# The Gains from Dimensionality

De baten van dimensionaliteit

Thesis

to obtain the degree of Doctor from the
Erasmus University Rotterdam
by command of the
rector magnificus

prof.dr. R.C.M.E. Engels

and in accordance with the decision of the Doctorate Board.

The public defense shall be held on

Thursday, July 5, 2018 at 13:30 hours

by

DIDIER NIBBERING

born in Capelle aan den IJssel, The Netherlands

**Erasmus University Rotterdam**

## Doctorate Committee

| | |
|---|---|
| **Promotor:** | Prof. dr. R. Paap |
| **Other members:** | Prof. dr. A.C.D. Donkers |
| | Prof. dr. D. Fok |
| | Prof. dr. F.R. Kleibergen |
| **Copromotor:** | Dr. M. van der Wel |

ISBN: 978 90 361 0520 0



This book is no. 716 of the Tinbergen Institute Research Series, established through cooperation between Thela Thesis and the Tinbergen Institute. A list of books which already appeared in the series can be found in the back.

# Acknowledgments

Although only my name is on its cover, many people have contributed to this dissertation.

First of all, I would like to thank my advisors. From the very beginning of my PhD track they shared many of their ideas and suggestions, and gave me the freedom to work on my own ideas. They managed to cope with my impatience, probably also by encouraging me to attend many conferences far away from Rotterdam. This dissertation greatly benefited from the devotion, flexibility, and honesty of Richard Paap. I am grateful for the enthusiasm of Michel van der Wel, his eye for detail, and his help in understanding the academic world.

Secondly, I am deeply indebted to Tom Boot. As coauthor of two chapters he contributed significantly to this dissertation. Our projects proved to me that research is the most fun when you explore new econometric horizons together. Thank you Tom, for your open-mindedness, guidance, and friendship.

Thanks to Bas Donkers, Dennis Fok, and Frank Kleibergen for reading the manuscript of this thesis and for your recommendations to improve it.

I thank my colleagues for the discussions about our research challenges and all other not directly related topics. This thesis greatly improved by the open door policy by many of my colleagues, but especially by my paranympf Bart Keijsers. Thank you for your listening ear and your helpful thoughts. Thanks to my other paranympf, Matthijs Oosterveen, for showing me that doing research can go hand in hand with the serious things in life.

I am grateful to many other people who supported me with their friendship. This dissertation especially benefited from the interest of Job Overbeek in my research. Your inability to think inside the box makes you the perfect sparring partner.

Thank you, Henk and Debbie, for teaching me to dream big. Your trust in me is my most valuable asset.

Thank you Evi, for supporting my ideas, from which writing this thesis is not even the craziest one..

# Contents

# Chapter 1

# Introduction

The rapid increase in available data is a promising development for empirical economic research along several dimensions. For instance, the analysis of large macroeconomic data sets may provide new insights in interactions between economic quantities. The increasing numbers of predictors may lead to improvements in economic forecast accuracy. Data on a large number of individual characteristics potentially leads to a better understanding of the drivers of observed behavioral patterns.

However, econometricians are challenged by the dimensions of many of these modern data sets, in which the number of explanatory variables approaches, or even exceeds, the number of available observations. The large number of macroeconomic indicators in time series data sets of Stock and Watson (2002) and McCracken and Ng (2016) have a limited number of observations due to the monthly frequency. Cross-sectional data sets on economic growth, as in Barro and Lee (1993), Sala-i-Martin (1997), and Fernandez et al. (2001), consist of a large number of explanatory variables for which the number of observations is bounded by the number of countries. Microeconomic data contain records of choices from a large number of alternatives which can potentially be explained by a large number of individual characteristics (Naik et al., 2008). The ratio of observations to explanatory variables is even smaller in studies on the relation between the human genome and later in life outcomes such as educational attainment by Rietveld et al. (2013).

Traditional econometric methods have a hard time in dealing with the dimensions of modern data sets. Popular estimation methods, such as ordinary least squares, are simply infeasible when the number of parameters exceeds the number of observations. Even when

parameter estimation is feasible, the large number of parameter estimates, potentially accompanied by high parameter estimation uncertainty, barely provides insight into the data. This thesis challenges the "curse of dimensionality" by introducing new ways to gain from dimensionality.

Credible assumptions are necessary to make the estimation of valuable relations in high-dimensional settings possible. This thesis examines two approaches. The first assumes that parameters can be clustered in groups with identical parameter values. The second imposes restrictions on the parameter magnitude.

The first approach, the idea of parameter clustering, is widely used to allow for flexible and parsimonous model specifications. Bonhomme and Manresa (2015) propose a parsimonious alternative to fixed effects and random effects models by modelling unobserved heterogeneity in panel data with fixed effects which are heterogenous across groups of individuals. Su et al. (2016) allow all regression parameters in panel data models to differ across groups. Mixture models are used to cluster individuals in groups with identical parameter values, in both panel data and cross-sectional data (Frühwirth-Schnatter, 2006). Regime switching models in the tradition of Hamilton (1989) cluster parameters over time in regimes with similar values. In choice data, researchers often cluster choice alternatives to a higher level when the choice set is large (Carson and Louviere, 2014). By aggregating dummy categories, categorical explanatory variables are also clustered.

The assumption that parameters can be clustered over certain dimensions, such as individuals, time, or categories, is valid in many applications. When analyzing a large number of individuals, one usually finds groups of individuals with homogeneous parameter values. While there is strong evidence that the behavior of macroeconomic variables changes over time, it is implausible that the economy changes in each time period. Decision makers facing a large choice set, probably do not have a strict preference ordering over all different choice alternatives.

The second approach to tackle the curse of dimensionality imposes restrictions on the magnitude of the parameters. For instance, a sparsity assumption restricts the number of nonzero parameters. The parameters in such a sparse model can be much easier to estimate and to interpret than in a model where the number of nonzero parameters is unrestricted. Even when the number of explanatory variables greatly exceeds the number of observations in a sparse model, the lasso estimator of Tibshirani (1996) is able to select the relevant explanatory

variables with high probability by setting the parameter values of the other variables equal to zero. Moreover, lasso also has been shown to have a high prediction accuracy under sparsity (Hastie et al., 2015).

Restricting the magnitude of parameters has been proven useful in different applications. Kock and Callot (2015) analyze a vector autoregressive model, a popular macroeconomic model in which the number of economic indicators as well as the number of lags is typically large. Since the indicators are observed at a quarterly frequency, the number of variables easily exceeds the number of observations. However, it is hard to argue that all lags of all variables are relevant for each economic indicator. Tibshirani et al. (2005) use thousands of genes to predict diseases like cancer, while they have less than hundred samples available. By assuming that only a small set of genes is related to the variable of interest, they can detect genes with predictive power. Belloni et al. (2010) estimate the effect of interest, for instance the effect of the initial level of GDP on the growth rates of GDP or the returns to schooling, while including all available control variables. By assuming that only a small set of controls affect the estimates of interest, the lasso can be used to select the relevant variables.

This thesis uses parameter clustering and restrictions on parameter magnitudes to achieve two different goals: accurate forecasting and valid inference in a data-rich environment. The next three chapters focus on economic forecasting. Chapter 2 examines the forecast performance of the Survey of Professional Forecasters. Next, Chapter 3 follows the first approach to deal with a high-dimensional parameter space. It clusters parameters over time to improve upon existing macro-economic forecasting models that account for time-varying effects. The fourth chapter uses the second approach, by assuming that the magnitude of the parameter values is small, and shows how machine learning methods can improve the accuracy of economic forecasts in this setting. The final two chapters of the thesis develop methods for inference in high-dimensional models. Chapter 5 introduces an inference toolbox for high-dimensional linear regression models. This toolbox relies on a sparsity assumption on the high-dimensional parameter vector, which fits in the second approach. Chapter 6 applies the first approach to enable inference in high-dimensional choice models, by clustering parameters over choice alternatives and explanatory dummy categories.

## Forecasting in a data-rich environment

One way to forecast in a data-rich environment, is to estimate parameters in a high-dimensional econometric model. Alternatively, the forecaster can rely on (other) professionals. Professional forecasters may use any model, but can also use other methods to incorporate the available information in their predictions.

In Chapter 2, based on Nibbering et al. (2017b), we examine what professional forecasters actually predict. We use spectral analysis and state space modeling to decompose economic time series into a trend, business-cycle, and irregular component. To examine which components are captured by professional forecasters, we regress their forecasts on the estimated components extracted from both the spectral analysis and the state space model. For both decomposition methods we find that the Survey of Professional Forecasters in the short run can predict almost all variation in the time series due to the trend and business-cycle, but the forecasts contain little or no significant information about the variation in the irregular component. A simple state space model, which is commonly used to estimate trends and cycles in time series, can produce almost the same predictions.

The finding from Chapter 2 that professional forecasters do not perform much better than models, is a good excuse to devote the next two chapters on improving econometric forecasting methods. The next chapter is about one of the main challenges in macroeconomic forecasting: How to take the instabilities over time into account? Typical macroeconometric models, vector autoregressions, contain a large number of (lagged) variables. When all these parameters are allowed to change over time, the parameter space easily increases to dimensions that make estimation infeasible. Therefore, feasible estimation methods restrict the law of motion of the parameters to only smooth continuous changes, or to only occasional jumps in parameter values.

However, the economy is characterized by a wide variety of shocks, which can be smooth or more abrupt. To account for both, Chapter 3 proposes a Bayesian infinite hidden Markov model to estimate time-varying parameters in a vector autoregressive model. The Markov structure allows for heterogeneity over time while accounting for state-persistence. By modeling the transition distribution as a Dirichlet process mixture model, parameters can vary over potentially an infinite number of regimes. The Dirichlet process however favours a parsimonious model without imposing restrictions on the parameter space. An empirical

application demonstrates the ability of the model to capture both smooth and abrupt parameter changes over time, and a real-time forecasting exercise shows excellent predictive performance even in large dimensional vector autoregressive models. Chapter 3 is based on Nibbering et al. (2017a).

The methods proposed in Chapter 3 allow for a very flexible model that includes a large set of economic indicators without restricting the form of heterogeneity over time. However, due to the large number of time-varying parameters, the estimation routine is complex. This fits in with a more general trend in econometrics, where methods become more and more involved to use all available information in a flexible way to arrive at an estimate or a forecast. The question is whether this hinge to complexity pays off. Chapter 4 shows that surprisingly simple methods achieve a forecast accuracy that is at least similar to the performance of sophisticated econometric methods when the predictor set is large.

Chapter 4, based on Boot and Nibbering (2017a), discusses a novel approach to obtain accurate forecasts in high-dimensional regression settings. Random subspace methods construct forecasts from random subsets of predictors or randomly weighted predictors. We provide a theoretical justification for these strategies by deriving bounds on their asymptotic mean squared forecast error, which are highly informative on the scenarios where the methods work well. Monte Carlo simulations confirm the theoretical findings and show improvements in predictive accuracy relative to widely used benchmarks. The predictive accuracy on monthly macroeconomic FRED-MD data increases substantially, with random subspace methods outperforming all competing methods for at least 66% of the series.

## Inference in a data-rich environment

The final two chapters do not focus on forecasting, but develop feasible parameter estimation methods for high-dimensional models with in-sample analysis as main goal. Firstly, we consider a simple linear regression model where the number of variables exceeds the number of available observations. Since the covariance matrix of the regressors is singular in this case, ordinary least squares is not feasible.

Chapter 5 is based on Boot and Nibbering (2017b) and examines this curse of dimensionality in a linear regression model. We introduce an asymptotically unbiased estimator for the full high-dimensional parameter vector. The estimator is accompanied by a closed-

form expression for the covariance matrix of the estimates that is free of tuning parameters. This enables the construction of confidence intervals that are valid uniformly over the parameter vector. Estimates are obtained by using a scaled Moore-Penrose pseudoinverse as an approximate inverse of the singular empirical covariance matrix of the regressors. The approximation induces a bias, which is then corrected for using the lasso. Regularization of the pseudoinverse is shown to yield narrower confidence intervals under a suitable choice of the regularization parameter.

The number of parameters increases with the number of variables in a simple linear regression model. However, the number of parameters in a standard multinomial choice model increases linearly with the number of choice alternatives and the number of explanatory variables. Since many modern applications involve large choice sets with categorical explanatory variables, which enter the model as large sets of binary dummies, the number of parameters easily approaches the sample size. This may result in overfitting and a substantial amount of parameter uncertainty.

Chapter 6 considers the setting of modern choice data sets. This chapter, which is based on Nibbering (2017), proposes methods for data-driven parameter clustering over outcome categories and explanatory dummy categories in a multinomial probit setting. A Dirichlet process mixture encourages parameters to cluster over the categories, which favours a parsimonious model specification without a priori imposing model restrictions. Simulation studies and an application to a data set of holiday destinations show a decrease in parameter uncertainty and an enhancement of the parameter interpretability, relative to a standard multinomial choice model.

# Chapter 2

# What Do Professional Forecasters Actually Predict?

*Joint work with Richard Paap and Michel van der Wel*

## 2.1 Introduction

Econometric models cannot accurately predict events when developers of the models fail to include information about main drivers of the outcomes. The global financial crisis is an example of the failure of models to account for the actual evolution of the real-world economy (Colander et al., 2009). Besides econometric models also surveys of forecasters provide predictions about key economic variables. Although professional forecasters cannot predict one-off events, like natural disasters, they may be quicker in taking into account interpretations of news and various expert opinions than econometric models before they form a final prediction. Fiscal, political, or weather conditions can be reasons for experts to arrive at predictions different from model-based forecasts. According to the amount of attention these surveys receive, they are perceived to contain useful information about the economy (as Ghysels and Wright (2009) note).

In this chapter we examine what professional forecasters actually are able to predict. Do they explain movements in economic time series which can also be explained by regular components like a trend or a business-cycle, or also a part of the irregular component, which can hardly be predicted by econometric models and non-experts? To address this question,

we decompose 5 key economic variables (GDP, the GDP deflator, unemployment, industrial production and housing starts) of the US economy in three components. Subsequently we examine whether panelists of the Survey of Professional Forecasters can explain the variation in the time series due to the different estimated components.

To decompose the economic variables we apply two commonly used methods in the literature to extract trends and business-cycles from time series. First we apply the Baxter and King (1999) low-pass filter which Baxter (1994) uses for the decomposition of exchange rates series into a trend, business-cycle, and irregular component. Second, we also decompose the time series into trend, cycle, and an irregular component using a state space model which is studied by Harvey (1985). Since each decomposition relies on different assumptions, we perform both methods and assess whether the results are robust. The low-pass filter and state space model are used to estimate the trend and cycle as precisely as possible, and are not considered as the true data generating process for the observed time series. Next, we regress the forecasts of the professional forecasters on the estimated components in both the spectral analysis and the state space model. We deal with the presence of a unit root in the forecasts and the estimated trend by using the framework of Park and Phillips (1989). To account for two-step uncertainty in the standard errors we implement the Murphy and Topel (2002) procedure.

Our results show that the professional forecasters can predict almost all variation in the time series due to the trend and the business-cycle components in the short-run, but explain little or even nothing of the variation in the irregular component. The small amount of variation in the irregular components that the professional forecasters capture may explain why some businesses and policymakers rely on professional forecasters. Both approaches to decompose the time series lead to approximately the same results in the forecast regressions. For larger forecast horizons, prediction of the cyclical component becomes worse. The results look very similar if we replace the professional forecasts by simple time series model forecasts. Professional forecasters perform slightly better with respect to root mean squared prediction error than a structural time series model, which is commonly used to estimate trends and cycles in time series. The difference is however only significant in a particular sample period. Finally, results suggest that professional forecasters seem to explain the realized values in the current period, which is already published, instead of explaining irregular events in the future.

Although forecast performance is a widely debated topic, we are, to the best of our knowledge, the first to assess forecasts from the perspective of 'what' is predicted instead of 'how good' the actual values are predicted. Hyndman and Koehler (2006) state that "despite two decades of papers on measures of forecast error" the recommended measures still have some fundamental problems. Moreover, all these measures are relative and have to be compared to a benchmark model. By assessing whether a significant amount of variation of the different components of a time series can be explained, no benchmark forecast is needed. Leitch and Ernesttanner (1995) show that conventional forecast evaluation criteria have little to do with the profitability of forecasts, which determines why firms spends millions of dollars to purchase professional forecasts. These firms may believe that experts have information about irregular movements in the future which cannot be predicted by econometric models.

The performance of professional forecasts have been subject to a number of studies. Thomas et al. (1999), Mehra (2002), and Gil-Alana et al. (2012) show that forecast surveys outperform benchmark models for forecasting inflation. These papers focus on the relative strength of expert forecasts in comparison to other forecast methods. In a comprehensive study, Ang et al. (2007) also show that professional forecasters outperform other forecasting methods in predicting inflation by means of relative measures and combinations of forecast methods. Instead of focusing on the relative strength of expert forecasts, we question what professional forecasters actually predict. Moreover, where other studies focus only on forecasting inflation, we also consider other key variables of the US economy. Franses et al. (2011) examine forecasts of various Dutch macroeconomic variables and conclude that expert forecasts are more accurate than model-based forecasts. Other papers show limited added value of professionals' forecasts. Franses and Legerstee (2010) show that in general experts are worse than econometric models in forecasting sales at the stock keeping unit level. Isiklar et al. (2006) find that professional forecasts of Consensus Economics do not include all available new information. Coibion and Gorodnichenko (2012) and Coibion and Gorodnichenko (2015) find persistence in the forecast errors for the GDP deflator of the Survey of Professional Forecasters. In a comparison between forecasts of professional forecasters and their long-run expectations, Clements (2015) finds little evidence that the forecasts of the Survey of Professional Forecasters are more accurate than forecasting the trend. Billio et al. (2013) show that the performance trade-off between a white noise model

and professional forecasts in predicting returns differs over time. There is also a literature that uses professional forecasts to improve models. For instance, Kozicki and Tinsley (2012) incorporates survey data in a model for inflation to have timely information on structural change, Mertens (2016) estimates trend inflation with the help of survey expectations, and Altug and Çakmaklı (2016) claim superior predictive power of models for inflation incorporating survey expectations.

The outline of this chapter is as follows. Section 2.2 explains the decomposition methods of the economic time series and the forecast regressions of the professional forecasts on the estimated components. Section 2.3 describes the economic time series and the corresponding forecasts from the Survey of Professional Forecasters, on which we apply the methods. Section 2.4 discusses the results obtained from the time series decompositions and the forecast regressions. Section 2.5 provides comparisons between professional and model-based forecasts to provide more insight in the results. We conclude with a discussion in Section 2.6.

## 2.2 Methods

To examine what professional forecasters actually forecast, we decompose the historical values for the predicted time series into three components; a trend, business-cycle, and irregular component. Since most macroeconomic surveys provide seasonally adjusted data, we consider seasonally adjusted time series and hence do not model the seasonal component. However, we argue that our methodology can easily be extended to seasonally unadjusted data. There are two common methods in the literature for decomposing time series; filters in the frequency domain and state space modeling in the time domain. Since each method relies on different assumptions (Harvey and Trimbur, 2003), we perform both methods and assess whether the results correspond with each other. In Section 2.2.1, we discuss the filtering of different components from the time series in a spectral analysis. Section 2.2.2 deals with the trend-cycle decomposition in a state space framework. Finally, Section 2.2.3 assesses the forecast regression, where we regress the professional forecasts on both the estimated components in the spectral analysis and on the estimated components in the state space framework. The estimated coefficients in these forecast regressions indicate which components can be explained by the professional forecasters.

### 2.2.1 Spectral Analysis

We consider the model

$$y_t = \mu_t + c_t + \varepsilon_t, \tag{2.1}$$

where $y_t$ is the observed time series, $\mu_t$ represents the trend, $c_t$ the business-cycle, and $\varepsilon_t$ the irregular component. In other words, we have a slow-moving component, an intermediate component, and a high-frequency component. We isolate these different frequency bands by a low-pass filter derived by Baxter and King (1999). They obtain the component time series by applying moving averages to the observed time series. The time series in a specific frequency band can be isolated by choosing the appropriate weights in the moving average.

The filter produces a new time series $x_t$ by applying a symmetric moving average to the filtered time series $y_t$:

$$x_t = \sum_{k=-K}^{K} a_k y_{t-k}, \tag{2.2}$$

with weights $a_k = a_{-k}$ specified as

$$a_k = b_k + \theta, \tag{2.3}$$

$$b_k = \begin{cases} \omega/\pi & \text{if } k = 0 \\ \sin(k\omega)/(k\omega) & \text{if } k = 1, ..., K, \end{cases} \tag{2.4}$$

where

$$\theta = \left(1 - \sum_{k=-K}^{K} b_k\right) / (2K+1) \tag{2.5}$$

is the normalizing constant which ensures that the low-pass filter places unit weight at the zero frequency. We denote the low-pass filter by $LP_K(p)$ where $K$ is the lag parameter for which $K = 12$ is assessed as appropriate for quarterly data by Baxter and King (1999). This means that we use twelve leads and lags of the data to construct the filter, so three years of observations are lost at the beginning and the end of the sample period. The periodicity $p$ of cycles is a function of the frequency $\omega$: $p = 2\pi/\omega$. We follow Baxter and King (1999) in the

definition of the business-cycle as cyclical components of no less than six quarters and fewer than 32 quarters in duration, and assign all components at lower frequency to the trend and higher frequencies to the irregular component. Thus, the filtered trend equals $LP_{12}(32)$ and the filtered business-cycle $LP_{12}(6) - LP_{12}(32)$. The filtered irregular component equals the original time series $y_t$ minus the filtered trend and filtered business-cycle component. Note that the low-pass filter "filters" two-sided estimates for the components which can also be referred to as smoothed estimates of the time series components. It is possible to apply the low-pass filter to seasonally unadjusted data by adding an additional frequency band to the band-pass filter.

Beside the Baxter and King filter there are more filtering methods in the frequency domain which can be used for extracting the trend and the business-cycle component from a time series. For example, Christiano and Fitzgerald (2003) and Pollock (2000) also propose frequency filters suitable for decomposing time series in three components. In our application we will show that these filters provide similar results as the Baxter and King filter.

### 2.2.2 State Space Model

Although the Baxter and King filter is a simple and effective methodology in extracting trends and cycles from time series, it does not allow for making any statistical inference on the components. Therefore we also estimate the components in a model-based approach, in which we obtain confidence intervals for the estimated component series. Moreover, we can estimate the periodicity of the cycle within the model instead of arbitrarily choosing the frequency bands. However, estimation of the model parameters must be feasible, and also in the time domain we have to make assumptions on the functional form of the model.

A well-known model-based approach in time series decomposition is the state space framework based on the basic structural time series model of Harvey (1990). After including a cyclical component representing the business-cycle, we consider the following model;

$$y_t = \mu_t + c_t + \varepsilon_t, \qquad \varepsilon_t \sim N(0, \sigma_\varepsilon^2), \tag{2.6}$$

where $y_t$ is the observed time series, $\mu_t$ represents the trend, $c_t$ the business-cycle, and $\varepsilon_t$ the irregular component with variance $\sigma_\varepsilon^2$. The trend component is specified by the local linear

trend model

$$\mu_{t+1} = \mu_t + \nu_t + \xi_t, \qquad \xi_t \sim N(0, \sigma_\xi^2), \tag{2.7}$$

$$\nu_{t+1} = \nu_t + \zeta_t, \qquad \zeta_t \sim N(0, \sigma_\zeta^2), \tag{2.8}$$

where $\nu_t$ represents the slope of the trend, and $\sigma_\xi^2$ and $\sigma_\zeta^2$ are the variances of the shocks. We opt for a smooth stochastic trend specification as in, for example, Durbin and Koopman (2012), by restricting $\sigma_\xi^2$ to zero. The business-cycle component is represented by the following relations

$$c_{t+1} = \rho c_t \cos\lambda + \rho c_t^* \sin\lambda + \kappa_t, \qquad \kappa_t \sim N(0, \sigma_\kappa^2), \tag{2.9}$$

$$c_{t+1}^* = -\rho c_t \sin\lambda + \rho c_t^* \cos\lambda + \kappa_t^*, \qquad \kappa_t^* \sim N(0, \sigma_\kappa^2), \tag{2.10}$$

where the unknown coefficients $\rho, \lambda,$ and $\sigma_\kappa^2$ represent the damping factor, the cyclical frequency, and the cycle error term variance, respectively. The period of the cycle equals $2\pi/\lambda$ and we impose the restrictions $0 < \rho < 1$ and $0 < \lambda < \pi$. For seasonally unadjusted data we can add a cycle with a seasonal frequency to the state space model, to obtain an extra component which captures the seasonal variation.

We estimate the unknown parameters $(\sigma_\varepsilon^2, \sigma_\xi^2, \sigma_\zeta^2, \sigma_\kappa^2, \rho, \lambda)$ in a state space framework;

$$y_t = Z\alpha_t + \varepsilon_t, \qquad \varepsilon_t \sim N(0, \sigma_\varepsilon^2), \tag{2.11}$$

$$\alpha_{t+1} = T\alpha_t + \eta_t, \qquad \eta_t \sim N(0, Q), \tag{2.12}$$

where the observation equation relates the observation $y_t$ to the unobserved state vector $\alpha_t$, which contains the trend and the cycle. This vector is modeled in the state equation. We use Kalman filtering and smoothing to obtain maximum likelihood parameter estimates and estimates for the state vector components (see, e.g., Durbin and Koopman (2012)).

Where the objective of the estimation routine is to minimize the observation noise $\varepsilon_t$ relative to the trend and the cycle, we are in this case also interested in the irregular component. So instead of allocating all variance in the time series to the trend and cycle components, the observation noise has to capture the irregular movement. To prevent the variance of the observation noise $\sigma_\varepsilon^2$ from going to zero, we fix it to the value of the variance of the esti-

mated irregular component in the low-pass filter.[1] As we show in Section 2.4.2, our results are robust with respect to alternative values for the variance of the observation noise.

### 2.2.3 Forecast Regression

Both the spectral analysis and the state space model yield a decomposition of the actual values in the historical time series. From here we investigate how the professional forecasts are related to the components of the historical time series by the regression equation

$$f_{t+h|t} = \beta_0 + \beta_1 \hat{\mu}_{t+h|T} + \beta_2 \hat{c}_{t+h|T} + \beta_3 \hat{\varepsilon}_{t+h|T} + v_{t+h}, \tag{2.13}$$

where $f_{t+h|t}$ is the professional forecast for $h$ time periods ahead conditional on the information known in time period $t$. The $\hat{\mu}_{t+h|T}$ represents the estimated trend, $\hat{c}_{t+h|T}$ the estimated business-cycle, and $\hat{\varepsilon}_{t+h|T}$ the irregular component. We consider the irregular component, which is constructed as the observed time series from which the estimated trend and cycle is removed, as estimate for the irregular variation in the observed time series. The components are estimated using the series $y_t$ for $t = 1, \ldots, T$, where each $y_t$ contains the observed value just before sending the survey in time period $t + h$. When the professional forecasters perfectly predict the actual values, we have $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3) = (0, 1, 1, 1)$ as the estimated components add up to the actual values. The coefficient $\beta_0$ accounts for a potential forecast bias in case the coefficients of the estimated components equal one.

It is good to emphasize that we do not consider the models in (2.1) and (2.6) as the true data generating process for the observed time series. The purpose of these models is to estimate the trend and cycle as precisely as possible. The irregular component is what is left over in the actual series after reasonable estimates of the trend and cycle have been removed. The structural time series model (2.6) imposes that the trend and cycle components are independent. As the trend/cycle estimates follow from filters, the estimated irregular component may be serially correlated as well. The persistence in our estimated irregular components is however very low. We want to address whether, despite the assumed absence of persistence in this component, professional forecasters can possibly predict some of the variation in the irregular component due to their expert information.

[1] Stock and Watson (1998) develop estimators and confidence intervals for the parameters in a state space model where the maximum likelihood estimator of the variance of the stochastic trend has a large point mass at zero. Our situation is different, as we restrict the variance of the observation noise.

Since many economic time series exhibit trending behavior, we expect a stochastic trend in the series of professional forecasts. We explicitly model a unit root in the local linear trend model in the state space framework. Unless professional forecasters have done a very poor job, there is a long-run relationship between the stochastic trend of the economic time series and the predicted values for this variable. So, we expect that the forecasts and the estimated trend are cointegrated. To examine this conjecture, we test in our empirical analysis for cointegration between the professional forecasts and the estimated trend with the Engle and Granger (1987) residual-based cointegration test.

In case of cointegration, we have in (2.13) a regression with cointegrated variables $f$ and $\hat{\mu}$ together with the $I(0)$ variables $\hat{c}$ and $\hat{\varepsilon}$. Park and Phillips (1989) show that in this situation the parameters can be consistently estimated with ordinary least squares. They also provide asymptotically chi-squared distributed Wald test statistics for inference on the estimated parameters (Park and Phillips, 1989, p. 108). We test whether the estimated parameters are individually equal to the values in a perfect forecast. Moreover, we test the null hypothesis of perfectly predicted values, that is $\beta = (0, 1, 1, 1)$.

The standard errors of the estimated coefficients in (2.13) do not account for the uncertainty in the regressors. Due to the fact that the regressors are estimates we may encounter heteroskedasticity in the residuals. Therefore we opt for White standard errors when the components are estimated in the spectral analysis (White, 1980). One of the benefits of the state space model is that here we do obtain estimates of the uncertainty in the model parameters. We can exploit the estimated parameter uncertainty in the state space framework by implementing the Murphy and Topel (2002) procedure for computing two-step standard errors. Adjusting the standard covariance matrix of the forecast regression parameters with the state space model parameter covariance matrix results in asymptotically correct standard errors.

It might be appealing to simultaneously estimate the historical time series components using (2.6)–(2.10) and the forecast regression coefficients in (2.13) by including the forecast regression in the state space framework. In this way we directly estimate standard errors for the estimated forecast regression coefficients, without the concern that we ignore the uncertainty in the estimated components. However, this approach allows the forecasts to influence the estimates of the components of the historical time series, which leads to incorrect inference. For this reason we do not consider this simultaneous set-up.

Finally, we want to stress that we use regression (2.13) to infer the correlations between the components of the historical time series and the predictions. We do not assume that forecasters really use the estimated components to arrive at their predictions, or make any other assumption about the generating process of predictions. Hence, we do not intend to make causality statements.

## 2.3 Data

We apply the methods of Section 2.2 to the well-documented and open database of the Survey of Professional Forecasters. We focus on key variables of the US economy which are available over a long period. We consider real-time data of nominal GDP, GDP deflator, unemployment, industrial production index, and housing starts. The forecasts for the Survey of Professional Forecasters are provided by the Federal Reserve Bank of Philadelphia.

To determine the information sets of the forecasters at the moment of providing the forecasts, we consider the timing of the survey. The quarterly survey, formerly conducted by the American Statistical Association and the National Bureau of Economic Research, began in the last quarter of 1968 and was taken over by the Philadelphia Fed in the second quarter of 1990. We collect data up to the second quarter of 2014. Table 2.1 shows all relevant information concerning the timing of the survey since it is conducted by the Philadelphia Fed. There is still some uncertainty about the timing before mid 1990 but the Philadelphia Fed assumes that it is similar to the timing afterwards. Based on this information we suppose that all panelists in the survey are informed about the actual values of the predicted variables up to and including the previous quarter. We use the same information set for constructing the model-based forecasts. Although the exact day of the month on which forecasters have to submit their predictions differs over the surveys, our results in Section 2.4.2 turn out to be robust to the differences in timing and to the takeover of the survey by the Philadelphia Fed.

Since the individual forecasters in the survey have limited histories of responses and forecasters may switch identification numbers, we mainly use time series of mean forecasts for the level of economic variables for which the data set includes observations over the whole survey period. The forecasts of the survey panelists are averaged in each time period. Beside the forecasts, the database of the Survey of Professional Forecasters also provides the real-time quarterly historical values corresponding to the predicted series. These historical

**Table 2.1:** Timing Survey of Professional Forecasters 1990:Q3 to present

| Survey | Questionnaires Sent | Last Quarter in Panelists' Information Sets | Deadline Submissions | Results Released |
|---|---|---|---|---|
| Q1 | End of January | Q4 | Middle of February | Late February |
| Q2 | End of April | Q1 | Middle of May | Late May |
| Q3 | End of July | Q2 | Middle of August | Late August |
| Q4 | End of October | Q3 | Middle of November | Late November |

The first three columns of this table provide the dates on which the survey for the current quarter is sent to the panelists and the last quarter of the series of actual historical values that is in the panelists' information set at this moment. The last two columns indicate when the forecasts for the current quarter must be submitted and when the results of these forecasts are released.

values are included in the information sets of the panelists, before they receive the survey for the next quarter. Therefore, the Real-Time Data Set for Macroeconomists (Croushore and Stark, 2001) could contain different values when there is a new release of the data after the survey is send but before the deadline for returning it. We assess the predictive performance against the time series decompositions of the real-time historical values provided by the survey.

Table 2.2 lists the series, which are all seasonally adjusted. The unemployment rate, index of industrial production, and housing starts are averaged over the underlying monthly levels. The base year for the GDP deflator and the index of industrial production changed several times in the considered sample period. We rescale the time series to base year 1958 in case of the GDP deflator and 1957-1959 in case of the index of industrial production. All base year changes, temporal aggregation, and a detailed explanation of the Survey of Professional Forecasters can be found in the documentation of the Federal Reserve Bank of Philadelphia.[2]

In this chapter we consider the logarithm of all historical time series and forecasts multiplied by one hundred. Figure 2.1 shows these key variables of the US economy. The solid line corresponds to the historical time series and the dashed dotted line to the difference between the historical values and the predictions by the Survey of Professional Forecasters. We recognize an upward trend in nominal GDP, GDP deflator, and industrial production index.

[2] http://www.philadelphiafed.org/research-and-data/real-time-center/survey-of-professional-forecasters/spf-documentation.pdf

**Table 2.2:** Variables Description

| Variable | Description |
|---|---|
| NGDP | Annual rate nominal GDP in billion dollars.<br>Prior to 1992 nominal GNP. |
| PGDP | GDP deflator with varying base years.<br>Prior to 1996 GDP implicit deflator and prior to 1992 GNP deflator. |
| UNEMP | Unemployment rate in percentage points. |
| INDPROD | Index of industrial production with varying base years. |
| HOUSING | Annual rate housing starts in millions. |

This table provides a short summary of each variable. All variables are seasonally adjusted.

The latter two also show some cyclical movements. From unemployment and housing we cannot directly identify a trend, but we see clear cyclical patterns in these series.

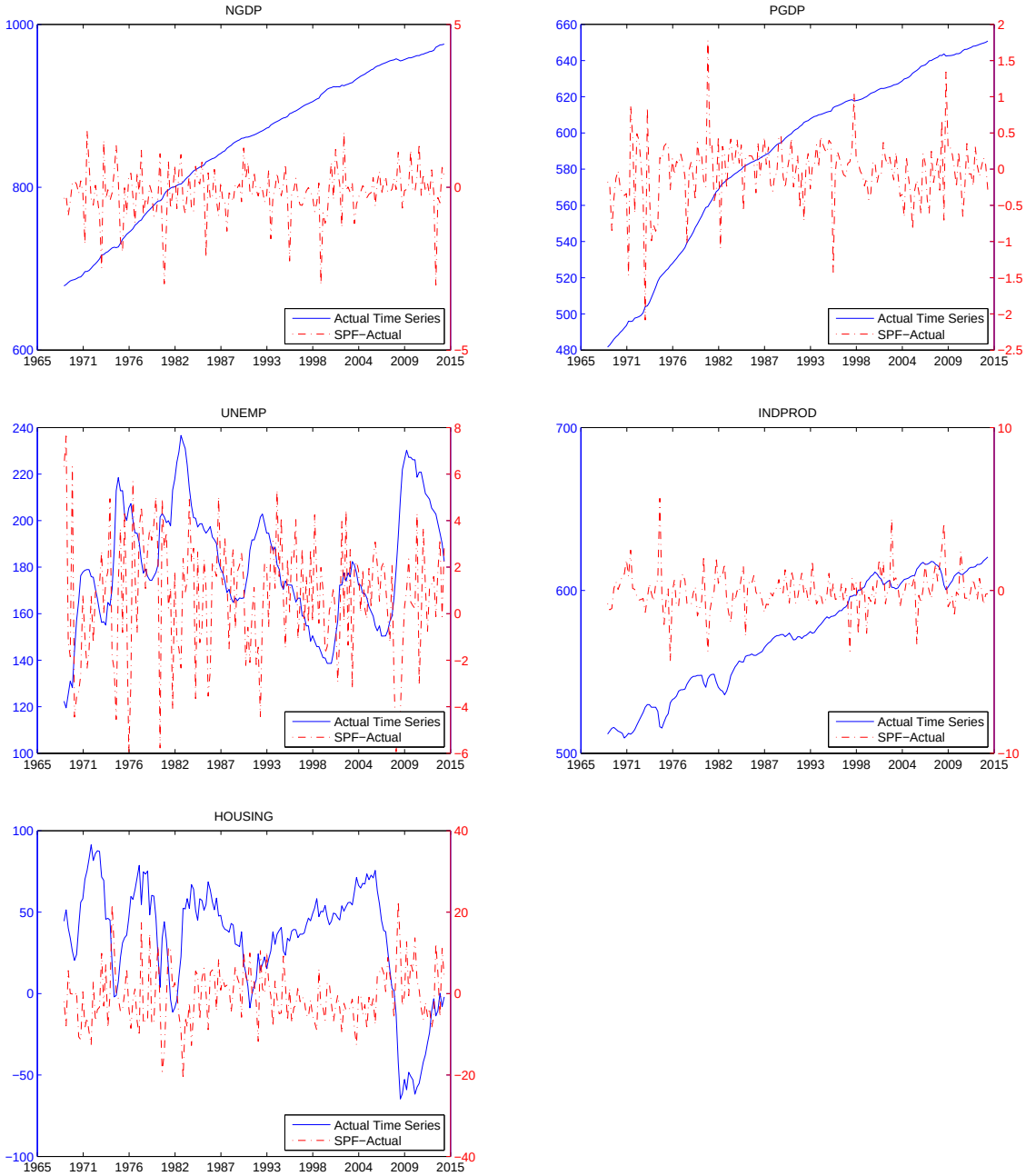
Table 2.3 shows the forecast bias for each variable computed as the average over the difference between the predictions of the survey of professional forecasters and the real-time historical values over different forecast horizons. A positive bias means that the professional forecasters on average overestimate the actual values. For the NGPD and PGDP series, the bias is almost always negative but small compared to the standard deviation. For the other series the bias is in most of the cases positive.

**Table 2.3:** Forecast Bias Estimates

| horizon | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| NGDP | -0.165<br>(0.750) | -0.277<br>(1.252) | -0.324<br>(1.709) | -0.332<br>(2.144) | -0.170<br>(2.590) |
| PGDP | -0.012<br>(0.446) | -0.025<br>(0.710) | -0.024<br>(0.997) | -0.025<br>(1.327) | 0.276<br>(1.241) |
| UNEMP | 0.799<br>(2.438) | 1.112<br>(5.587) | 0.787<br>(8.438) | 0.111<br>(11.434) | -0.110<br>(13.917) |
| INDPROD | -0.039<br>(1.284) | 0.122<br>(2.425) | 0.369<br>(3.493) | 0.704<br>(4.406) | 1.031<br>(5.080) |
| HOUSING | -0.391<br>(7.085) | 1.059<br>(12.192) | 3.141<br>(16.147) | 5.262<br>(19.948) | 6.628<br>(23.053) |

This table shows the forecast bias and the standard deviation in parentheses for each variable over different horizons. The bias is computed as the average over the difference between the predictions of the Survey of Professional Forecasters and the actual historical values. A positive bias means that the forecasters on average overestimate the actual values. Due to missing values, the estimation sample starts at 1974Q4 for $h = 5$.

**Figure 2.1:** Historical Time Series and the Survey of Professional Forecasters

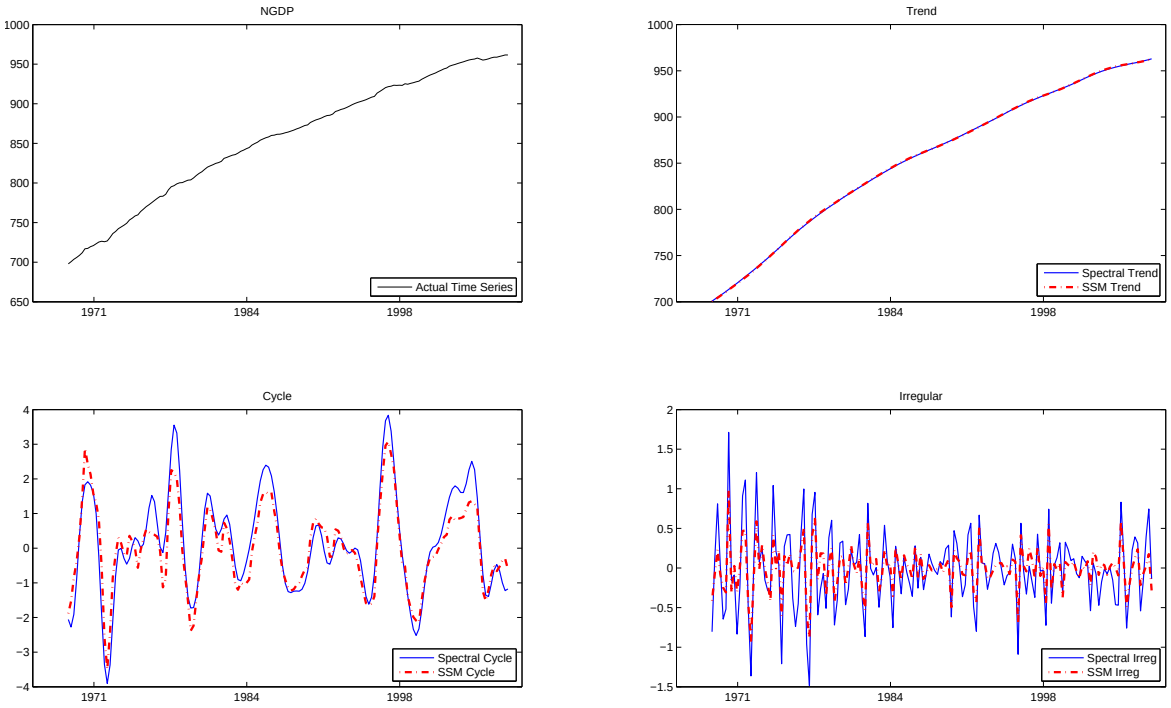



Historical time series (blue solid line, left axis) graphs together with the differences between the predictions of the Survey of Professional Forecasters and the actual values (red dashed dotted line, right axis). The figure shows the nominal GDP, GDP deflator, unemployment, industrial production index, and housing starts, respectively. The time series are log transformed and multiplied by one hundred.

## 2.4 Results

In this section we discuss the results of the analysis of the predictions of the Survey of Professional Forecasters. First, we consider the decomposition of the actual time series

**Figure 2.2:** Decomposition of Nominal GDP




Nominal GDP decomposed in a trend, a cycle, and an irregular component by the low-pass filters and the state space model. The first window shows one hundred times the logarithm of the actual values in the historical time series and the other windows show the components estimated in the low-pass filters by a blue solid line and the components estimated in the state space model by a red dashed dotted line.

based on both the frequency and time domain analysis. Second, we examine the relation between the professional forecasts and the estimated components. We first consider one-step ahead predications based on the mean of the professional forecasts, followed by the same analysis based on individual forecasts. We end this section by considering multiple-step ahead forecasts.

### 2.4.1 Time Series Decomposition

Figure 2.2 shows nominal GDP decomposed in a trend, a cycle, and an irregular component by the low-pass filters and the state space model. For all components the two time series follow roughly the same pattern. The fact that the two methods, which rely on different assumptions, result in approximately the same decomposition indicates that the estimated decompositions are reliable. We conclude the same for the other time series, that is GDP deflator, unemployment, industrial production index, and housing starts, for which Figure 2.7 up to Figure 2.10 can be found in Appendix 2.A.

**Table 2.4:** State Space Model Parameter Estimates

| | Estimate (Std. error) | | | | | Implied |
|---|---|---|---|---|---|---|
| | $\sigma_\varepsilon$ | $\sigma_\zeta$ | $\sigma_\kappa$ | $\lambda$ | $\rho$ | Cycle |
| NGDP | 0.489 | 0.142<br>(0.055) | 0.577<br>(0.091) | 0.330<br>(0.083) | 0.910<br>(0.034) | 19 |
| PGDP | 0.241 | 0.131<br>(0.030) | 0.218<br>(0.043) | 0.314<br>(0.035) | 0.954<br>(0.020) | 20 |
| UNEMP | 2.152 | 0.360<br>(0.194) | 3.791<br>(0.298) | 0.218<br>(0.019) | 0.978<br>(0.013) | 29 |
| INDPROD | 0.908 | 0.070<br>(0.037) | 1.454<br>(0.116) | 0.250<br>(0.029) | 0.948<br>(0.018) | 25 |
| HOUSING | 5.241 | 0.309<br>(0.181) | 6.721<br>(0.688) | 0.188<br>(0.028) | 0.965<br>(0.016) | 33 |

This table shows the parameter estimates in the state space model where the variance of the observation noise $\sigma_\varepsilon^2$ is fixed to the variance of the irregular component estimated by the low-pass filter. The $\sigma_\varepsilon$ represents the standard deviation of the observation noise, $\sigma_\zeta$ the second order trend error term standard deviation, $\sigma_\kappa$ the cycle error term standard deviation, $\lambda$ the cyclical frequency, and $\rho$ the damping factor. The standard errors of the estimates are reported in parentheses. The last column presents the period of the cycle (in quarters), implied by the $\lambda$ estimate.

Table 2.4 shows the state space model parameter estimates. Almost all parameter estimates are significant. The estimated period of the cycle in GDP equals nineteen quarters, which lies in the business-cycle period interval defined by Baxter and King. Except for housing starts (33 quarters), this is also the case for all other variables.

### 2.4.2 Forecast Regression

As discussed in Section 2.2.3, for correct inference of the forecast regression parameters in (2.13) the forecasts should be cointegrated with the estimated trend. Table 2.5 shows the Engle-Granger cointegration test results on both the estimated trend in the spectral analysis as the estimated trend in the state space model for the one-step ahead forecasts. The null hypothesis of no cointegration is rejected at a 5% significance level in all cases, except for the trend in the GDP deflator resulting from the spectral analysis. Hence, we have to be more careful interpreting the results of the forecast regression for this variable. For the other four variables we can straightforwardly use the Park and Phillips (1989) test statistics.

We include the estimated components in the forecast regression equation (2.13) with $h = 1$ to examine how the professional one-step ahead forecasts are related to the different components. Table 2.6 shows the results based on the estimated components in the spec-

**Table 2.5:** Cointegration Tests Forecast and Trend Time Series

| | Spectral Analysis | | | State Space Model | | |
|---|---|---|---|---|---|---|
| | $\tau$-stat. | lags | $p$-value | $\tau$-stat. | lags | $p$-value |
| NGDP | -5.806 | 1 | 0.000 | -6.136 | 1 | 0.000 |
| PGDP | -2.973 | 0 | 0.123 | -3.941 | 0 | 0.011 |
| UNEMP | -5.538 | 1 | 0.000 | -4.397 | 1 | 0.003 |
| INDPROD | -5.978 | 1 | 0.000 | -4.814 | 1 | 0.001 |
| HOUSING | -3.977 | 2 | 0.010 | -3.791 | 1 | 0.017 |

This table shows the Engle-Granger residual-based cointegration test of the null hypothesis of no cointegration against the alternative of cointegration. The professional one-step ahead forecast is the dependent variable and an intercept is included. The MacKinnon (1996) $p$-values are reported and the lag length is specified as the number of lagged differences in the test equation determined by the Schwarz criterion. The first four columns show the results based on the estimated trend in the spectral analysis and the last three columns the results based on the estimated trend in the state space model.

tral analysis and Table 2.7 shows the results based on the state space model. Due to the lag parameter in the spectral analysis, the filtered series start after twelve quarters from the beginning of the sample period and end twelve quarters before the end of the sample period. To make the results comparable we also exclude these observations from the estimated series in the state space framework, which results in a sample period from the last quarter of 1971 to the second quarter of 2011. Table 2.14 in Appendix 2.B reports the results of the spectral analysis based on the Christiano-Fitzgerald and the Butterworth filter. Since the outcomes are very similar, we only discuss the results based on the Baxter and King decomposition here.

Tables 2.6 and 2.7 show the estimated coefficients for each component with the standard errors in parentheses and the Wald test statistic on the null hypothesis that the coefficient is equal to the weight expected in a perfect forecast. That is, the intercept is tested against zero and the components against one. These Wald test statistics are asymptotically chi-squared distributed with critical value 3.842 at the 5% significance level. The asterisks indicate whether a coefficient significantly differs from the value that is expected in a perfect forecast. The first six columns of each table show the forecast regression for each variable with intercept, and the last four columns the results without intercept ($\beta_0 = 0$).

The first six columns of Table 2.6 show that the trend and cycle components receive a weight close to one. Although some of these estimates significantly differ from one due

**Table 2.6:** Forecast Regressions ($h = 1$) Based On Spectral Analysis

| | Estimate (Std. error) | | | | Estimate (Std. error) | | |
|---|---|---|---|---|---|---|---|
| | intercept | trend | cycle | irreg. | trend | cycle | irreg. |
| NGDP | −1.178 | 1.001 | 0.954 | 0.249* | 1.000* | 0.959 | 0.248* |
| | (0.620) | (0.001) | (0.037) | (0.149) | (0.000) | (0.038) | (0.154) |
| | 3.613 | 2.752 | 1.505 | 25.494 | 10.051 | 1.150 | 23.802 |
| PGDP | −0.197 | 1.000 | 0.990 | −0.132* | 1.000 | 0.992 | −0.133* |
| | (0.505) | (0.001) | (0.037) | (0.173) | (0.000) | (0.039) | (0.174) |
| | 0.153 | 0.120 | 0.080 | 42.95 | 0.839 | 0.045 | 42.302 |
| UNEMP | 1.318 | 0.997 | 0.949* | 0.581* | 1.004* | 0.945* | 0.587* |
| | (1.960) | (0.011) | (0.016) | (0.104) | (0.001) | (0.015) | (0.102) |
| | 0.452 | 0.067 | 9.966 | 16.208 | 18.975 | 13.982 | 16.418 |
| INDPROD | −3.491 | 1.006 | 0.938* | 0.441* | 1.000 | 0.939* | 0.440* |
| | (1.936) | (0.003) | (0.030) | (0.168) | (0.000) | (0.030) | (0.166) |
| | 3.251 | 3.194 | 4.386 | 11.122 | 0.102 | 4.246 | 11.401 |
| HOUSING | 2.555* | 0.919* | 0.888* | 0.239* | 0.973* | 0.847* | 0.252* |
| | (0.880) | (0.022) | (0.038) | (0.119) | (0.010) | (0.036) | (0.119) |
| | 8.423 | 13.960 | 8.832 | 40.781 | 6.847 | 18.427 | 39.817 |

This table shows the parameter estimates in forecast regression (2.13) of the professional forecasts on the low-pass filter decomposition, with and without intercept. White standard errors are reported in parentheses together with Wald test statistics on the null hypothesis that the coefficient is equal to the weight expected in a perfect forecast. An asterisk ($*$) denotes that the coefficient significantly differs from the weight expected in a perfect forecast at the 5% significance level.

to the small standard errors, we can say that the professional forecasters can predict most of the variation caused by a trend and a business-cycle. However, the parameter estimates corresponding to the irregular component differ significantly from one while having large standard errors. Moreover, some of the weights of the irregular components do significantly differ from zero, which means that the professional forecasters still seem to capture a bit of the irregular movements in the time series.

When the weights of the estimated components equal one, the estimated intercept accounts for a potential bias in the level of the forecasts. Because most variables are on average underestimated by the professional forecasters, we estimate in most cases a negative intercept. The estimated weights of the components do not change much when we do not include an intercept; the estimated weights for the trend and the cycle are close to one and the weights for the irregular component are similar as before (last four columns of Table 2.6). Moreover, unreported results show that fixing the coefficients of the trend and cycle components to one, barely changes the results with respect to the estimated weights of the irregular components.

Table 2.7 shows the results based on the estimated components in the state space model. We find almost the same results. The estimated weights for the trend and cycle components

are again close to one. However, it is remarkable that in case of the state space analysis all estimated weights for the irregular components are negative and in about half of the cases even significantly different from zero. The Wald test on the null hypothesis that the professional forecasters perfectly predict is again rejected for all variables with $p$-values equal to 0.000. Some of the estimated weights for the trend and cycle components differ significantly from one, for example GDP deflator and housing starts.

One could argue that the results in Tables 2.6 and 2.7 can be different before the Philadelphia Fed took over the survey compared to the period thereafter. However, including a dummy for the period after the take-over is almost never significant on a five percent level and does not significantly change the estimated coefficients of the components, and is therefore omitted from the reported forecast regressions. We also account for the varying calender dates for the survey deadline as well for the release dates of the survey results. These dates are documented from the moment The Fed took over the survey. For this sample period we include a dummy indicating whether the amount of days between the last release and the next deadline is above or below the median. Again we do not find significant estimates and hence we decide to omit this dummy. To make sure that our results are robust against definition changes, we also perform the analysis of Table 2.6 on the first differences of the series, where the first differences are constructed using the vintages in the real-time dataset. Appendix 2.C shows the results. We find that not all the weights of the business-cycle are as close to one as we found for the level data, but the weights of the irregular components are again significantly different from one.

Where we reported the White standard errors and corresponding Wald statistics in case of the components estimated in the spectral analysis in Table 2.6, in Table 2.7 ordinary standard errors and Wald statistics are reported. These standard errors do not take into account that the regressors are estimates. Since we obtain an estimated covariance matrix of the estimated parameters in the state space framework, we can adjust the ordinary standard errors for the uncertainty in the regressors. Table 2.8 shows the effect of the uncertainty in the estimated components in the state space model on the results of the forecast regression by reporting the two-step standard errors and corresponding Wald statistics.

The second column of Table 2.8 shows that the standard errors of the intercepts are now even larger. However, the forecast bias for nominal GDP, industrial production index, and housing starts is still significantly different from zero. Where the weights for the trend and

**Table 2.7:** Forecast Regressions ($h = 1$) Based On State Space Model

| | Estimate (Std. error) | | | | Estimate (Std. error) | | |
|---|---|---|---|---|---|---|---|
| | intercept | trend | cycle | irreg. | trend | cycle | irreg. |
| NGDP | −1.242* | 1.001 | 1.063 | −0.596* | 1.000* | 1.061 | −0.587* |
| | (0.553) | (0.001) | (0.044) | (0.194) | (0.000) | (0.045) | (0.196) |
| | 5.049 | 3.794 | 2.009 | 67.910 | 10.242 | 1.861 | 65.503 |
| PGDP | −0.316 | 1.001 | 1.096* | −0.804* | 1.000 | 1.100* | −0.805* |
| | (0.387) | (0.001) | (0.042) | (0.171) | (0.000) | (0.042) | (0.171) |
| | 0.666 | 0.627 | 5.242 | 111.429 | 0.100 | 5.757 | 111.773 |
| UNEMP | 0.015 | 1.004 | 0.980 | −0.024* | 1.004* | 0.980 | −0.024* |
| | (2.082) | (0.011) | (0.011) | (0.190) | (0.001) | (0.011) | (0.189) |
| | 0.000 | 0.145 | 3.073 | 29.212 | 21.326 | 3.139 | 29.413 |
| INDPROD | −3.708* | 1.006* | 0.989 | −0.443* | 0.999 | 0.988 | −0.436* |
| | (1.689) | (0.003) | (0.020) | (0.229) | (0.000) | (0.021) | (0.231) |
| | 4.821 | 4.724 | 0.300 | 39.817 | 0.126 | 0.362 | 38.506 |
| HOUSING | 4.520* | 0.866* | 0.971 | −0.381* | 0.975* | 0.939* | −0.340* |
| | (1.240) | (0.032) | (0.020) | (0.136) | (0.011) | (0.018) | (0.141) |
| | 13.292 | 17.822 | 2.086 | 103.030 | 5.243 | 10.927 | 90.524 |

This table shows the parameter estimates in forecast regression (2.13) of the professional forecasts on the state space model decomposition, with and without intercept. Standard errors are reported in parentheses together with Wald test statistics on the null hypothesis that the coefficient is equal to the weight expected in a perfect forecast. An asterisk ($*$) denotes that the coefficient significantly differs from the weight expected in a perfect forecast at the five percent significance level.

cycle components of housing starts significantly differ from one in case of ordinary standard errors, they do not significantly differ from one when we do not include an intercept and account for uncertainty in the estimated components. The weights of the irregular components are still significantly different from one. It remains remarkable that all estimated weights for the irregular components are negative and that still some of these effects are significantly different from zero. The forecast regressions in the last four columns still have a few trend and cycle coefficients significantly different from one due to small standard errors, for example, for nominal GDP, GDP deflator, and unemployment. In general, the conclusions do not change much when we account for two-step uncertainty. The Survey of Professional Forecasters can predict one-step ahead almost all variation in the time series due to a trend and a business-cycle, but predict little of the variation caused by the irregular component.

### 2.4.3 Individual Forecasts Analysis

As discussed in Section 2.3, it is difficult to analyse the performance of individual forecasters in the survey, since they have limited histories of responses and forecasters may switch identification numbers. However, we can analyze the individual predictions by pooling them

**Table 2.8:** Forecast Regressions ($h = 1$) With Two-Step Standard Errors

| | Estimate (Std. error) | | | | Estimate (Std. error) | | |
|---|---|---|---|---|---|---|---|
| | intercept | trend | cycle | irreg. | trend | cycle | irreg. |
| NGDP | −1.242* | 1.001 | 1.063 | −0.596* | 1.000* | 1.061 | −0.587* |
| | (0.553) | (0.001) | (0.046) | (0.232) | (0.000) | (0.047) | (0.233) |
| | 5.048 | 3.793 | 1.858 | 47.390 | 10.241 | 1.725 | 46.237 |
| PGDP | −0.316 | 1.001 | 1.096* | −0.804* | 1.000 | 1.100* | −0.805* |
| | (0.387) | (0.001) | (0.045) | (0.192) | (0.000) | (0.044) | (0.192) |
| | 0.666 | 0.627 | 4.650 | 88.747 | 0.100 | 5.037 | 88.822 |
| UNEMP | 0.015 | 1.004 | 0.980 | −0.024* | 1.004* | 0.980 | −0.024* |
| | (2.098) | (0.012) | (0.012) | (0.212) | (0.001) | (0.011) | (0.211) |
| | 0.000 | 0.143 | 2.903 | 23.428 | 21.264 | 2.958 | 23.539 |
| INDPROD | −3.708* | 1.006* | 0.989 | −0.443* | 1.000 | 0.988 | −0.436* |
| | (1.689) | (0.003) | (0.02) | (0.261) | (0.000) | (0.021) | (0.263) |
| | 4.818 | 4.722 | 0.298 | 30.571 | 0.126 | 0.359 | 29.823 |
| HOUSING | 4.520* | 0.866* | 0.971 | −0.381* | 0.975 | 0.939 | −0.340* |
| | (1.719) | (0.045) | (0.044) | (0.146) | (0.013) | (0.044) | (0.152) |
| | 6.917 | 8.827 | 0.428 | 88.956 | 3.669 | 1.902 | 77.655 |

This table shows the parameter estimates in forecast regression (2.13) of the professional forecasts on the state space model decomposition, with and without intercept. Standard errors and Wald test statistics account for two-step uncertainty and are computed based on the Murphy and Topel (2002) procedure. Standard errors are reported in parentheses together with Wald test statistics on the null hypothesis that the coefficient is equal to the weight expected in a perfect forecast. An asterisk ($*$) denotes that the coefficient significantly differs from the weight expected in a perfect forecast at the five percent significance level.

all in one forecast regression. Table 2.9 shows the results of a forecast regression on all individual forecasts, that is all forecasts over the sample period without averaging over the forecasts from the different panelists in each time period.

We find that the weights corresponding to the trend and the cycle are also close to one when we consider all individual forecasts, instead of the mean of the survey. The estimated parameter of the irregular component is in most cases closer to zero than to one. Since the regressions include a large number of observations (5784) the standard errors become small and almost every weight is significantly different from the weight in a perfect forecast on a five percent significance level.[3] In sum, the findings are in line with the results based on the mean of the Survey of Professional Forecasters.

[3]Unreported results show that a weighted regression where we weight with the number of forecasters to account for time variation in the number of forecasters produces similar results. Results can be obtained from the authors upon request.

**Table 2.9:** Forecast Regressions ($h = 1$) Based On Individual Forecasts

| | Estimate (Std. error) | | | | Estimate (Std. error) | | |
|---|---|---|---|---|---|---|---|
| | intercept | trend | cycle | irreg. | trend | cycle | irreg. |
| NGDP | −1.213* (0.113) | 1.001* (0.000) | 0.939* (0.008) | 0.253* (0.027) | 1.000* (0.000) | 0.949* (0.008) | 0.248* (0.028) |
| PGDP | −21.752* (0.136) | 1.042* (0.000) | 0.745* (0.010) | 0.055* (0.042) | 1.005* (0.000) | 1.135* (0.027) | −0.175* (0.107) |
| UNEMP | 1.529* (0.464) | 0.995 (0.003) | 0.952* (0.004) | 0.628* (0.026) | 1.004* (0.000) | 0.947* (0.004) | 0.635* (0.025) |
| INDPROD | −28.892* (0.371) | 1.054* (0.001) | 0.929* (0.006) | 0.448* (0.033) | 1.003* (0.000) | 0.944* (0.009) | 0.426* (0.039) |
| HOUSING | 2.298* (0.240) | 0.922* (0.006) | 0.888* (0.009) | 0.257* (0.027) | 0.970* (0.002) | 0.855* (0.008) | 0.268* (0.026) |

This table shows the parameter estimates in forecast regression (2.13), of the professional forecasts on the low-pass filter decomposition, with and without intercept. The regressions include all 5784 individual forecasts over the sample period without averaging over the forecasts from the different panelists in each time period. For additional information, see the note following Table 2.6.

### 2.4.4 Multi-step-ahead Forecasts

So far, our results are based on one-step-ahead predictions of the Survey of Professional Forecasters. To examine whether our findings also hold for multi-step-ahead forecasts, we perform the forecast regressions for different forecast horizons. The Survey of Professional Forecasters provides forecasts up to five quarters ahead.

Table 2.10 shows the results of the forecast regressions for $h = 5$ based on spectral analysis. Appendix 2.D shows the results for $h = 2, \ldots, 4$. We find that for all forecast horizons, the trend component receives a weight close to one and the weights corresponding to the irregular component are closer to zero than to one. The parameter estimates corresponding to the cycle decrease with the forecast horizon, and the forecast bias increases in the forecast horizon. In sum, we find that the professional forecasters are able to predict the trend over a longer horizon, but the forecasters are less able to produce unbiased forecasts and capture variation in the business-cycle when the forecast horizon increases.

## 2.5 Further Results

In this section we perform some extra analyses to shed light on our results and provide more insight on the value of the professional forecasts. First, we assess the robustness of the fixed variance of the irregular component in the state space framework against a range of values.

**Table 2.10:** Forecast Regressions Based On Spectral Analysis for $h = 5$

| | Estimate (Std. error) | | | | Estimate (Std. error) | | |
|---|---|---|---|---|---|---|---|
| | intercept | trend | cycle | irreg. | trend | cycle | irreg. |
| NGDP | −3.338 | 1.004 | −0.040* | −0.118* | 1.000 | −0.061* | −0.098* |
| | (3.144) | (0.004) | (0.142) | (0.451) | (0.000) | (0.135) | (0.462) |
| | 1.127 | 1.203 | 54.006 | 6.133 | 0.404 | 62.255 | 5.650 |
| PGDP | 3.543 | 0.995 | 0.742 | −0.166* | 1.000* | 0.805 | −0.201* |
| | (3.752) | (0.006) | (0.162) | (0.497) | (0.000) | (0.143) | (0.493) |
| | 0.892 | 0.767 | 2.543 | 5.500 | 6.259 | 1.852 | 5.946 |
| UNEMP | 9.921 | 0.945 | 0.139* | −0.501* | 0.999 | 0.103* | −0.446* |
| | (9.295) | (0.052) | (0.115) | (0.436) | (0.005) | (0.108) | (0.436) |
| | 1.139 | 1.110 | 56.251 | 11.855 | 0.018 | 68.515 | 10.978 |
| INDPROD | −10.240 | 1.019 | −0.119* | 0.089* | 1.002* | −0.117* | 0.087* |
| | (7.682) | (0.013) | (0.098) | (0.393) | (0.001) | (0.101) | (0.392) |
| | 1.777 | 2.104 | 129.042 | 5.376 | 12.459 | 123.125 | 5.435 |
| HOUSING | 20.545* | 0.553* | 0.076* | −0.012* | 1.000 | −0.406* | 0.107* |
| | (2.756) | (0.065) | (0.115) | (0.195) | (0.023) | (0.105) | (0.308) |
| | 55.579 | 46.629 | 64.149 | 26.838 | 0.000 | 178.401 | 8.389 |

This table shows the parameter estimates in forecast regression (2.13) with $h = 5$, of the professional forecasts on the low-pass filter decomposition, with and without intercept. Due to missing values, the estimation sample starts at 1974Q4. For additional information, see the note following Table 2.6.

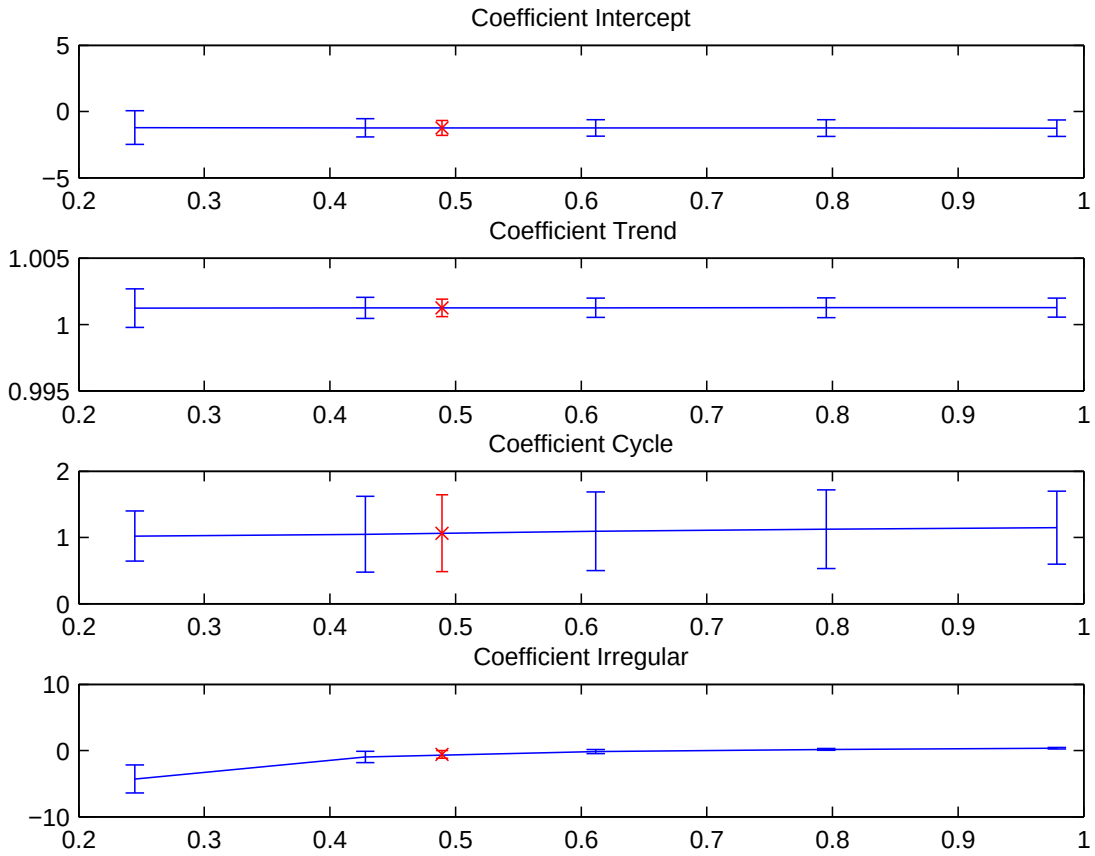
Next, we compare the forecasts of a basic time series model with the professional forecasts with respect to their ability to forecast the irregular component and with respect to accuracy. Finally, we examine the forecast regression in Section 2.4.2 with lagged trend, cycle and irregular components.

### 2.5.1 Sensitivity to Fixed Variance

To estimate the components in the state space framework, the variance of the irregular component is fixed to the value of the variance of the estimated irregular component in the low-pass filter. To assess how the forecast regression results are affected by this restriction, we perform a sensitivity analysis on the value of the variance of the irregular component. Figure 2.3 shows the sensitivity of the estimated coefficients in the forecast regression of nominal GDP based on the estimated components in the state space model. Figure 2.11 up to Figure 2.14 in Appendix 2.E show the sensitivities for the other time series.

The figure shows the values of the estimated coefficients with error bands of one standard error, for different values of the standard deviation of the estimated irregular component. The asterisks show the estimated coefficients at the value of the standard deviation of the estimated irregular component in the low-pass filter. The coefficients of the intercept, trend,

**Figure 2.3:** Sensitivity Analysis Fixed Variance Irregular Component




Sensitivity of the estimated coefficients in the forecast regression of nominal GDP to the standard deviation of the estimated irregular component in the state space framework. The (blue) lines show the value of the estimated coefficients with error bands of one standard error, for different values of the standard deviation of the estimated irregular component. The error bands are constructed with two-step standard errors. The (red) asterisks show the estimated coefficients at the value of the standard deviation of the estimated irregular component in the low-pass filter.

and business-cycle show hardly any differences over the interval. The coefficient of the irregular component seems to deviate more from the weight expected in a perfect forecast when the standard deviation of the estimated irregular component decreases. So the choice to fix the variance of the irregular component is not likely to influence the results found in the forecast regressions.

### 2.5.2 Model-based Forecast Decomposition

Based on the forecast regressions we find that the mean of the Survey of Professional Forecasters only explains little of the time series variation due to the irregular component. This is surprising when we presume that professional forecasters may adapt faster and be more flex-

**Table 2.11:** $AR(p)$ Model Forecast Regressions ($h = 1$) for Nominal GDP

| | Estimate (Std. error) | | | | Estimate (Std. error) | | |
|---|---|---|---|---|---|---|---|
| | intercept | trend | cycle | irreg. | trend | cycle | irreg. |
| NGDP | −2.154* | 1.002* | 0.971 | 0.010* | 1.000 | 0.980 | 0.009* |
| | (0.783) | (0.001) | (0.051) | (0.163) | (0.000) | (0.054) | (0.172) |
| | 7.572 | 7.565 | 0.321 | 36.643 | 0.497 | 0.138 | 33.031 |
| PGDP | −1.222 | 1.002 | 0.983 | −0.117* | 1.000 | 0.996 | −0.122* |
| | (0.778) | (0.001) | (0.061) | (0.315) | (0.000) | (0.060) | (0.321) |
| | 2.466 | 2.449 | 0.078 | 12.618 | 0.054 | 0.003 | 12.172 |
| UNEMP | 2.439 | 0.986 | 1.029 | −0.026* | 0.999 | 1.023 | −0.015* |
| | (4.348) | (0.024) | (0.047) | (0.351) | (0.002) | (0.046) | (0.349) |
| | 0.315 | 0.351 | 0.386 | 8.520 | 0.178 | 0.244 | 8.462 |
| INDPROD | −1.322 | 1.002 | 1.055 | 0.309* | 1.000 | 1.055 | 0.309* |
| | (3.039) | (0.006) | (0.054) | (0.251) | (0.000) | (0.054) | (0.250) |
| | 0.189 | 0.187 | 1.042 | 7.582 | 0.036 | 1.045 | 7.638 |
| HOUSING | 1.834 | 0.962 | 1.032 | 0.024* | 1.003 | 0.991 | 0.034* |
| | (1.370) | (0.031) | (0.062) | (0.170) | (0.017) | (0.060) | (0.173) |
| | 1.792 | 1.455 | 0.260 | 33.081 | 0.029 | 0.022 | 31.051 |

This table shows the parameter estimates in forecast regression (2.13) of the $AR(p)$ model forecasts on the low-pass filter decomposition, with and without intercept. White standard errors are reported in parentheses together with Wald test statistics on the null hypothesis that the coefficient is equal to the weight expected in a perfect forecast. An asterisk ($*$) denotes that the coefficient significantly differs from the weight expected in a perfect forecast at the five percent significance level.

ible than pure model-based prediction methods. However, we do not expect an econometric model to capture the irregular component. To investigate this conjecture, we regress model forecasts on the estimated components of the historical time series.

We generate forecasts with an autoregressive model of order $p$, $AR(p)$, for the first difference of the log series estimated on a moving window of ten years of quarterly observations. The order $p$ is selected for each forecasting period by means of the Schwartz information criterion on the moving window. The model is estimated using the latest vintage of real-time historical data available at the moment of forecasting using a similar approach as in the previous section.

Table 2.11 shows the forecast regression results of the one-step-ahead predictions in the sample from the last quarter of 1971 to the first quarter of 2013. The overall picture resembles the results in Section 2.4.2. Both the estimated weights of the components estimated in the spectral analysis as the estimated weights of the components estimated in the state space model show that the model-based predictions can only explain the trend and cycle components. The forecasts do not contain any information about the irregular component and the weight is negative in case one opts for a state space model approach to decompose the time series.

**Table 2.12:** Mean Squared Prediction Errors

| | Last 40 quarters | | | Last 20 quarters | | |
|---|---|---|---|---|---|---|
| | SPF | SSM | DM | SPF | SSM | DM |
| NGDP | 0.749 | 0.884 | −3.618* | 0.830 | 0.844 | −0.194 |
| PGDP | 0.554 | 0.562 | −0.731 | 0.573 | 0.584 | −0.503 |
| UNEMP | 2.299 | 4.462 | −5.739* | 1.972 | 4.508 | −2.387* |
| INDPROD | 1.136 | 1.453 | −2.521* | 0.816 | 1.171 | −1.216 |
| HOUSING | 6.516 | 8.673 | −3.706* | 7.083 | 9.082 | −1.246 |

This table shows the mean squared prediction error of the one-step ahead predictions of the Survey of Professional Forecasters (SPF) and the state space model (SSM), together with the Diebold and Mariano (1995) test statistic. We have real-time data from 1947Q1 to 2014Q1 from which we use an expanding window in the state space model to predict the last 40 quarters. Mean squared prediction errors are reported over all predictions and the predictions for the last 20 quarters.

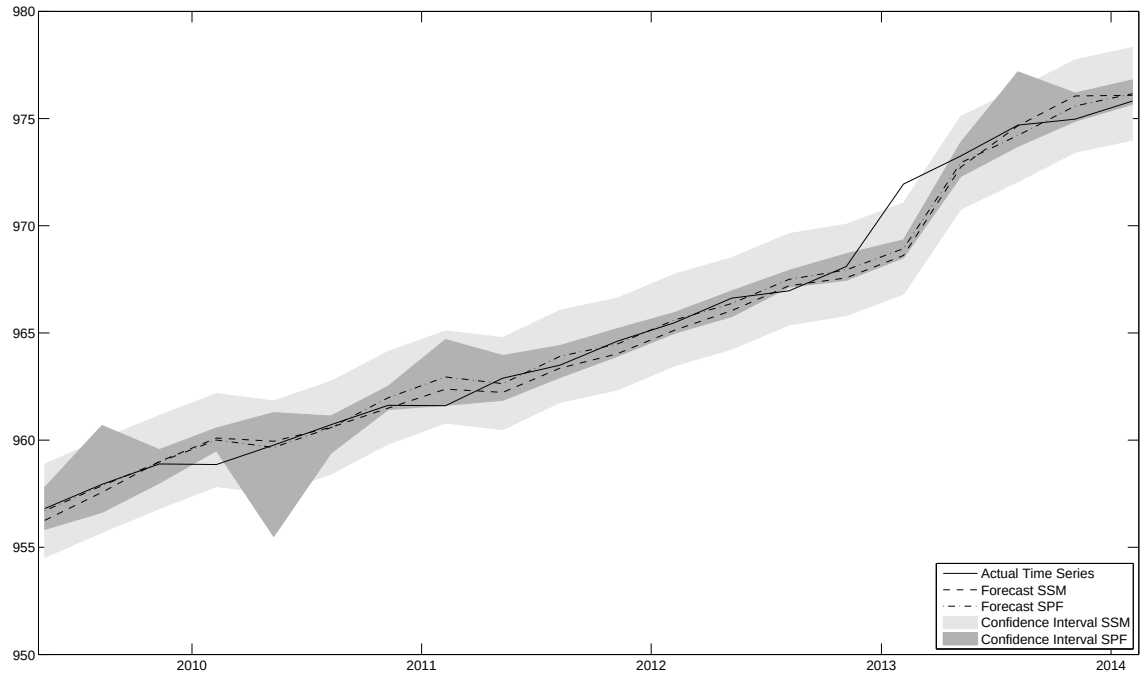
### 2.5.3 Forecast Accuracy

Our previous results show that the professional forecasters predict little of the irregular component. To investigate the value-added of professional forecasts, we compare them to simple model-based predictions. We obtain these predictions from the Kalman filter in the state space model (2.6)–(2.10) in which we do not fix a signal-to-noise ratio. So the irregular component estimated by the state space model is allowed to go to zero.

We generate the one-step-ahead predictions in the sample from 1980Q4 up to 2014Q2 using an expanding window consisting of the the latest vintage of real-time historical data available at the moment of forecasting. The first estimation sample starts at 1947Q1. The data is provided in the Real-Time Data Set for Macroeconomists of the Federal Reserve Bank of Philadelphia. We account for changing base years in the GDP deflator and the industrial production index by scaling all data in the Real-Time Data Set and the Survey of Professional Forecasters by the value for 1980Q4 in the latest vintage available at the moment of forecasting.

Table 2.12 shows the mean squared prediction errors for the forecasts of the state space model and the Survey of Professional Forecasters. Except for PGDP, the state space model is significantly outperformed on a five percent significance level by the professional forecasters based on all predictions. Although the professional forecasters cannot capture all variation in the irregular component, they probably do a better job in forecasting the trend and the business-cycle than the state space model over the whole time period. When we only

**Figure 2.4:** Model-Based and Professional Forecasts NGDP



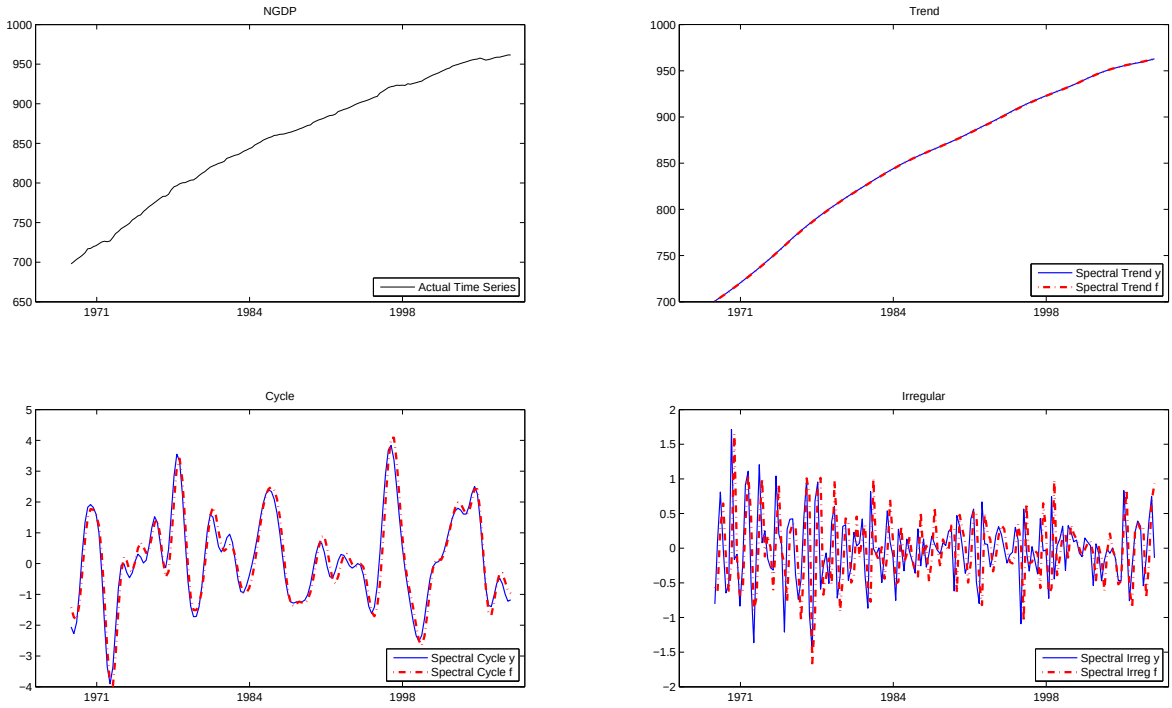

Nominal GDP predictions of the state space model (dashed line) and the Survey of Professional Forecasters (dashed dotted line) together with the actual time series (solid line). The corresponding gray surfaces represent the constructed confidence intervals of the predictions. The jump in 2013 can be explained by a change in the Bureau of Economic Analysis (BEA) definition of GDP.

consider the predictions for the last 20 quarters, the state space model is only significantly outperformed for unemployment but again the professional forecasters are more accurate in terms of MSPE.

Figure 2.4 shows the nominal GDP forecasts,[4] the confidence intervals and the actual historical time series for the evaluation period including the last five years of the sample. The confidence interval for the Survey of Professional Forecasters is constructed by the lowest and highest individual forecast and the state space prediction comes along with a covariance matrix from which we retrieve two times the standard deviation. The two predictions are very close to each other and follow an almost identical pattern. Where the constructed confidence interval of the professional forecasts seems narrower over the whole evaluation period, it has some outliers, while the confidence interval of the state space predictions is quite stable. Overall, the structural time series model produces almost the same predictions as the Survey of Professional Forecasters.

[4]Due to a change in the Bureau of Economic Analysis (BEA) definition of GDP, Figure 2.4 shows a jump in 2013. However, our results are robust to these kind of changes, as the analysis on the first differences of the series in Table 2.6 shows. We thank an anonymous referee for pointing this out.

**Figure 2.5:** Decomposition of Nominal GDP in Spectral Analysis



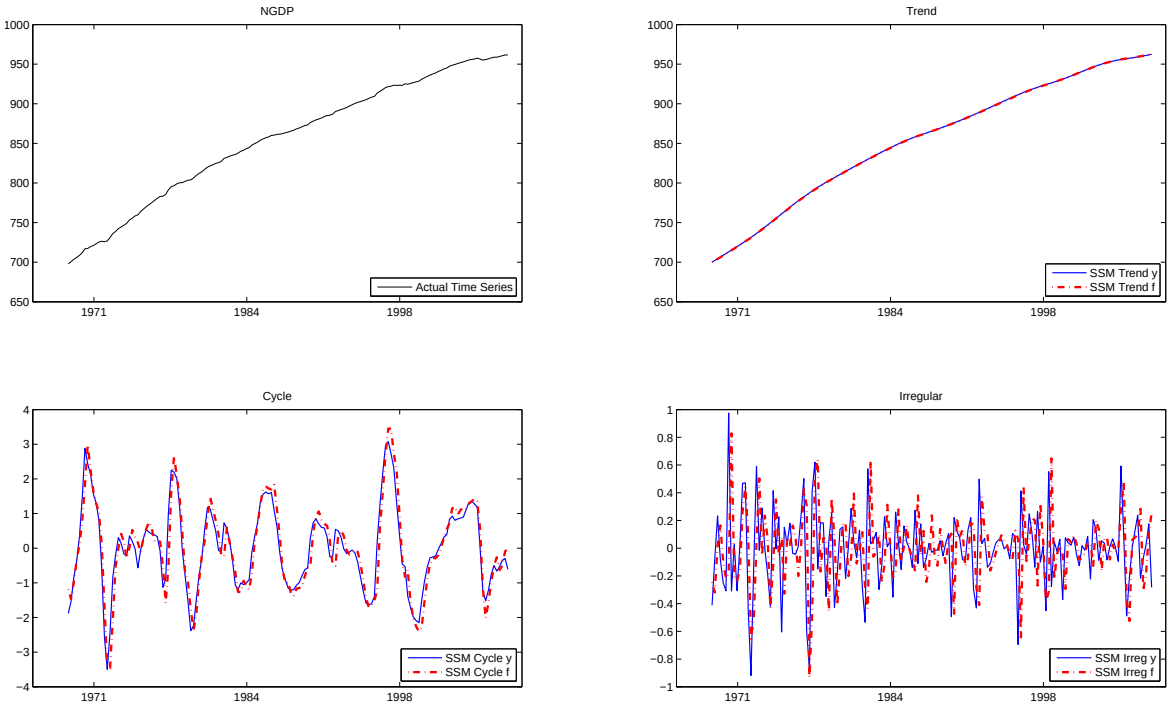

The historical time series and the mean of the forecasts of the Survey of Professional Forecasters for Nominal GDP decomposed in a trend, a cycle, and an irregular component by the low-pass filters. The first window shows one hundred times the logarithm of the actual values in the historical time series and the other windows show the estimated components of the actual historical time series by a blue solid line and the estimated components of the mean of the forecasts of the Survey of Professional Forecasters by a red dashed dotted line.

### 2.5.4 Forecast Regression with Lagged Components

To shed light on the information in the professional forecasts, we now also consider the time series decomposition of the mean of the Professional Forecasters. Figure 2.5 and Figure 2.6 show these decompositions together with the decomposition of the historical time series based on a spectral analysis and a state space model, respectively. In both figures, the business-cycle and the irregular component estimated from the forecasts seem to lag behind these components estimated from the historical time series.

Since the decompositions of the mean of the forecasts of the Survey of Professional forecasters suggest that the forecasts are biased towards lagged values of nominal GDP, we regress the professional forecasts on the lagged values of the components estimated from the historical time series. Table 2.13 shows the results for all series. Due to small standard errors the weights of the lagged estimated trend and cycle sometimes differ significantly from one, but the weights of the irregular component do not significantly differ from one, except for

**Figure 2.6:** Decomposition of Nominal GDP in State Space Model




The historical time series and the mean of the forecasts of the Survey of Professional Forecasters for Nominal GDP decomposed in a trend, a cycle, and an irregular component by the state space model. The first window shows one hundred times the logarithm of the actual values in the historical time series and the other windows show the estimated components of the actual historical time series by a blue solid line and the estimated components of the mean of the forecasts of the Survey of Professional Forecasters by a red dashed dotted line.

housing starts. This suggests that the professional forecasters explain the value of the series the current period, which is already published, instead of explaining irregular events in the future.

## 2.6 Conclusion

In this chapter we have examined what professional forecasters actually explain. We use a spectral analysis and a state space model to decompose economic time series into three components; a trend, a business-cycle, and an irregular component. Thereafter we examine which components are explained by the Survey of Professional Forecasters in a regression of the mean forecasts on the estimated components of the actual historical time series. We run these regressions based on the components estimated by the low-pass filters in the spectral analysis and the components estimated in a state space model. Both approaches lead to approximately the same results. For most time series we cannot reject that the mean of the

**Table 2.13:** Forecast Regressions ($h = 1$) on Lagged Estimated Components

| | Estimate (Std. error) | | | | Estimate (Std. error) | | |
|---|---|---|---|---|---|---|---|
| | intercept | trend | cycle | irreg. | trend | cycle | irreg. |
| NGDP | 6.610* | 0.994* | 0.989 | 0.951 | 1.002* | 0.958 | 0.953 |
| | (0.495) | (0.001) | (0.025) | (0.094) | (0.000) | (0.041) | (0.155) |
| | 178.283 | 110.964 | 0.207 | 0.272 | 610.593 | 1.081 | 0.092 |
| PGDP | −9.641* | 1.022* | 0.998 | 1.002 | 1.002* | 1.008 | 0.999 |
| | (1.029) | (0.002) | (0.006) | (0.010) | (0.000) | (0.004) | (0.007) |
| | 87.871 | 100.860 | 0.136 | 0.054 | 487.419 | 3.751 | 0.020 |
| UNEMP | −3.270 | 1.024 | 0.896* | 0.541 | 1.006* | 0.905* | 0.528 |
| | (3.554) | (0.020) | (0.031) | (0.247) | (0.002) | (0.029) | (0.247) |
| | 0.847 | 1.459 | 11.560 | 3.462 | 12.175 | 10.747 | 3.657 |
| INDPROD | 0.245 | 1.001 | 0.974* | 0.943 | 1.001* | 0.974* | 0.943 |
| | (4.915) | (0.010) | (0.013) | (0.031) | (0.000) | (0.013) | (0.031) |
| | 0.002 | 0.004 | 4.268 | 3.425 | 38.798 | 4.319 | 3.385 |
| HOUSING | −0.123 | 0.971 | 0.866* | 0.558* | 0.968* | 0.868* | 0.557* |
| | (0.863) | (0.021) | (0.033) | (0.077) | (0.009) | (0.032) | (0.077) |
| | 0.020 | 1.986 | 16.395 | 32.577 | 13.202 | 16.674 | 33.377 |

This table shows the parameter estimates in forecast regression (2.13) of the professional forecasts on the lagged values of the low-pass filter decomposition, with and without intercept. White standard errors are reported in parentheses together with Wald test statistics on the null hypothesis that the coefficient is equal to the weight expected in a perfect forecast. An asterisk ($*$) denotes that the coefficient significantly differs from the weight expected in a perfect forecast at the five percent significance level.

professional forecasts predicts the variation in the trend and the business-cycle, but there is little or no predictive power for the variation in the irregular component. A simple state space model, which is commonly used to estimate trends and cycles in time series, produces almost the same predictions.

The results suggest that both econometric models and the mean of the professional forecasts contain little information about the variation in the irregular component. This result is not surprising when professional forecasters also use model-based techniques to construct their predictions and the irregular component is characterized by weak persistence. Both econometric models and professional forecasters perform well in capturing the trend and the business-cycle. The fact that in some cases the professional forecasters also capture a small amount of the variation in the irregular components, may explain why some businesses and policymakers rely on professional forecasters.

Since the time series in the database of the Survey of Professional Forecasters are already seasonally adjusted, the time series decompositions are limited to a trend, cycle and irregular component. An interesting topic for future research is to analyze whether professional forecasters are able to predict seasonal variation by extending our analysis with a seasonal component and seasonally unadjusted data.

## 2.A Time Series Decompositions

In Subsection 2.4.1 we show nominal GDP decomposed in a trend, a cycle, and an irregular component by the low-pass filters and the state space model. Here we show the decompositions of the other variables; GDP deflator (PGDP), unemployment (UNEMP), industrial production index (INDPROD), and housing starts (HOUSING). Each figure corresponding to a variable consist of four windows. The first window shows the actual values in the historical time series and the other windows show the components estimated in the low-pass filters by a blue solid line and the components estimated in the state space model by a red dashed dotted line. All variables are log transformed and multiplied by hundred.

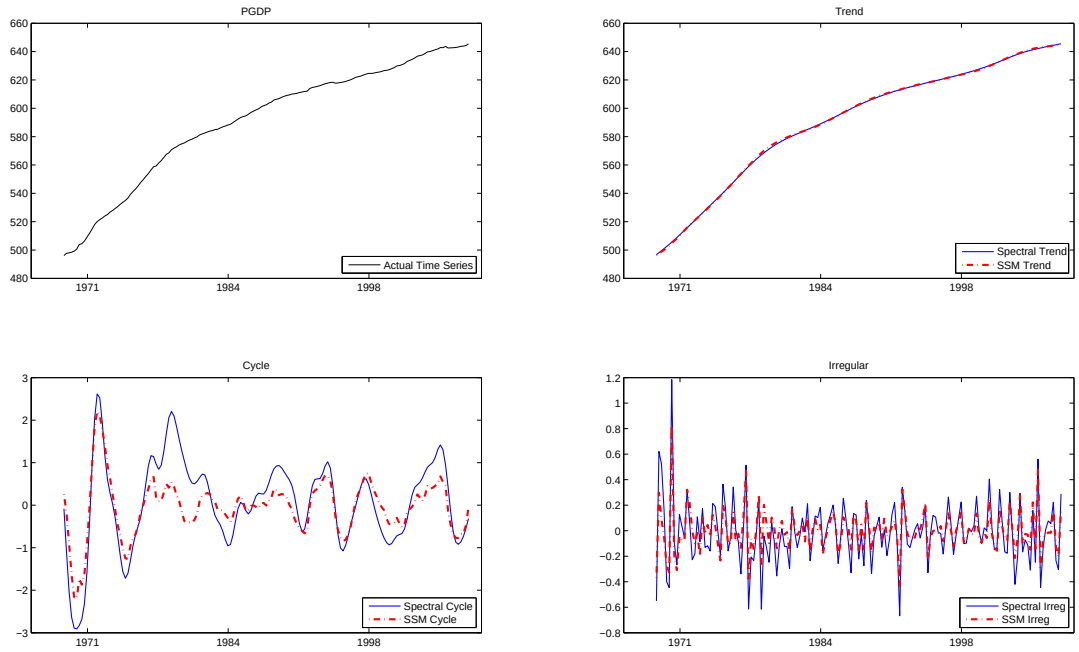



**Figure 2.7:** GDP deflator

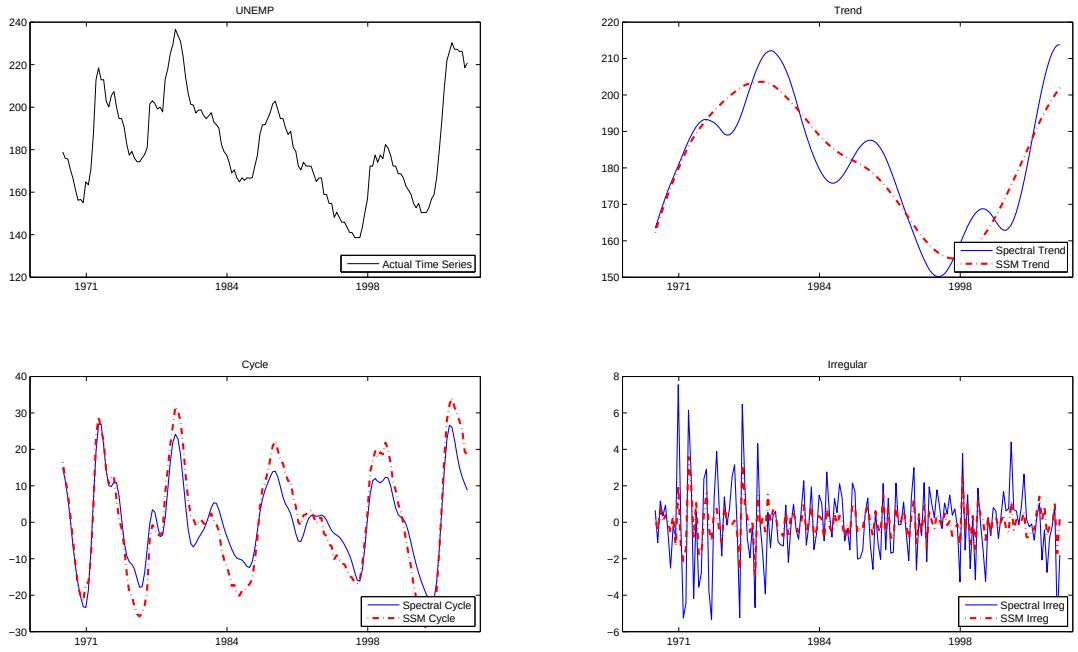



**Figure 2.8:** Unemployment

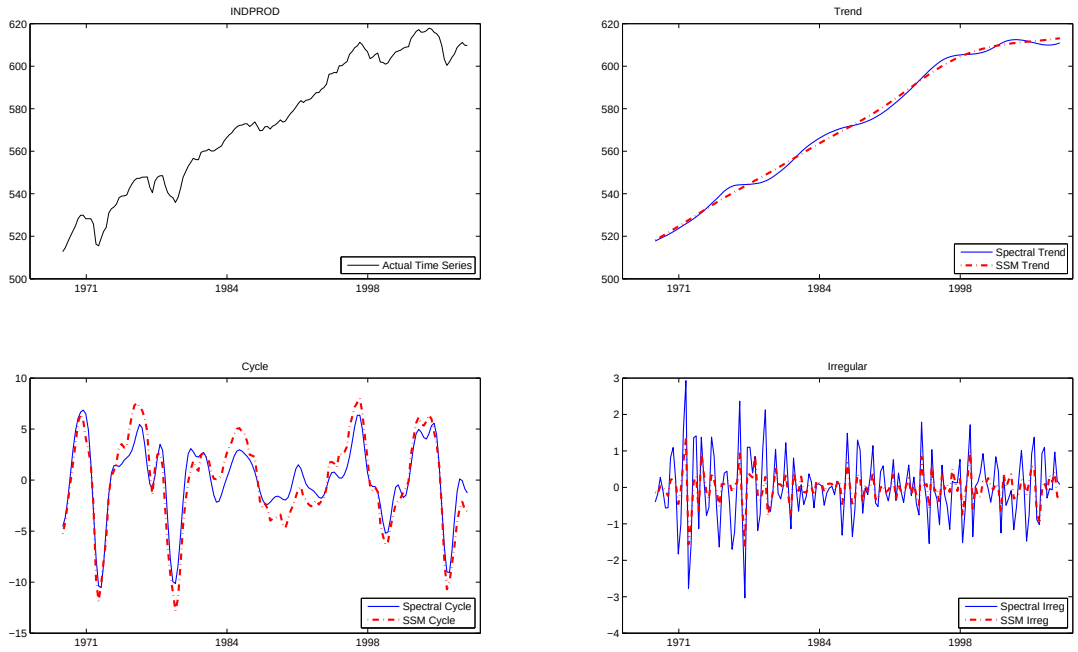



**Figure 2.9:** Industrial Production Index

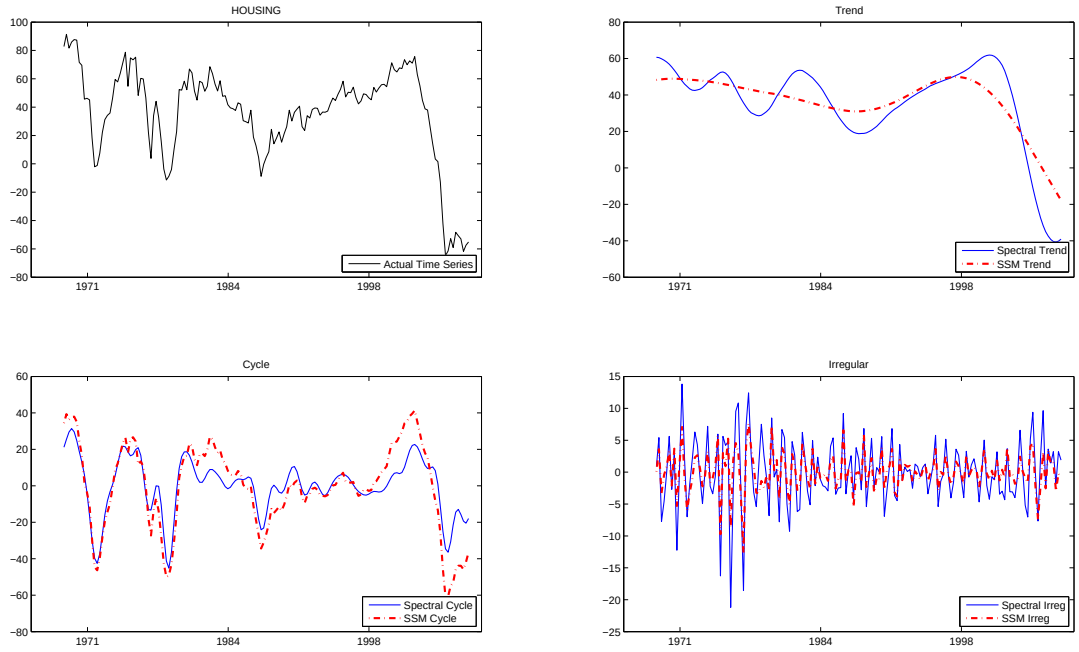



**Figure 2.10:** Housing Starts

# 2.B Alternative Frequency Filters

**Table 2.14:** Forecast Regressions ($h = 1$) Based On Alternative Frequency Filters

| | Estimate (Std. error) | | | | Estimate (Std. error) | | |
|---|---|---|---|---|---|---|---|
| | intercept | trend | cycle | irreg. | trend | cycle | irreg. |
| | Christiano-Fitzgerald filter | | | | | | |
| NGDP | −0.749 | 1.001 | 0.969 | 0.182* | 1.000* | 0.971 | 0.182* |
| | (0.482) | (0.001) | (0.036) | (0.138) | (0.000) | (0.037) | (0.140) |
| | 2.420 | 1.514 | 0.715 | 35.188 | 11.163 | 0.639 | 34.043 |
| PGDP | −0.729 | 1.001 | 1.003 | −0.139* | 1.000 | 1.001 | −0.138* |
| | (0.376) | (0.001) | (0.040) | (0.158) | (0.000) | (0.042) | (0.168) |
| | 3.759 | 3.646 | 0.004 | 51.905 | 1.070 | 0.000 | 46.166 |
| UNEMP | 1.280 | 0.997 | 0.959* | 0.508* | 1.004* | 0.958* | 0.508* |
| | (1.404) | (0.008) | (0.014) | (0.099) | (0.001) | (0.014) | (0.100) |
| | 0.832 | 0.128 | 9.073 | 24.839 | 20.765 | 9.328 | 24.370 |
| INDPROD | −1.451 | 1.003 | 0.953 | 0.425* | 1.000 | 0.954 | 0.426* |
| | (1.493) | (0.003) | (0.024) | (0.162) | (0.000) | (0.024) | (0.161) |
| | 0.944 | 0.935 | 3.687 | 12.654 | 0.016 | 3.594 | 12.777 |
| HOUSING | 1.232 | 0.944* | 0.908* | 0.230* | 0.964* | 0.908* | 0.230* |
| | (0.661) | (0.014) | (0.027) | (0.115) | (0.008) | (0.027) | (0.116) |
| | 3.471 | 16.455 | 11.897 | 44.493 | 19.145 | 11.768 | 44.338 |
| | Butterworth filter | | | | | | |
| NGDP | −0.753 | 1.001 | 1.004 | 0.017* | 1.000* | 1.004 | 0.017* |
| | (0.475) | (0.001) | (0.039) | (0.170) | (0.000) | (0.040) | (0.172) |
| | 2.515 | 1.580 | 0.010 | 33.365 | 11.735 | 0.013 | 32.490 |
| PGDP | −0.716* | 1.001 | 1.038 | −0.320* | 1.000 | 1.039 | −0.320* |
| | (0.363) | (0.001) | (0.040) | (0.152) | (0.000) | (0.041) | (0.162) |
| | 3.881 | 3.762 | 0.892 | 75.130 | 1.164 | 0.882 | 66.659 |
| UNEMP | 0.678 | 1.001 | 0.961* | 0.457* | 1.004* | 0.959* | 0.458* |
| | (1.735) | (0.009) | (0.016) | (0.102) | (0.001) | (0.016) | (0.102) |
| | 0.153 | 0.004 | 5.672 | 28.376 | 21.873 | 6.988 | 27.953 |
| INDPROD | −1.476 | 1.003 | 0.959 | 0.330* | 1.000 | 0.959 | 0.329* |
| | (1.483) | (0.003) | (0.030) | (0.198) | (0.000) | (0.030) | (0.198) |
| | 0.991 | 0.983 | 1.945 | 11.433 | 0.016 | 1.806 | 11.496 |
| HOUSING | 1.168 | 0.946* | 0.928* | 0.095* | 0.967* | 0.916* | 0.099* |
| | (0.680) | (0.015) | (0.036) | (0.139) | (0.009) | (0.036) | (0.138) |
| | 2.953 | 12.278 | 4.001 | 42.503 | 14.606 | 5.458 | 42.474 |

This table shows the parameter estimates in forecast regression (2.13) of the professional forecasts on the Christiano-Fitzgerald filter decomposition and the Butterworth filter decomposition. For notes, see Table 2.6.

# 2.C Forecast Regressions First Differences

**Table 2.15:** Forecast Regressions ($h = 1$) Based On Spectral Analysis on First Differences

| | Estimate (Std. error) | | | Estimate (Std. error) | |
|---|---|---|---|---|---|
| | intercept | cycle | irreg. | cycle | irreg. |
| NGDP | 0.430* | 0.646* | 0.169* | 0.847* | 0.171* |
| | (0.071) | (0.043) | (0.056) | (0.019) | (0.063) |
| | 36.544 | 68.780 | 216.222 | 62.220 | 174.791 |
| PGDP | 0.356* | 0.619* | −0.002* | 0.848* | −0.002* |
| | (0.043) | (0.048) | (0.085) | (0.031) | (0.119) |
| | 68.330 | 63.377 | 138.426 | 24.196 | 71.160 |
| UNEMP | 0.852* | 0.761* | 0.518* | 0.770* | 0.517* |
| | (0.122) | (0.030) | (0.054) | (0.033) | (0.064) |
| | 48.381 | 63.980 | 80.745 | 47.402 | 56.536 |
| INDPROD | 0.247* | 0.531* | 0.221* | 0.577* | 0.222* |
| | (0.073) | (0.040) | (0.077) | (0.037) | (0.083) |
| | 11.443 | 135.813 | 102.068 | 129.573 | 87.411 |
| HOUSING | −0.953* | 0.444* | 0.361* | 0.468* | 0.362* |
| | (0.334) | (0.052) | (0.053) | (0.053) | (0.054) |
| | 8.152 | 112.131 | 143.583 | 102.285 | 140.718 |

This table shows the parameter estimates in forecast regression (2.13) with first differences of the series instead of levels, of the professional forecasts on the low-pass filter decomposition, with and without intercept. Since we take first differences, the trend is removed from the forecast regression. For additional information, see the note following Table 2.6.

# 2.D Multi-step-ahead Forecasts

**Table 2.16:** Forecast Regressions Based On Spectral Analysis for $h = 2$

| | Estimate (Std. error) | | | | Estimate (Std. error) | | |
|---|---|---|---|---|---|---|---|
| | intercept | trend | cycle | irreg. | trend | cycle | irreg. |
| NGDP | −2.317* | 1.002* | 0.820* | −0.362* | 1.000* | 0.827* | −0.374* |
| | (0.933) | (0.001) | (0.065) | (0.196) | (0.000) | (0.066) | (0.207) |
| | 6.168 | 4.999 | 7.640 | 48.545 | 9.255 | 6.857 | 44.203 |
| PGDP | −1.085 | 1.002 | 0.991 | −0.504* | 1.000 | 1.002 | −0.520* |
| | (0.941) | (0.002) | (0.059) | (0.227) | (0.000) | (0.062) | (0.236) |
| | 1.331 | 1.290 | 0.024 | 43.901 | 0.514 | 0.001 | 41.508 |
| UNEMP | 3.951 | 0.984 | 0.812* | −0.227* | 1.006* | 0.800* | −0.211* |
| | (3.758) | (0.021) | (0.035) | (0.222) | (0.002) | (0.033) | (0.221) |
| | 1.105 | 0.546 | 29.244 | 30.634 | 8.066 | 37.562 | 30.036 |
| INDPROD | −4.779 | 1.009 | 0.766* | −0.207* | 1.000 | 0.765* | −0.210* |
| | (3.456) | (0.006) | (0.053) | (0.246) | (0.000) | (0.053) | (0.240) |
| | 1.913 | 2.025 | 19.653 | 24.055 | 0.837 | 19.796 | 25.360 |
| HOUSING | 6.323* | 0.852* | 0.638* | −0.300* | 0.986 | 0.537* | −0.268* |
| | (1.247) | (0.031) | (0.060) | (0.129) | (0.015) | (0.056) | (0.128) |
| | 25.722 | 22.887 | 36.707 | 102.017 | 0.865 | 69.448 | 98.129 |

This table shows the parameter estimates in forecast regression (2.13) with $h = 2$, of the professional forecasts on the low-pass filter decomposition, with and without intercept. For additional information, see the note following Table 2.6.

**Table 2.17:** Forecast Regressions Based On Spectral Analysis for $h = 3$

| | Estimate (Std. error) | | | | Estimate (Std. error) | | |
|---|---|---|---|---|---|---|---|
| | intercept | trend | cycle | irreg. | trend | cycle | irreg. |
| NGDP | −3.557* | 1.004* | 0.589* | −0.396* | 1.000* | 0.594* | −0.413* |
| | (1.240) | (0.001) | (0.094) | (0.257) | (0.000) | (0.093) | (0.267) |
| | 8.222 | 7.104 | 19.319 | 29.536 | 4.436 | 18.917 | 27.960 |
| PGDP | −2.924 | 1.005 | 0.961 | −0.517* | 1.000 | 0.989 | −0.531* |
| | (1.524) | (0.003) | (0.095) | (0.318) | (0.000) | (0.102) | (0.346) |
| | 3.679 | 3.723 | 0.170 | 22.740 | 0.187 | 0.011 | 19.523 |
| UNEMP | 4.160 | 0.981 | 0.632* | −0.336* | 1.004 | 0.620* | −0.318* |
| | (5.931) | (0.033) | (0.060) | (0.323) | (0.003) | (0.058) | (0.322) |
| | 0.492 | 0.310 | 37.578 | 17.080 | 1.776 | 43.577 | 16.758 |
| INDPROD | −4.581 | 1.009 | 0.507* | −0.451* | 1.001 | 0.506* | −0.455* |
| | (4.816) | (0.008) | (0.074) | (0.299) | (0.000) | (0.074) | (0.293) |
| | 0.905 | 1.072 | 44.229 | 23.500 | 3.630 | 45.111 | 24.704 |
| HOUSING | 10.425* | 0.790* | 0.387* | −0.277* | 1.010 | 0.220* | −0.227* |
| | (1.739) | (0.042) | (0.076) | (0.156) | (0.018) | (0.069) | (0.166) |
| | 35.925 | 25.219 | 64.405 | 66.677 | 0.340 | 128.258 | 54.812 |

This table shows the parameter estimates in forecast regression (2.13) with $h = 3$, of the professional forecasts on the low-pass filter decomposition, with and without intercept. For additional information, see the note following Table 2.6.

**Table 2.18:** Forecast Regressions Based On Spectral Analysis for $h = 4$

| | Estimate (Std. error) | | | | Estimate (Std. error) | | |
|---|---|---|---|---|---|---|---|
| | intercept | trend | cycle | irreg. | trend | cycle | irreg. |
| NGDP | −4.839* | 1.005* | 0.283* | −0.134* | 1.000 | 0.285* | −0.138* |
| | (1.507) | (0.002) | (0.105) | (0.279) | (0.000) | (0.104) | (0.296) |
| | 10.313 | 9.321 | 46.936 | 16.468 | 1.505 | 46.996 | 14.797 |
| PGDP | −5.390* | 1.009* | 0.891 | −0.422* | 1.000 | 0.936 | −0.404* |
| | (2.160) | (0.004) | (0.134) | (0.386) | (0.000) | (0.150) | (0.428) |
| | 6.226 | 6.395 | 0.663 | 13.591 | 0.037 | 0.185 | 10.790 |
| UNEMP | 5.300 | 0.971 | 0.413* | −0.357* | 1.000 | 0.397* | −0.336* |
| | (7.917) | (0.044) | (0.085) | (0.382) | (0.004) | (0.080) | (0.381) |
| | 0.448 | 0.417 | 48.170 | 12.636 | 0.007 | 56.375 | 12.272 |
| INDPROD | −3.291 | 1.007 | 0.223* | −0.084* | 1.001* | 0.222* | −0.087* |
| | (5.810) | (0.010) | (0.085) | (0.320) | (0.000) | (0.084) | (0.316) |
| | 0.321 | 0.487 | 84.426 | 11.484 | 8.790 | 86.014 | 11.840 |
| HOUSING | 15.014* | 0.715* | 0.171* | −0.090* | 1.032 | −0.072* | −0.012* |
| | (2.164) | (0.051) | (0.081) | (0.161) | (0.020) | (0.073) | (0.202) |
| | 48.149 | 31.553 | 104.963 | 45.799 | 2.492 | 213.014 | 25.032 |

This table shows the parameter estimates in forecast regression (2.13) with $h = 4$, of the professional forecasts on the low-pass filter decomposition, with and without intercept. For additional information, see the note following Table 2.6.

# 2.E Sensitivity Analysis

In Subsection 2.4.2 we show the sensitivity of the estimated coefficients in the forecast regression of nominal GDP to the standard deviation of the variance of the estimated irregular component in the state space framework. Here we show the sensitivities of the coefficients of the components of the other variables; GDP deflator (PGDP), unemployment (UNEMP), industrial production index (INDPROD), and housing starts (HOUSING). Each figure corresponding to a variable consists of four windows; the coefficients of the intercept, trend, business-cycle, and irregular component. The blue lines indicate the value of the estimated coefficient with error bands of one standard error, for different values of the standard deviation of the variance of the estimated irregular component. The error bands are constructed with two-step standard errors. The red asterisks show the estimated coefficient at the value of the standard deviation of the variance of the estimated irregular component in the low-pass filter.

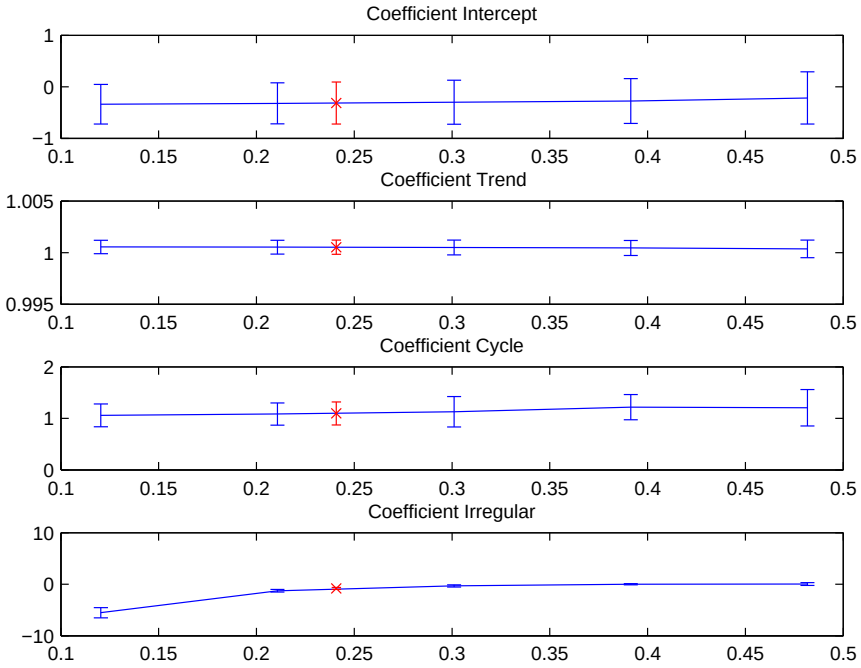



**Figure 2.11:** GDP deflator

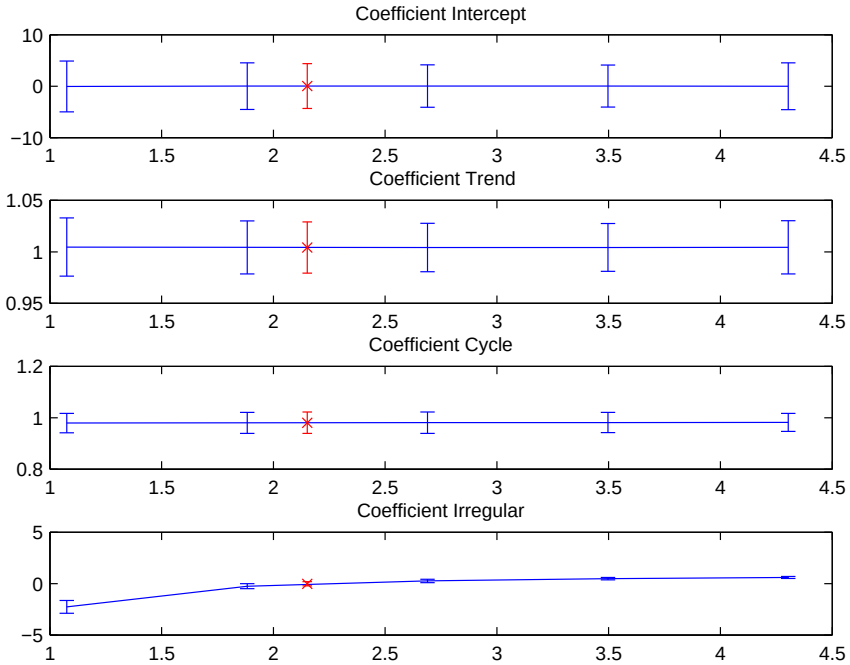



**Figure 2.12:** Unemployment

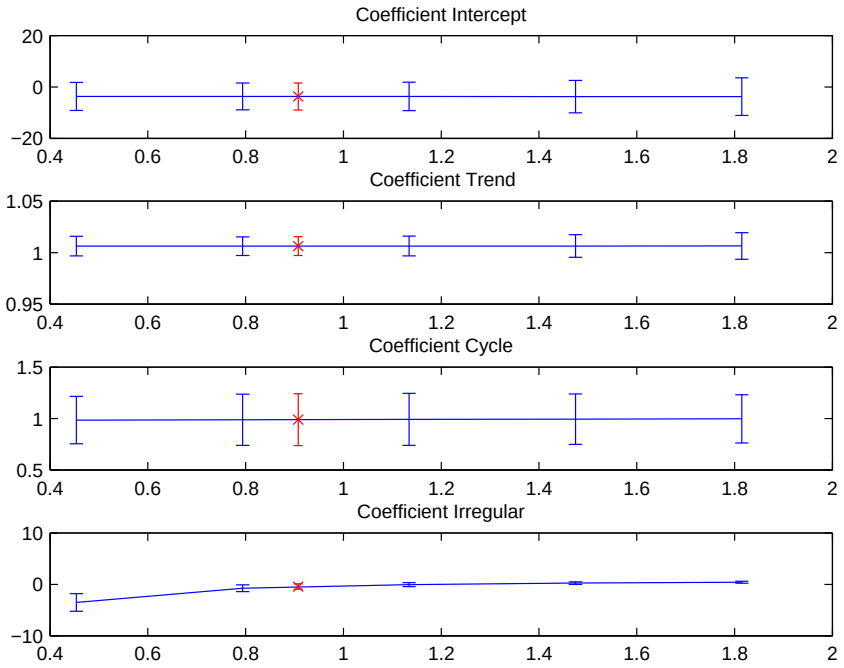



**Figure 2.13:** Industrial Production Index

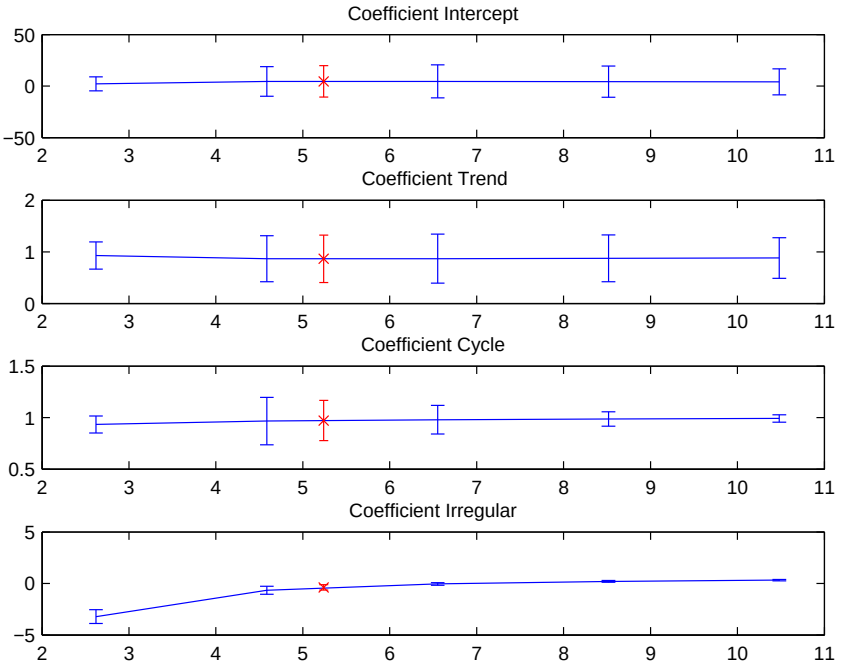



**Figure 2.14:** Housing Starts

# Chapter 3

# A Bayesian Infinite Hidden Markov Vector Autoregressive Model

*Joint work with Richard Paap and Michel van der Wel*

## 3.1 Introduction

Many researchers study estimation methods for time-varying parameters in vector autoregressive (VAR) models. The large sets of variables and hence parameters considered in these models compared to the number of available observations, increase the complexity of estimating time-varying parameters. Therefore, feasible estimation methods rely invariably on a set of model restrictions. To manage the number of time-varying parameter estimates, Cogley and Sargent (2005) impose the instantaneous relations among the VAR variables to be time-invariant. Chib et al. (2006) assume a factor structure for the covariance matrix. Primiceri (2005) imposes parameters to evolve smoothly over time, by modelling the evolution of the parameters in the coefficient and covariance matrices as random walks. Koop and Korobilis (2013) make use of forgetting factors to model time-variation in the parameters.

Parameter breaks in the models proposed by abovementioned papers are drawn from the same distribution. However, there is a wide variety in shapes and magnitudes of shocks to the economy, from abrupt shocks following rapid shifts in policy to smoother changes due to learning of economic agents. The model of Primiceri (2005), as a well-known example, imposes a break in each time period by modelling the law of motion by random walks. This

approach does not allow for the presence of occasional jumps in parameter values, and the continuous changes imply a linear increase in parameter uncertainty over time. Alternatively, Sims and Zha (2006) (among others) capture time variation with a finite number of regimes in a Markov switching framework. These discrete break models are able to model shifts in policy but cannot account for smoother changes. Moreover, the number of regimes needs to be arbitrarily fixed before parameter estimation, ignoring the uncertainty in the number of breaks. Univariate models that do account for different types of breaks, such as Pesaran et al. (2006) and Giordani et al. (2007), do not account for state-persistence, and cannot easily be scaled up to (large dimensional) multivariate models.

In this chapter we contribute to the literature of time-varying parameter vector autoregressive models by proposing a semi-parametric Bayesian model which accounts for heterogeneous parameters. Both the autoregressive parameters and the covariances of the innovations in the model are allowed to change over time without imposing any restrictions on the parameter space to make estimation feasible. We employ a hidden Markov chain in combination with a Dirichlet process to allow for time-varying parameters. The Dirichlet process mixture encourages parameters to cluster in regimes with similar values. This feature favours a parsimonious model, which is a huge advantage in modelling parameter heterogeneity in structural time series models which, due to large sets of variables, already suffer from the curse of dimensionality. Moreover, the Dirichlet process mixture allows parameters to be drawn from different distributions over time by a potentially infinite number of regimes, which makes it possible to model abrupt breaks together with smoother changes. In contrast to parameter estimation in models with a fixed finite number of switching regimes, we can estimate parameter values, state values, and the number of regimes along with their uncertainty, together in one round. Furthermore, the hidden Markov structure accounts for state-persistence, as is often encountered in macroeconomic data.

We illustrate the performance of the model in an extensive empirical application on a monetary VAR. We show the ability of the model to capture heterogeneity over time together with both abrupt shocks and smooth changes in a structural analysis. We especially find posterior evidence for time-varying volatility. In a real-time forecasting exercise we compare the forecast performance of a small VAR (3 variables) and a large VAR (10 variables) infinite hidden Markov model to time-varying parameter VAR benchmarks. The infinite hidden Markov model with a time-varying covariance matrix shows for most forecast horizons and

variables the best performance based on various evaluation measures. The relative performance of the infinite hidden Markov model benefits from the combination of both smooth and abrupt changes in parameter estimates, and its ability to switch to a stable regime from an explosive state, both in-sample and out-of-sample. Policy making based on models that only allow for smooth changes may be delayed in intervening after an abrupt shock. Discrete break models tend to be more prone to overreact to smooth changes or ignore small changes. For in-sample analysis, the infinite hidden Markov model allows for a more balanced deployment of policy instruments. Out-of-sample, it enables policy makers to anticipate more accurately to future changes in the economy.

Next to the empirical contribution, the technical contribution of our approach is threefold. First, we generalize the infinite hidden Markov model to a multivariate setting, and construct a novel alternative to existing restrictive time-varying parameter VARs. Note that Hou (2017) proposed a multivariate generalization of the infinite hidden Markov model in parallel. The infinite hidden Markov VAR model builds upon work of Jochmann (2015), Song (2014), and Bauwens et al. (2017). They bring a semi-parametric Bayesian model, developed by Fox et al. (2011) for speaker diarization, to the univariate time series literature. This results in autoregressive models with an infinite number of regimes. Bauwens et al. (2017) show the superiority in forecast performance on macroeconomic time series relative to univariate models with fixed parameters. We extend the result of Bauwens et al. (2017) to analyse not only the predictive performance compared to multivariate fixed parameter models, but also to often used time-varying parameter VAR models.

Second, we contribute to the growing literature on estimating large time-varying parameter VARs. Recent studies show that increasing the dimensions of the VAR model improves forecasting and structural analysis (Carriero et al., 2015b). Bańbura et al. (2010), Koop (2013), Carriero et al. (2015a), and Giannone et al. (2015) estimate large VARs but do not account for parameter change, while Cogley and Sargent (2005), Primiceri (2005), Chib et al. (2006), Clark (2012), and Clark and Ravazzolo (2015) find convincing evidence for time-varying parameters in small VAR models. Only a few papers try to bridge the gap between large and time-varying systems. Koop and Korobilis (2013) use a semi-Bayesian approach which imposes restrictions on the parameter space and is unsuitable for policy analysis as parameters are not estimated but selected from a small grid of different values using forgetting factors. Carriero et al. (2015b) model time-varying volatility by only a single common un-

observed factor, and for high-dimensional models they have to rely on a misspecified model. Since the infinite hidden Markov model estimates time-variation relatively parsimoniously, it can handle high-dimensional VAR systems without restricting the parameter space.

Third, the infinite hidden Markov model accounts for uncertainty in the underlying break processes but reduces the parameter uncertainty relative to other time-varying parameter models (Song, 2014). Traditional regime-switching models capture time-variation by a fixed finite number of regimes (Hamilton, 1989), which ignores the uncertainty around the number of regimes. Chopin and Pelgrin (2004) take this uncertainty into account by jointly estimating the parameters and the number of in-sample regimes. Moreover, traditional Markov switching models assume that future states are always equal to one of the estimated in-sample regimes, which results in inaccurate forecasts in case of new out-of-sample regimes. The infinite hidden Markov model estimates the number of regimes and allows for new regimes out-of-sample. Other researchers model heterogeneity over time by change-point models (Chib, 1998; Koop and Potter, 2007; Liu et al., 2017) or impose parameters to change each time period, for example Primiceri (2005). In these models, different states cannot reoccur over time, which inevitably results in a loss of estimation efficiency. The infinite hidden Markov model reduces the parameter uncertainty by estimating parameters on data over all similar states, also when observations are separated from each other by break points.

The remainder of this chapter is as follows. Section 3.2 discusses the model specification and explains parameter inference by Bayesian methods. Section 3.3 explains the empirical application of the methods on a monetary VAR. It introduces the data, discusses how we use the model for monetary policy analysis, and performs a forecasting exercise. We conclude with a discussion in Section 3.4.

## 3.2 Methods

This section discusses the specification and parameter estimation of the infinite hidden Markov VAR model. Section 3.2.1 introduces the baseline specification of the reduced form of a time-varying parameter VAR. From here, we explain how we capture the parameter heterogeneity over time by constructing regimes with homogeneous parameter values. The regimes and parameter values are estimated by Bayesian methods. In Section 3.2.2, we specify the

prior distributions and set up a Markov Chain Monte Carlo (MCMC) sampler. Moreover, we show how we can sample from the predictive density.

### 3.2.1 Model Specification

Consider the reduced form of a time-varying vector autoregressive model of order $l$

$$y_t = B_t x_t + \varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}(0, \Sigma_t), \qquad t = 1, \ldots, T, \tag{3.1}$$

where $y_t$ is a $p \times 1$ vector of observed endogenous time series, $B_t$ is a $p \times (k = 1 + pl)$ matrix with time-varying coefficients, and $\varepsilon_t$ are heteroskedastic independent disturbances with covariance matrix $\Sigma_t$. The $k \times 1$ vector $x_t = [1, y'_{t-1}, \ldots, y'_{t-l}]'$ includes an intercept and the endogenous variables up to lag $l$ as explanatory variables.

Both the coefficient matrix $B_t$ and the covariance matrix $\Sigma_t$ in (3.1) contain time-varying parameters. Equivalently, we can say that the parameters in $B_t$ and $\Sigma_t$ vary over an infinite number of regimes, where the number of regimes equals the number of time periods $T$ when each time period has a different parameter value. Within the regimes the parameters are assumed to be time-invariant but across regimes the parameters are allowed to be different. We can write (3.1) as

$$y_t = B_{s_t} x_t + \varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}(0, \Sigma_{s_t}), \tag{3.2}$$

where $s_{1:T} = \{s_1, \ldots, s_T\}$ takes integer values indicating the regime at time $t$. While there is strong evidence that the behavior of macroeconomic variables changes over time, it is implausible that the economy changes in each time period with probability one. Therefore, we can specify a potentially more parsimonious model by modelling the transition probability of moving from one state to another (Hamilton, 1989). Since we specify a transition probability distribution running over an infinite number of regimes, this does not prevent (3.2) to have different regimes for each time period.

We let the regime indicators $s_{1:T}$ follow a first-order Markov chain, where $\pi_{ij}$ denotes the transition probability of moving from state $i$ to state $j$ under the constraint that $\sum_{j=1}^{J} \pi_{ij} = 1$ for all $i$, where $J$ possibly goes to infinity. So each state $i$ has a state-specific transition distribution $\pi_i$ over $J$ states; $s_t \sim \pi_{s_{t-1}}$, where $\pi_i = (\pi_{i1}, \ldots, \pi_{iJ})$ . When the number of

states is possibly infinite, we potentially have infinitely many parameters in the state transition matrix $\pi = (\pi_1', \ldots, \pi_J')'$. Since estimating all these parameters is infeasible, we follow the framework of Teh et al. (2012), and implicitly integrate out the transition parameters by specifying the transition distributions $\pi_i$ as a Dirichlet process mixture model,

$$\pi_i | a, H \sim DP(a, H_i),$$

where $DP$ denotes a Dirichlet process distribution (Ferguson, 1973), the scalar $a = \alpha + \kappa$ is the concentration parameter, $H_i = \frac{\alpha\beta + \kappa\delta_i}{\alpha+\kappa}$ the base distribution, $\delta_i$ denotes a unit-mass measure concentrated at $i$, and $\alpha$ captures dispersion. The base distribution is constructed by the global transition distribution $\beta = (\beta_1, \ldots, \beta_J)$ and scaled by the persistence parameter $\kappa$ to account for state-dependence (Fox et al., 2011). When $\kappa = 0$, a standard Dirichlet process mixture model is recovered, that does not take state-persistence into account.

The Dirichlet process can be seen as a mixture over the transition probability distributions of $J$ current states, where the state-specific transition probability distribution runs over $J$ states in the next period, and $J$ possibly goes to infinity (Escobar and West, 1995). So, not only the global transition distribution runs over an infinite number of states, the Dirichlet process is also an infinite discrete distribution over state-specific transition distributions. Since the expectation of the Dirichlet process equals the base distribution, states tend to have similar transition distributions; $E[\pi_{ij}] = \frac{\alpha\beta_j + \kappa 1(i=j)}{\alpha+\kappa}$, where $1(A)$ denotes an indicator variable that equals one if event $A$ occurs and zero otherwise. An amount $\kappa > 0$ is added to the $i$th component of $\alpha\beta$, such that the expected probability of self-transition is increased by an amount proportional to $\kappa$. Moreover, element $j$ of $H_i$, that is $H_{ij}$, can be interpreted as prior mean for the transition probabilities into state $j$. The variance of the Dirichlet process equals $H_i(1 - H_i)/(a + 1)$, from which we infer that $\alpha$ indeed controls the dispersion around the prior mean across rows of the transition matrix.

Conditional on the regimes in the previous time periods, the regime indicator $s_t$ can be equal to the current regime of time period $t - 1$, an existing regime realized more than one period back in time, or switch to a new regime. In the latter case, new parameters values in the added row and column to the transition matrix are generated by a base distribution. The base distribution of the transition parameters is the scaled global transition distribution $\beta$,

defined as

$$\beta_j = \nu_j \prod_{l=1}^{j-1}(1-\nu_l), \qquad \nu_j|\gamma \sim Beta(1,\gamma), \quad j = 1,2,\ldots,$$

where $\beta = \{\beta_j\}_{j=1}^{\infty}$ is defined as a probability mass function on a countably infinite set. This is known as a stick-breaking construction, which can also be written as $\beta \sim Stick(\gamma)$. The expected number of represented hidden states is governed by $\gamma$, by controlling how concentrated the probability mass will be across the columns of the transition matrix. Switching to a new regime also implies that new values of the model parameters $\theta_{s_t} = \{B_{s_t}, \Sigma_{s_t}\}$ have to be generated. The base distribution of the model parameters is denoted by

$$\{B_{s_t}, \Sigma_{s_t}\} \sim H_\theta(\Theta),$$

where $\Theta$ is a set of hyperparameters in the base distribution.

We can summarize the complete model specification by the following equations,

$$y_t = B_{s_t} x_t + \varepsilon_t, \tag{3.3}$$

$$\varepsilon_t \sim \mathcal{N}(0, \Sigma_{s_t}), \tag{3.4}$$

$$s_t|s_{t-1}, \{\pi_i\}_{i=1}^{\infty} \sim \pi_{s_{t-1}}, \tag{3.5}$$

$$\pi_i|\alpha,\kappa,\beta \sim DP\left(\alpha+\kappa, \frac{\alpha\beta+\kappa\delta_i}{\alpha+\kappa}\right), \tag{3.6}$$

$$\beta_j = \nu_j \prod_{l=1}^{j-1}(1-\nu_l), \tag{3.7}$$

$$\nu_j|\gamma \sim Beta(1,\gamma), \quad j = 1,2,\ldots, \tag{3.8}$$

$$\{B_{s_t}, \Sigma_{s_t}\} \sim H_\theta(\Theta). \tag{3.9}$$

Equations (3.3) and (3.4) specify the reduced form of the time-varying vector autoregressive model, where the parameters $\theta_{s_t} = \{B_{s_t}, \Sigma_{s_t}\}$ vary over an infinite number of regimes. To retrieve the different regimes we use a hidden Markov chain in combination with a Dirichlet process mixture model. Equation (3.5) specifies the hidden Markov model by introducing a first-order Markov chain with transition probability matrix $\pi$. The transition probability distribution of $\pi_i$ is specified as a Dirichlet process mixture in (3.6)-(3.8). The $\kappa$ parameter

captures the persistence in macroeconomic data by controlling the probability that parameters remain constant between time periods. Equation (3.9) concludes, and provides the base distribution of the model parameters, $H_\theta$, parameterized by the hyperparameters $\Theta$.

For ease of notation we follow the Markov-switching literature and specify one regime switching process for all model parameters. To obtain a potential efficiency gain we can easily extend to a model with different regime-switching processes for the parameters in the coefficient matrix and the parameters in the covariance matrix, within the infinite hidden Markov framework. However, a regime switch in this model does not necessarily mean that all parameters change. For example, a regime switch can either be the result of a change in the parameters in the covariance (coefficient) matrix while the coefficient (covariance) parameters remain constant, or a change in all parameter values.

### 3.2.2 Bayesian Inference

To estimate the parameters $\theta_{s_t} = \{B_{s_t}, \Sigma_{s_t}\}$ we rely on the Markov Chain Monte Carlo (MCMC) algorithm for the infinite hidden Markov model derived by Fox et al. (2011). Bauwens et al. (2017) apply a variant of this sampler in a univariate econometric time series context.

Although there are sampling algorithms that can deal with an infinite number of regimes (which are also derived and discussed by Fox et al. (2011)), these algorithms suffer in general from slow mixing rates. Therefore, we opt for a sampler which truncates the number of possible states to a fixed degree $L$, the so called degree $L$ weak limit approximation (Ishwaran and Zarepour, 2002). This sampler fixes the number of states $J$ in the state transition matrix $\pi$ and the global transition distribution $\beta$ to $L$. When $L$ equals the number of time periods $T$ in the sample, the truncated model is in practice equal to the full Dirichlet process mixture model. However, smaller values for $L$ improve computational time significantly and when $L$ is large enough, the approximation error is negligible.

The degree $L$ weak limit approximation fosters models with less than $L$ regimes while allowing for new regimes, bounded by $L$, when new data are observed. Since the state assignments and the number of different states with nonzero assigned observations can differ over different sample iterations, the posterior distributions of the estimated parameters can be different for each time period while the sampler finds only a small number of different

regimes. In each iteration, the sampler draws for each of the $L$ states the transition probabilities, which are used to sample the state assignments for each observation. Due to small transition probabilities, some states can stay empty. Since the state labels may switch over different MCMC iterations (label switching) we cannot identify regime-specific posterior quanities, see, for example, Frühwirth-Schnatter (2001). Therefore, we report observation-specific posterior results which are identified, see Geweke (2007a).

**Prior Distributions**

The degree $L$ weak limit approximation induces finite Dirichlet distribution priors on $\beta$ and $\pi_i$,

$$\begin{aligned}\beta|\gamma &\sim Dir(\gamma/L, \ldots, \gamma/L),\\ \pi_i|\alpha, \beta, \kappa &\sim Dir(\alpha\beta_1, \ldots, \alpha\beta_i + \kappa, \ldots, \alpha\beta_L),\end{aligned}$$

where $\beta$ and $\pi_i$ are $L$-dimensional row vectors and $Dir$ denotes the finite Dirichlet distribution. We let the data determine the number of states with $L$ as an upper bound and the number of regime-switches by treating the hyperparameters of the transition distributions $\{\gamma, \alpha, \kappa\}$ as unknown. We place priors on these hyperparameters,

$$\alpha + \kappa \sim \text{Gamma}(a_\alpha, b_\alpha), \qquad \gamma \sim \text{Gamma}(a_\gamma, b_\gamma), \qquad \rho = \frac{\kappa}{\alpha + \kappa} \sim \text{Beta}(c_\rho, d_\rho).$$

The parameters in the prior distributions of the concentration parameters, denoted by $\{a_\alpha, b_\alpha, a_\gamma, b_\gamma\}$, control the dispersion with respect to their base distributions and thereby the expected number of states. The prior beliefs about the number of regime-switches are captured by $\{c_\rho, d_\rho\}$. A relatively low value for $c_\rho$ corresponds to rapid switches from one state to another. Increasing $c_\rho$ leads to higher state-persistence.

The prior on the parameters $\theta_{s_t} = \{B_{s_t}, \Sigma_{s_t}\}$ is a Normal-inverse-Wishart,

$$\text{vec}(B_{s_t})|\Sigma_{s_t}, \Theta \sim \mathcal{N}(\text{vec}(b_B), V_B \otimes \Sigma_{s_t}), \qquad \Sigma_{s_t}|\Theta \sim \mathcal{IW}(\nu_\Sigma, S_\Sigma), \tag{3.10}$$

where the vec$(A)$ operator stacks the columns of matrix $A$ and $\Theta$ is the collection of hyperparameters for $\theta$, $\{b_B, V_B, \nu_\Sigma, S_\Sigma\}$. So the infinite hidden Markov model allows for an elegant

conjugate prior structure in which we can put prior beliefs about the model parameters in the coefficient and covariance matrices directly in the prior distribution in (3.10). For instance, we can control the prior probability mass at stationary VARs, by shrinking the coefficients to zero with values close to zero in $b_B$ and relatively small in $V_B$.

Moreover, the prior specification for the parameters in covariance matrix $\Sigma_{s_t}$ in (3.10) does not rely on a factorization of the covariance matix, as is common in time-varying parameter vector autoregressive models in the tradition of Primiceri (2005),

$$\Sigma_{s_t} = A_{s_t} D_{s_t} A'_{s_t}, \tag{3.11}$$

where $A_{s_t}$ is a lower triangular matrix and $D_{s_t}$ a diagonal matrix. Using this specification of $\Sigma_{s_t}$, the priors are specified elementwise on the elements in $A_{s_t}$ and $D_{s_t}$. Hence, the ordering of the variables in the model has an impact on the implied prior for $\Sigma_{s_t}$ and therefore also on the joint posterior of all the model parameters.

### Posterior Distribution

Fox et al. (2011) derive a sample algorithm applicable for Bayesian parameter inference in the infinite hidden Markov model. We extend the sampling steps to the multivariate econometric time series context of the time-varying parameter VAR models and present the resulting sampling steps:

**Step 1.** Set the truncation level $L$ of possible hidden Markov states. Sample an initial draw for the hyperparameters of the transition distributions from their priors and do the same for the hyperparameters in the base distribution $H_\theta$. Initialize the transition distributions $\beta$ and $\pi_i$ by drawing from their $L$-dimensional Dirichlet priors.

**Step 2.** Sample the regime indicators $s_{1:T}$ using backward messages $m_{t,t-1}(i)$ from the state assignment probabilities in time period $t$ to state $i$ in time period $t-1$.

**(a)** First, work sequentially backwards in time. For each $i = 1, \ldots, L$, $m_{T+1,T}(i) = 1$ and

$$m_{t,t-1}(i) = \sum_{j=1}^{L} \pi_{ij} \mathcal{N}(y_t; B_j x_t, \Sigma_j) m_{t+1,t}(j), \qquad t = T, \ldots, 2,$$

where $\mathcal{N}(y; \mu, \Sigma)$ denotes the probability density function of the multivariate Normal distribution with mean $\mu$ and covariance matrix $\Sigma$.

**(b)** Second, work sequentially forward in time and initialize the number of transitions from state $i$ to $j$ observed in the state vector $s_{1:T}$, $n_{ij} = 0$ with $i, j = 1, \ldots, L$. For each $j = 1, \ldots, L$, compute the probability that observation $y_t$ is assigned to state $j$

$$f_j(y_t) = \pi_{s_{t-1},j} \mathcal{N}(y_t; B_j x_t, \Sigma_j) m_{t+1,t}(j), \qquad t = 1, \ldots, T,$$

sample the regime indicators from a discrete distribution,

$$s_t \sim \sum_{j=1}^{L} f_j(y_t) 1(s_t = j), \qquad t = 1, \ldots, T,$$

and increment $n_{s_{t-1},s_t}$.

**Step 3.** Sample auxiliary variables $m$, $w$, and $\bar{m}$ to simplify the resampling of $\beta$.

**(a)** For $i = 1, \ldots, L$ and $j = 1, \ldots, L$ set $m_{ij} = 0$. For $k = 1, \ldots, n_{ij}$ sample $x_k \sim \text{Bernoulli}(\frac{\alpha\beta_j + \kappa 1(i=j)}{i - 1 + \alpha\beta_j + \kappa 1(i=j)})$ and increment $m_{ij}$ if $x_k = 1$.

**(b)** For $i = 1, \ldots, L$ sample $w_i \sim \text{Binomial}(m_{ii}, \frac{\rho}{\rho + \beta_i(1-\rho)})$. Set $\bar{m}_{ij} = m_{ij}$ if $i \neq j$ and $\bar{m}_{ij} = m_{ij} - w_i$ if $i = j$.

**Step 4.** Sample the global transition distribution

$$\beta \sim \text{Dir}(\gamma/L + \sum_i \bar{m}_{i1}, \ldots, \gamma/L + \sum_i \bar{m}_{iL}).$$

**Step 5.** Sample the transition distribution $\pi$. For $i = 1, \ldots, L$ sample

$$\pi_i \sim \text{Dir}(\alpha\beta_1 + n_{i1}, \ldots, \alpha\beta_i + \kappa + n_{ii}, \ldots, \alpha\beta_L + n_{iL}).$$

**Step 6.** Sample the regime parameters $\theta$ for $j = 1, \ldots, L$. Let $x_j$ be the $t_j \times k$ matrix with rows $x_{s_t=j}$ and $t_j$ the number of observations in state $s_t$. Define $y_j$ as a $p \times t_j$ matrix.

Sample the model parameters

$$\bar{B} = (x_j' x_j + V_B^{-1}), \qquad \bar{b} = (y_j x_j + b_B V_B^{-1})\bar{B}^{-1},$$
$$\bar{S} = S_\Sigma + (y_j' - x_j \bar{b}')'(y_j' - x_j \bar{b}') + (\bar{b} - b_B)V_B^{-1}(\bar{b} - b_B)',$$
$$\Sigma_j | y_j, \Theta \sim \mathcal{IW}(\nu_\Sigma + t_j, \bar{S}), \qquad \text{vec}(B_j) | y_j, \Sigma_j, \Theta \sim \mathcal{N}(\text{vec}(\bar{b}), \bar{B}^{-1} \otimes \Sigma_j).$$

**Step 7.** Sample the hyperparameters of the transition distributions $\gamma$, $\alpha$, and $\kappa$.

**(a)** Sample auxiliary variables $r_i \sim \text{Beta}(\alpha + \kappa + 1, \sum_j n_{ij})$ and $s_i \sim \text{Bernoulli}(\frac{\sum_j n_{ij}}{\sum_j n_{ij} + \alpha + \kappa})$ for $i = 1, \ldots, L$ to simplify the posterior distributon of $\alpha + \kappa$.
Sample $\alpha + \kappa \sim \text{Gamma}(a_\alpha + \sum_i \sum_j m_{ij} - \sum_i s_i, (\frac{1}{b_\alpha} - \sum_i \log r_i)^{-1})$.

**(b)** Sample $\rho = \frac{\kappa}{\alpha + \kappa} \sim \text{Beta}(c_\rho + \sum_i w_i, d_\rho + \sum_i \sum_j m_{ij} - \sum_i w_i)$.

**(c)** Sample auxiliary variables $r \sim \text{Beta}(\gamma + 1, \sum_i \sum_j \bar{m}_{ij})$ and $s \sim \text{Bernoulli}(\frac{\sum_i \sum_j \bar{m}_{ij}}{\sum_i \sum_j \bar{m}_{ij} + \gamma})$. Compute $\bar{K} = \sum_k 1(\sum_i \bar{m}_{ij} > 0)$ and Sample $\gamma \sim \text{Gamma}(a_\gamma + \bar{K} - s, (\frac{1}{b_\gamma} - \log r)^{-1})$.

**Step 8.** Go to step 2.

### Predictive Densities

To construct a predictive density, we again make use of the degree $L$ weak limit approximation. When $L$ is assumed to be much larger than the number of in-sample regimes, the infinite hidden Markov model takes out-of-sample parameter breaks into account by allowing for new regimes out-of-sample. Here we show how future values are sampled from their predictive densities, together with the potentially new regimes, and the corresponding future model parameter values.

We simulate the predictive densities of $y_{T+h}$ for different horizons $h$ by iterating over the auto-regressive equation in (3.2), in each iteration of the sampler, using the parameter draws obtained in that sample iteration. In iteration $(i)$ of the sampler we have,

$$y_{T+h}^{(i)} = B_{s_{T+h}^{(i)}}^{(i)} x_{T+h-1}^{(i)} + \varepsilon_{T+h}^{(i)}, \qquad \varepsilon_{T+h}^{(i)} \sim \mathcal{N}(0, \Sigma_{s_{T+h}^{(i)}}^{(i)}),$$
$$s_{T+h}^{(i)} \sim \text{Multinomial}(\pi_{s_{T+h-1,1}}^{(i)}, \ldots, \pi_{s_{T+h-1,L}}^{(i)}),$$

where $B^{(i)}_{s^{(i)}_{T+h}}$ and $\Sigma^{(i)}_{s^{(i)}_{T+h}}$ are the parameter draws in iteration $(i)$ of the sampler, and $x^{(i)}_{T+h-1}$ is constructed from $y^{(i)}_{T+h-1}, \ldots, y^{(i)}_{T+h-l}$, where elements are replaced by in-sample observations when known. The regime indicators $s^{(i)}_{T+h}$ are sampled from a Multinomial distribution using the in-sample draws for the state transition probabilities, and $L$ is the number of clusters under the degree $L$ weak limit approximation. Since the degree $L$ weak limit approximation is assumed to be much larger than the number of estimated in-sample states, future parameter values can be drawn from new regimes which are not present in-sample.

## 3.3 Empirical Application

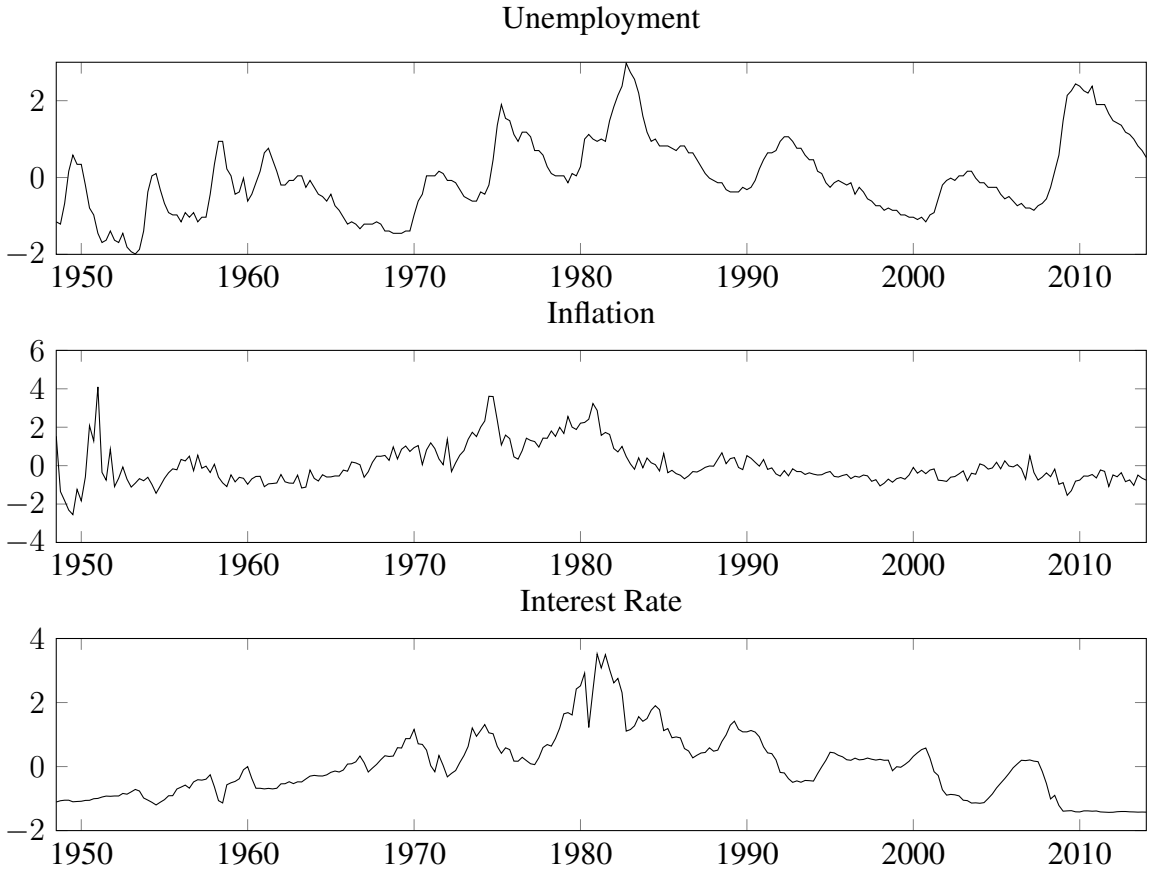
We apply the newly proposed model on a monetary VAR of the U.S. economy consisting of the unemployment rate, inflation rate and federal funds rate, in an in-sample and out-of-sample application. Section 3.3.1 introduces the data. In Section 3.3.2 we use the infinite hidden Markov model to study the effects of monetary policy shocks in a structural VAR model. Section 3.3.3 performs a real-time forecasting exercise with the a small monetary VAR model to assess the out-of-sample performance of the infinite hidden Markov model compared to benchmark models. In 3.3.3 we also analyse the forecast performance of a large dimensional VAR model.

We follow Primiceri (2005) and consider VAR models with two lags. Posterior results are based on 20,000 iterations of the MCMC sampler, from which the first 10,000 are discarded. Visual inspection shows that this number of iterations is enough for convergence. The number of possible states is truncated at 20 in the degree $L$ weak limit approximation.

### 3.3.1 Data

We use three macroeconomic time series of the U.S. economy, the unemployment rate, inflation rate, and interest rate, to construct a monetary VAR. The unemployment rate is the civilian rate of unemployment, and inflation is calculated as a function of the GDP deflator $P_t$ to obtain the annualized quarterly growth rate of prices; 400 times the first difference of the logarithm of $P_t$. Since the three month Treasury bill rate is available over a longer period of time than the federal funds rate, the interest rate is represented by the first.

**Figure 3.1:** Time Series Small Monetary VAR U.S. Economy




This figure shows the standardized quarterly data series as included in the small monetary VAR of the U.S. economy, with the unemployment rate, inflation rate, and interest rate as variables. The inflation series represents 400 times the first difference of the logarithm of the consumer price index for all urban consumers. The interest rate denotes the effective federal funds rate in percentages. The sample period runs from the first quarter of 1948 to the last quarter of 2015.

The real-time data for the unemployment rate and the GDP deflator are collected by the Federal Reserve Bank of Philadelphia. The three month Treasury bill rate is not subject to revisions and is available from the Federal Reserve Bank of St. Louis. The GDP deflator is available as quarterly time series and the unemployment rate and interest rate as monthly time series. We follow Cogley and Sargent (2002, 2005); Cogley et al. (2010); D'Agostino et al. (2013), by taking the value at the second month of the quarter for the unemployment and the value at the first month of the quarter for the interest rate, to obtain quarterly series for all three variables.

The first quarter of 1948 is the first time period for which all data is available. We consider data through 2015Q4. When we date a vintage as the last quarter for which all data are available, we have vintages from 1965Q4 to 2015Q4. We use all data in the most recent

**Table 3.1:** Parameters of Prior Distributions

| $a_\alpha$ | $b_\alpha$ | $a_\gamma$ | $b_\gamma$ | $c_\rho$ | $d_\rho$ | $\nu_\Sigma$ | $S_\Sigma$ | $b_B$ | $V_{B[i=j]}$ | $V_{B[i\neq j]}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 10 | 1 | 10 | 10 | 1 | $p+2$ | $\frac{1}{\nu_\Sigma} I_p$ | $0_{p\times k}$ | $\left(\frac{\lambda}{\text{lag}^2}\right)$ | 0 |

This table shows the parameters of the priors as discussed in Section 3.2.2, where $0_{p\times q}$ represents a zero matrix of size $p \times q$ and $I_q$ is the identity matrix of dimension $q$. The diagonal elements of $V_B$ are scaled by the lag order of the corresponding variables in $x_t$, where the lag length equals $\lambda$ for the intercept.

vintage for in-sample analysis. Figure 3.1 shows the standardized data series as included in the model.

Table 3.1 shows the prior parameter values of the model. We follow Fox et al. (2011) in the parameter values in the prior distributions on the hyperparameters of the transition distribution. We opt for a non-informative prior by choosing large scale parameters $\{b_\alpha, b_\gamma\}$ in the Gamma distributions. A relatively low value for $c_\rho$, that is $c_\rho = 10$, corresponds to rapid switches from one state to another. Setting $c_\rho = 1000$ leads to higher state-persistence.

Since the data is standardized, we can choose non-informative priors for the model parameters. The inverse-Wishart distribution of the covariance parameters has degrees of freedom equal to the number of variables in the model plus two and a scaled identity matrix as scale matrix. The prior mean of the coefficient parameters is set to zero, and we can control the shrinkage of the coefficient estimates towards zero by the shrinkage parameter $\lambda$ in the prior variance of the coefficient parameters. By scaling the prior variance of coefficient parameters by the lag order of the corresponding variable, we shrink coefficients estimates of higher lag order variables to zero, similar to the Minnesota prior (Doan et al., 1984).

For out-of-sample purposes discussed later, we try to avoid sampling many explosive VAR parameters by choosing a more tighter parametrization of the prior distribution of the coefficients. Although it is know that alternating between explosive and nonexplosive regimes can still produce non-explosive processes in the long-run (Francq and Zakoıan, 2001) we want to exclude having long periods of explosive regimes in our prediction sample. In sum, for in-sample analysis we set$\lambda = 1$ and $\lambda = 0.1$ in our forecasting exercise.

### 3.3.2 Structural VAR model

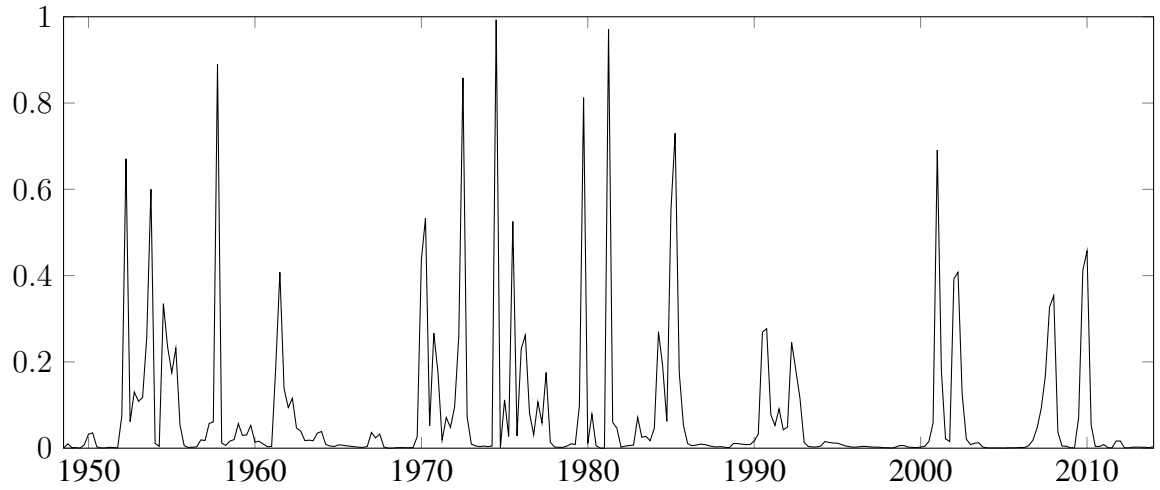
Since a time varying parameter VAR consists of a large amount of parameters, even in a small model with only three variables and two lags, it takes too much space to present estimation diagnostics of all parameters. Moreover, it is difficult to give an economic interpretation to each posterior distribution. Therefore, we discuss the stability of the VAR model over time, we show the posterior results for the variances of the structural shocks, and we report impulse response functions which summarize the economic implications of the estimated structural coefficients. To identify the structural parameters, we opt for a Cholesky decomposition in our application on a small monetary VAR in which the variables are ordered as {unemployment, inflation, interest rate} (Sims, 1980). However, alternative identification schemes, for instance, long-run restrictions or sign restrictions, can also be applied to the infinite hidden Markov VAR model.

#### Stability Diagnostics

We find compelling evidence of instability in the parameter estimates over time. The posterior probability for four different regimes equals 70%, for five regimes 28%, and the remaining probability mass is concentrated at six regimes. Figure 3.2 shows the posterior probability of a regime switch for each time period. Most of the breaks are detected before 1990. Thereafter, there is a more stable period which is followed by higher break probabilities corresponding to the Dot-com bubble in 2000 and the global financial crisis starting in 2007.

The instability is also reflected by the time variation in the posterior probability of an explosive system. When the largest absolute eigenvalue of the companion form of the reduced form VAR is larger than one in a specific time period, the system is in an explosive regime at that point in time. Figure 3.3 shows the time-varying posterior probability of an explosive regime. Most of the probability mass for the largest absolute eigenvalue is located below one. However, before the mid eighties we observe temporary increases in the probability of an explosive system.

**Figure 3.2:** Posterior Break Probabilities




This figure shows for each time period the posterior probability of switching regime in the current period compared to the previous period, computed as the sum of draws in the sampler with a regime-switch in a specific time period divided by the total number of draws.

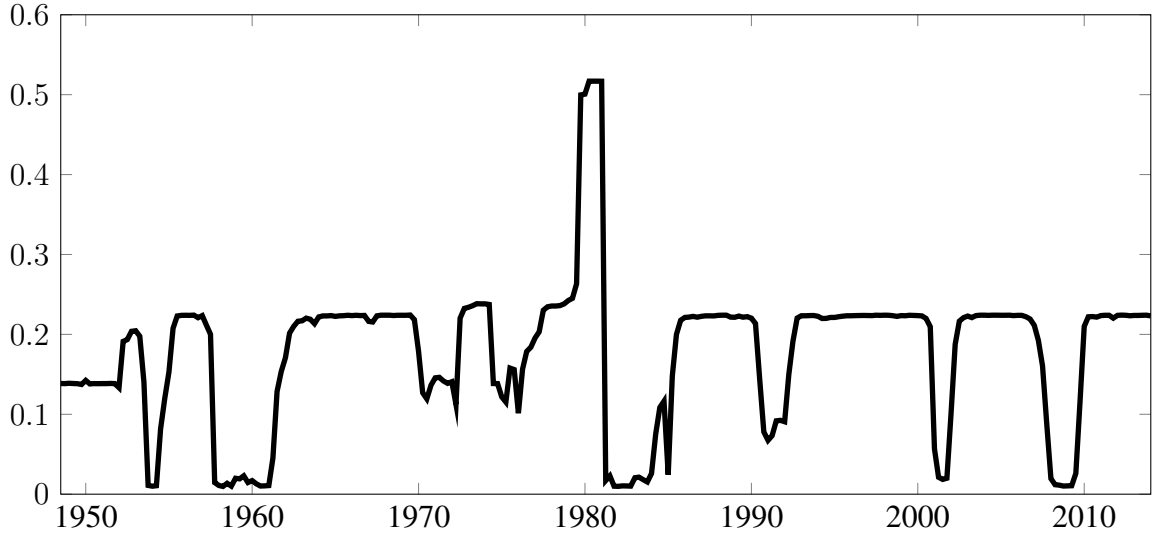
### Structural Variance

We define monetary policy shocks as interest rate responses to variables other than unemployment and inflation. The changes in relative importance of these shocks over time are displayed by the time-varying variance of the monetary policy shock. Figure 3.4 shows the time-varying variance of monetary policy shocks, together with shocks to the unemployment and inflation equation.

The confidence bands around the time-varying variance of monetary policy shocks provide evidence that there is variation over time. The first thing to note beside some peaks in variance in the early sixties and mid seventies, is the long period of high variance running from 1979 to 1983. This feature is well-known and can be attributed to a period with deviant monetary policy. After this high variance regime, the changes in variance are quite modest with only small exceptions.

The first panel of Figure 3.4 shows that the variance of unemployment shocks follows a pattern similar to the variance of monetary policy shocks. The confidence bounds also clearly support modelling time variation. The variance of the shocks to inflation, as showed in the second panel of Figure 3.4, seems to be less volatile over time. Apart from high variance regimes in the seventies, it behaves more stable in the rest of the sample period.

**Figure 3.3:** Largest Absolute Eigenvalues Companion Form




This figure shows the time-varying posterior probability of an explosive regime.

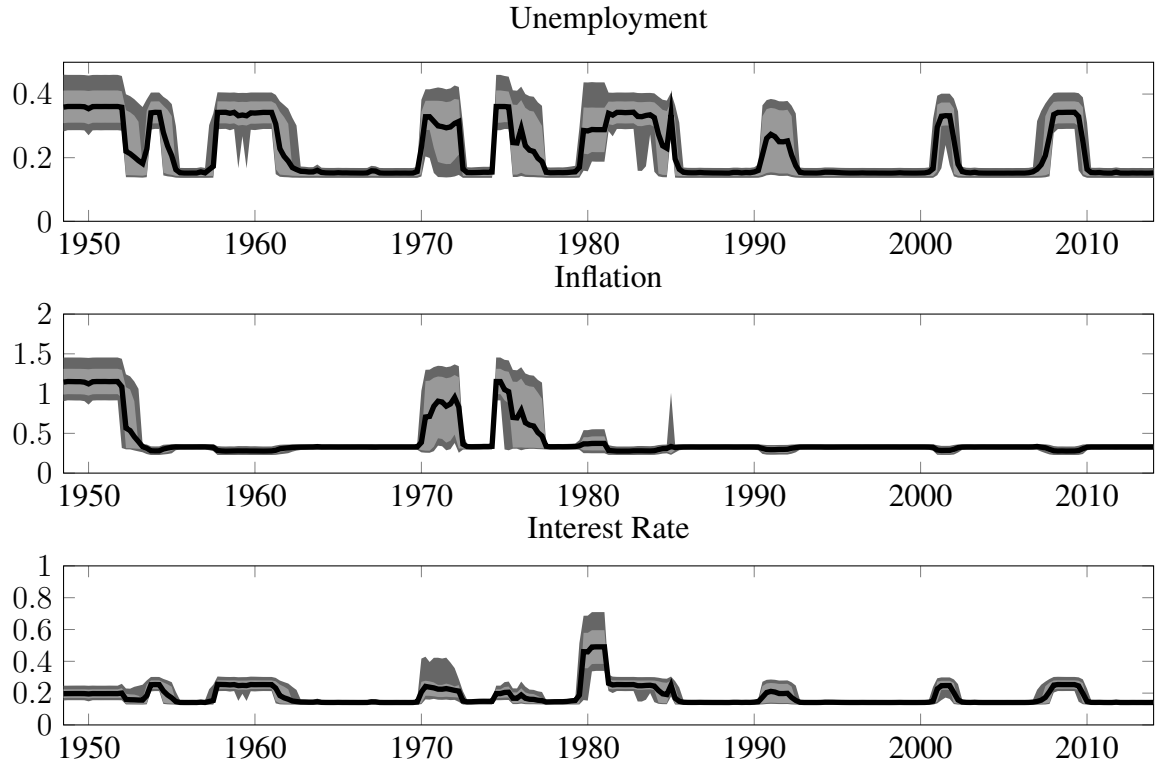
### Impulse Response Functions

Figure 3.5 shows the posterior mean of the impulse response functions of the SVAR to a unit monetary policy shock, conditional on the estimated regimes. The functions trace out the effect of the structural shocks over a time path of four years for each variable, conditional on the estimated states. Because of the time-varying parameters in the model, the time paths are different for each date the shock hits the system. So for each quarter in the estimation sample, we have the time path of the policy shock effect over the sixteen following quarters.

According to the posterior mean, the impulse responses differ in strength over time. In general, the effects in the period around 1970 and 1980 and the period after 2007 seem to be more severe. The initial reaction of unemployment can be both positive or negative, but the result of the shock after four years seems always to be positive. The impulse response of inflation differs also in strength over time, but in general it reacts positive to an interest shock and converges back to zero, which is known as the price puzzle. The interest rate also seems to converge to zero, after a sharp decline following the impulse of magnitude one.

The impulse response functions in Figure 3.5 suggest some variation over time. However, Figure 3.5 shows only the mean of the posterior distribution of the impulse responses. To get an idea about the uncertainty around the time variation we arbitrarily choose four different moments at which a monetary policy shock hits the system, and plot the posterior means together with the 68% and 90% of the posterior distributions. Figure 3.6 shows the impulse

**Figure 3.4:** Posterior Means of the Structural Variance Parameters

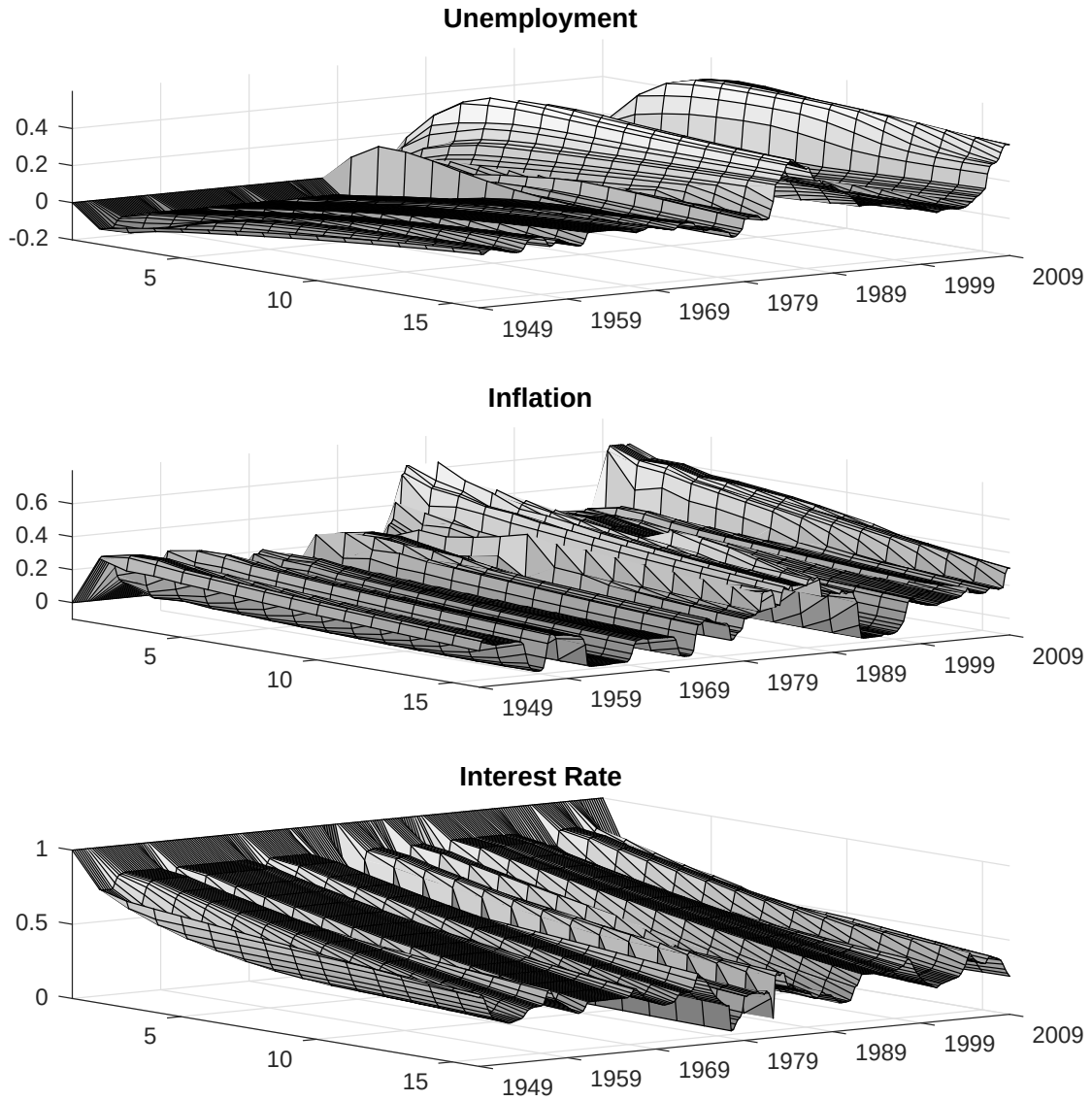



This figure shows the time-varying posterior means (solid line) of the structural variance parameters together with the 68% and 90% confidence bands. The panels show from top to bottom the variance of the residuals in the unemployment equation, inflation equation, and interest equation, respectively.

response functions to a monetary policy shock for the different shock dates; the second quarter of 1965, 1980, 1995, and 2010. When taking the whole posterior distribution of the impulse responses into account, we can hardly conclude that the responses are time-varying. Figure 3.6 shows that, apart from the impulse response of the interest rate in 1980, we can hardly find any differences in the shape of the response functions per variable, but the magnitude of the effect differs over time. However, due to the high uncertainty about the shape and magnitude of the impulse responses, we find no convincing posterior evidence for time-variation in impulse responses functions.

According to the confidence bounds there is little posterior evidence that the impulse responses of unemployment and inflation differ from zero. However, Table 3.2 shows that after some shock dates the probability mass clearly indicates a positive effect. With a posterior probability of approximately 80 percent, a monetary policy shocks results in a positive effect on unemployment after four years, for all different shock dates. For inflation, there seem to be only a short-term positive effect, also with probabilities close to 80 percent.

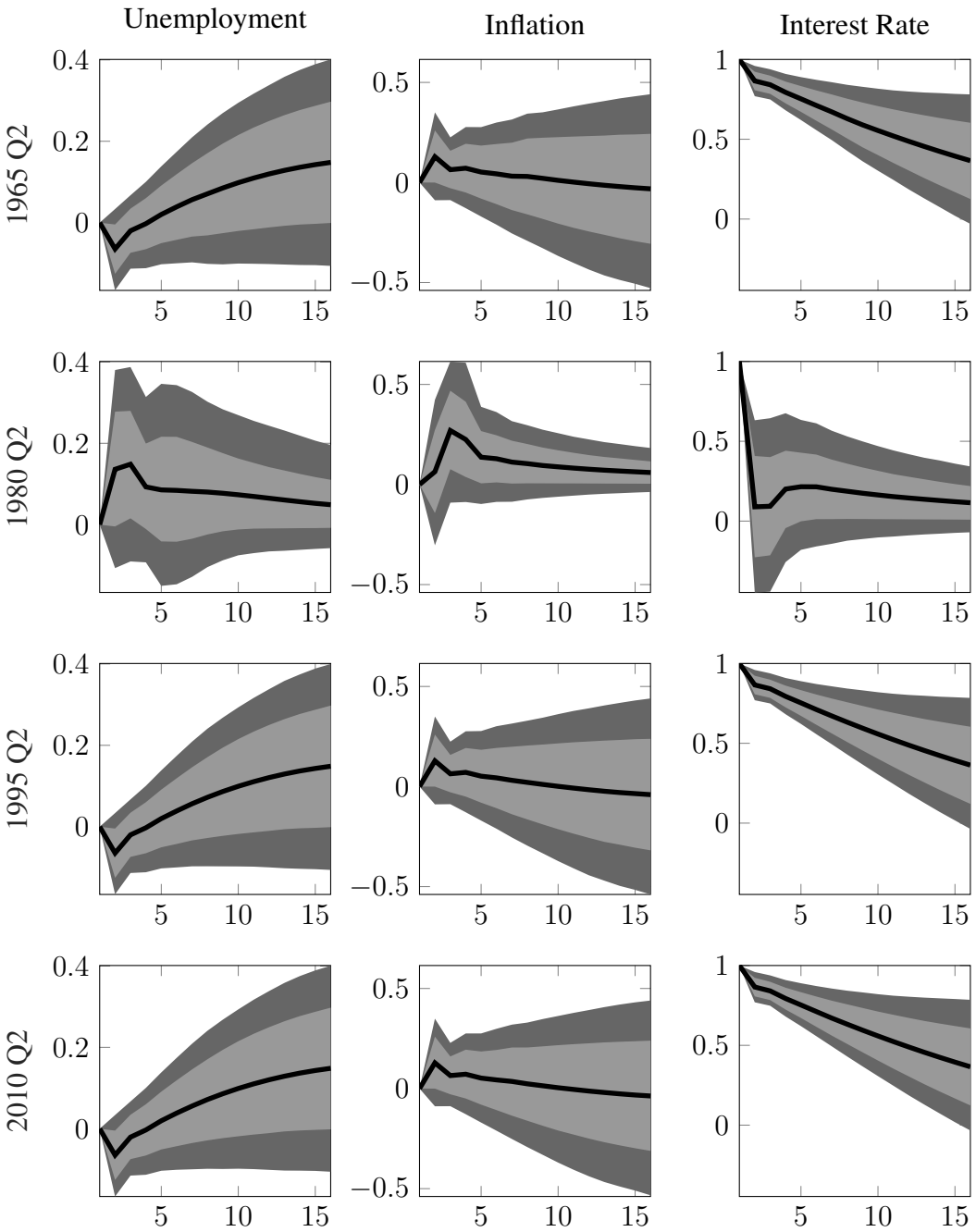
**Figure 3.5:** Impulse Response Functions




This figure shows the posterior means of the impulse response functions to a monetary policy shock. From top to bottom we have the responses in the unemployment equation, inflation equation, and interest equation. The y-axis runs from 0 to 16 and traces out the effect of the shock over a period of four years. The x-axis indicates at which date a shock hits the system. The z-axis shows the magnitude of the response. The monetary policy shock is defined as a one unit shock to interest rate.

The impulse response functions are constructed from a large number of coefficients which are allowed to be different over time. To decrease parameter uncertainty, we can impose all coefficients to be constant over time. However, this results also in time-invariant impulse response functions. Alternatively, we restrict only the long term impact to be con-

**Figure 3.6:** Impulse Response Functions




This figure shows the posterior means (solid line) of the impulse response functions together with the 68% and 90% confidence bands for different moments in time. The monetary policy shock is defined as a one unit shock to interest rate.

stant over time and account for variation in short term effects

$$y_t = c_{s_t} + \Pi_{s_t} y_{t-1} + D\Delta y_{t-1} + \varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}(0, \Sigma_{s_t}). \tag{3.12}$$

**Table 3.2:** Posterior Probability of Positive Impulse Response

| | Unemployment | | Inflation | | Interest Rate | |
|---|---|---|---|---|---|---|
| | 1 year | 4 year | 1 year | 4 year | 1 year | 4 year |
| 1965 Q2 | 49.4 | 83.9 | 71.7 | 46.4 | 100.0 | 93.4 |
| 1980 Q2 | 81.0 | 79.0 | 88.8 | 85.7 | 79.4 | 86.2 |
| 1995 Q2 | 49.3 | 83.9 | 71.6 | 44.6 | 100.0 | 93.1 |
| 2010 Q2 | 49.2 | 83.8 | 71.8 | 45.1 | 100.0 | 93.3 |

This table shows the posterior probabilities that the impulse response functions in Figure 3.6 are larger than zero, one year and four years after a monetary policy shock.

Appendix 3.A discusses the sample steps for this restricted model and Appendix 3.B shows the impulse response functions following from this more parsimonious model. However, despite of the decrease in parameter uncertainty, we still do not find posterior support for time-varying impulse response functions.

### 3.3.3 Forecasting Exercise

To asses the out-of-sample performance of the infinite hidden Markov VAR model, we perform a forecasting exercise in which we compare the predictive performance of the infinite hidden Markov model against benchmark models. We adopt a similar real-time forecasting framework as D'Agostino et al. (2013), who iteratively produce forecasts with the time-varying parameter VAR model of Primiceri (2005).

We start the forecasting exercise with an estimation sample running from 1948Q1 up to 1969Q4 of the vintage 1969Q4. We standardize the variables and estimate the model parameters on this sample. We compute with each model forecasts up to five quarters ahead outside the estimation window, from 1970Q1 to 1971Q1. After we have produced the forecasts based on the first estimation sample, we move one quarter ahead and re-estimate the model parameters using the standardized variables based on all the data in vintage 1970Q1. That means that we use an expanding window to estimate the model parameters. Again, we use each model to compute forecasts up to five quarters ahead. We repeat this procedure up to vintage 2014Q1 (as we need later vintages to evaluate forecasts up to five quarters ahead, as we discuss in Section 3.3.3). This exercise results in time series of 178 one-period-ahead forecasts from 1970Q1 to 2014Q2 and a time series of the same length containing

five-periods-ahead predictions from 1971Q1 to 2015Q2, since we compare the forecasts to data after 2 revisions.

Since there is evidence that large VAR models can improve in forecast performance upon small models, we also extend the small model to ten variables to construct forecasts of the unemployment rate, inflation rate, and interest rate. We add five variables from the real-time database of the Federal Reserve Bank of Philadelphia; M1 money stock, real gross domestic product, personal consumption expenditures, industrial production index, and imports of goods and services, and we add the S&P 500 index and the total borrowings of depository institutions from the Federal Reserve which are unrevised data from the Federal Reserve Bank of St. Louis. In case of monthly data we take the value of the third month and all data is included in the model as growth rates by taking 400 times the first difference of the logarithm. The flow variables real gross domestic product, personal consumption expenditures, and imports of goods and services are available as quarterly data.

**Forecasting Models**

We compare the predictive performance of variants of the infinite hidden Markov model against the predictive performance of a time-invariant Bayesian VAR, the time-varying parameter VAR model of Primiceri (2005), and the time-varying parameter VAR model of Koop and Korobilis (2013). The model of Primiceri (2005) is designed for the specific type of small monetary VAR studied in this forecast exercise. To avoid over-fitting and since a parsimonious model could potentially lead to a more efficient model, and therefore an increase in predictive performance, we not only forecast with the unrestricted infinite hidden Markov VAR model, but also with models in which we restrict coefficients or the covariance matrix to be time-invariant. The model in which we restrict both boils down to a linear Bayesian VAR without any time-variation included.

The linear Bayesian VAR model is identical to the model in (3.1), but now the parameters are fixed over time;

$$y_t = Bx_t + \varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}(0, \Sigma), \qquad t = 1, \ldots, T, \tag{3.13}$$

and the prior on the parameters $\theta = \{B, \Sigma\}$ is a Normal-inverse-Wishart as in (3.10) with values for the hyperparameters given in Table 3.1. We simulate from the predictive densities

of $y_{T+h}$ for different horizons $h$ as described for the infinite hidden Markov model in Section 3.2.2.

We specify and estimate the Primiceri model as outlined in Primiceri (2005). However, for a fair comparison with the other models, we follow Koop and Korobilis (2013) in not using a training sample prior. We generate forecast from the Primiceri model (denoted as P05) in the same way as Koop and Korobilis (2013) and D'Agostino et al. (2013). That means that we use iterated forecasts in the same way as discussed for the linear BVAR, and we allow for out-of-sample parameter change in the VAR.

Note that due to computational constraints and the fact that the estimation algorithm involves taking the inverse of large matrices, the model of Primiceri (2005) cannot be estimated in a stable way for a large VAR model. That is why Koop and Korobilis (2013) propose an alternative model that can handle large dimensional time-varying parameter VAR models by using forgetting factors to model time-variation. We implement this model, which we denote as KK13, as final benchmark. We take for the forgetting factors $\lambda = 0.99$, $\kappa = 0.96$, and $\gamma = 0.1$. This specification does not involve dynamic model averaging. Comparing averaged forecasts over different model dimensions and prior specifications against the forecasts of the other models, would also require the implementation of the dynamic model averaging technique for the competing models, which is beyond the scope of this chapter. Moreover, we set $\hat{\Sigma}_0 = I_p$ instead of using a training sample.

**Forecast Evaluation**

Following the framework of D'Agostino et al. (2013), we compare the forecasts for a particular time period with the third release of the figures for that time period. So, we evaluate predictions against numbers which may have been revised two times. This means that we evaluate the last one-year-ahead prediction against the numbers in vintage 2015Q4.

We evaluate point forecasts using the root mean squared prediction error (RMSPE) and the mean absolute prediction error (MAPE). For the first (second) metric we set the point forecast equal to the mean (median) of the predictive density. We evaluate the forecast performance of the whole predictive density with the average predictive densities (APD). Beside assessing predictions with the traditional Bayesian forecast performance measures, like the RMSPE and the APD, we follow Groen et al. (2013) in evaluating density forecasts based

on alternative measures such as the continuous ranked probability score (CRPS) and the quantile scores.

Where point forecasts emphasis the median or mean of the predictive density, there are often applications in which the tails of the predictive density are of special interest. For instance, in case VAR models are used for constructing impulse response functions to perform policy analysis. The outcomes of policy analysis are heavily affected by the tail behavior of predicted future outcomes. To also evaluate the performance of the tails of the predictive density we employ integrated weighted versions of Gneiting and Raftery (2007) average quantile scores (avQS). With the avQS-C, avQS-R, and avQS-L we evaluate the center, the right tail, and the left tail of the predictive density, respectively. The exact formulas of all evaluation measures are given in Appendix 3.C.

**Forecast Performance**

The real-time forecasting exercise results in forecasts of five different models[1], for all three variables included in the monetary VAR, over five horizons, for both a small monetary VAR and a large dimensional VAR. Tables 3.3, 3.4 and 3.5 show the values of the predictive performance measures for these forecasts. Underscores indicate the best performing model for a specific horizon and variable, according to a particular evaluation criterion. Except for the average predictive density, the best performing model is the one that produces forecasts with small values for the forecast performance statistics.

We find that in general, the infinite hidden Markov model with a time-varying covariance matrix shows the best performance, where the unrestricted infinite hidden Markov model is too flexible and the homogeneous linear VAR too restrictive. Table 3.3 shows that the unrestricted infinite hidden Markov model outperforms the other models in only one case and the linear BVAR results only once in the lowest RMSPE. Also on the predictive density evaluation measures in Tables 3.4 and 3.5, the unrestricted model shows in a few cases the best performance, but the linear BVAR is systematically outperformed.

The restricted infinite hidden Markov model results in all cases in the lowest RMSPE for unemployment. The results for inflation and interest rate vary per horizon and model

[1] Although we only present results for the unrestricted infinite hidden Markov model and a variant in which the coefficients are restricted to be time-invariant, Appendix 3.A shows that we can also restrict the covariance matrix, the long term impact matrix, or both. Forecast results for these models do not alter the main findings of our analysis and are available upon request.

**Table 3.3:** Forecasting Results RMSPE

| | | small VAR | | | | | large VAR | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | IHM | | | KK13 | P05 | IHM | | | KK13 |
| hor. | var. | $\theta_{s_t}$ | $\Sigma_{s_t}$ | $\theta$ | $\theta_t$ ∼ff | $\theta_t$ ∼rw | $\theta_{s_t}$ | $\Sigma_{s_t}$ | $\theta$ | $\theta_t$ ∼ff |
| 1 | UR | 0.554 | 0.386 | 0.444 | 0.387 | 0.407 | 0.398 | 0.337 | 0.357 | 0.363 |
| | Infl. | 1.497 | 1.465 | 1.516 | 1.460 | 1.433 | 1.486 | 1.449 | 1.498 | 1.375 |
| | IR | 1.035 | 0.869 | 0.899 | 0.847 | 0.864 | 0.924 | 0.849 | 0.871 | 0.811 |
| 2 | UR | 0.834 | 0.621 | 0.723 | 0.624 | 0.674 | 0.612 | 0.526 | 0.593 | 0.560 |
| | Infl. | 1.712 | 1.650 | 1.756 | 1.686 | 1.597 | 1.680 | 1.620 | 1.688 | 1.549 |
| | IR | 1.440 | 1.225 | 1.251 | 1.205 | 1.242 | 1.287 | 1.186 | 1.242 | 1.190 |
| 3 | UR | 1.049 | 0.805 | 0.939 | 0.825 | 0.930 | 0.847 | 0.734 | 0.844 | 0.763 |
| | Infl. | 1.893 | 1.838 | 1.982 | 1.859 | 1.745 | 1.890 | 1.895 | 1.903 | 1.720 |
| | IR | 1.730 | 1.435 | 1.508 | 1.474 | 1.551 | 1.620 | 1.499 | 1.571 | 1.502 |
| 4 | UR | 1.195 | 0.964 | 1.109 | 0.998 | 1.228 | 1.092 | 0.941 | 1.068 | 0.962 |
| | Infl. | 2.067 | 2.065 | 2.222 | 2.091 | 2.004 | 2.269 | 2.268 | 2.212 | 2.036 |
| | IR | 2.017 | 1.713 | 1.788 | 1.771 | 1.899 | 1.859 | 1.787 | 1.848 | 1.786 |
| 5 | UR | 1.301 | 1.100 | 1.236 | 1.150 | 1.636 | 3.474 | 1.120 | 1.248 | 1.131 |
| | Infl. | 2.096 | 2.247 | 2.383 | 2.332 | 2.374 | 4.191 | 2.455 | 2.343 | 2.235 |
| | IR | 2.624 | 1.961 | 2.030 | 2.068 | 2.313 | 4.846 | 2.036 | 2.054 | 2.017 |

This table shows the RMSPE for unemployment (UR), inflation (Infl.), and interest rate (IR) over five different horizons; from one-quarter ahead till five-quarters ahead. Forecasts are produced by the infinite hidden Markov model in which all model parameters $\theta_{s_t} = \{B_{s_t}, \Sigma_{s_t}\}$ change over regimes, a version with only a time-varying covariance matrix $\Sigma_{s_t}$, and the linear Bayesian VAR (column $\theta$). The benchmark time-varying parameter VAR models of Koop and Korobilis (2013), in which time-variation is governed by forgetting factors (ff), and Primiceri (2005), wherein time-variation is modelled as random walks (rw), are denoted as KK13 and P05, respectively. The left panel shows results in a small VAR and the right panel in a large VAR model. Underscores indicate the best performing model for a specific horizon, variable and model dimension.

dimension. The P05 model improves upon forecast accuracy in predicting inflation in the small VAR. Since the estimation procedure of this model cannot be scaled up to high dimensional models, this model is absent in the forecast comparison for the large VAR. The KK13 model replaces the P05 model as best performer for inflation here. The KK13 model and the restricted infinite hidden Markov model are close competitors on accurate point forecasts for interest rate. Table 3.6 in Appendix 3.D shows the MAPE, another point forecast evaluation criterium, which is more friendly against outliers. This metric shows results similar to the RMSPE, with the restricted infinite hidden Markov model being superior in forecasting unemployment, P05 consistently better in forecasting inflation in the small model, and in some cases KK13 shows the best performance on inflation and interest rate in the large VAR.

**Table 3.4:** Forecasting Results APD

| | small VAR | | | | | large VAR | | | |
|---|---|---|---|---|---|---|---|---|---|
| | IHM | | | KK13 | P05 | IHM | | | KK13 |
| hor. | $\theta_{s_t}$ | $\Sigma_{s_t}$ | $\theta$ | $\theta_t \sim$ff | $\theta_t \sim$rw | $\theta_{s_t}$ | $\Sigma_{s_t}$ | $\theta$ | $\theta_t \sim$ff |
| 1 | 0.569 | 1.924 | 0.469 | 1.352 | 1.675 | 0.171 | 0.335 | 0.187 | 0.323 |
| 2 | 0.266 | 0.678 | 0.223 | 0.470 | 0.555 | 0.057 | 0.153 | 0.060 | 0.048 |
| 3 | 0.170 | 0.387 | 0.143 | 0.242 | 0.289 | 0.069 | 0.065 | 0.031 | 0.010 |
| 4 | 0.119 | 0.249 | 0.106 | 0.137 | 0.157 | 0.018 | 0.042 | 0.021 | 0.002 |
| 5 | 0.090 | 0.166 | 0.083 | 0.084 | 0.095 | 0.090 | 0.022 | 0.015 | 0.001 |

This table shows the APD for five different horizons; from one-quarter ahead till five-quarters ahead. For additional information, see the note following Table 3.3.

When we take the whole predictive density into account, we find impressive results in favor of the restricted infinite hidden Markov model. Table 3.4 shows that this model beats all benchmark models for all horizons and model dimensions. The only competing model on the predictive density is the unrestricted infinite hidden Markov model. In the small VAR model, the linear BVAR shows the worst predictive performance of the benchmark models, followed by the KK13 and P05 models. However, unless for the one-period-ahead forecasts, the KK13 model cannot increase in the value of the predictive density relative to the linear BVAR in the large dimensional model.

The average predictive density is computed over the forecasts for all variables included in the model. However, out of all variables included in the large VAR model, we focus on unemployment, inflation, and interest rate. The continuous ranked probability score separately evaluates the predictive densities of each variable. Table 3.5 shows that the strong performance in density forecasts of the restricted infinite hidden Markov model is based on the forecasts for unemployment and interest rates. The P05 model does again a good job in predicting inflation in the small model. In the large dimensional model, the KK13 model only sometimes perform better on short-term density forecasts. Tables 3.7, 3.8, and 3.9 in Appendix 3.D show the quantile scores, which show in which part of the predictive densities which model performs best. In summary, the restricted infinite hidden Markov model in the small VAR performs often better than benchmarks model in the important left-tail.

Comparing the forecasts in the small VAR against the forecasts in the large VAR model, we find that in most cases the forecast quality deteriorates with larger model dimensions. Table 3.3 shows that only for short horizons there is an improvement in forecast accuracy

**Table 3.5:** Forecasting Results avCRPS

| | | small VAR | | | | | large VAR | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | IHM | | | KK13 | P05 | IHM | | | KK13 |
| hor. | var. | $\theta_{s_t}$ | $\Sigma_{s_t}$ | $\theta$ | $\theta_t \sim$ff | $\theta_t \sim$rw | $\theta_{s_t}$ | $\Sigma_{s_t}$ | $\theta$ | $\theta_t \sim$ff |
| 1 | UR | 0.288 | 0.199 | 0.243 | 0.204 | 0.215 | 0.215 | 0.181 | 0.199 | 0.196 |
| | Infl. | 0.819 | 0.793 | 0.842 | 0.788 | 0.775 | 0.814 | 0.797 | 0.822 | 0.755 |
| | IR | 0.512 | 0.406 | 0.468 | 0.420 | 0.430 | 0.492 | 0.442 | 0.462 | 0.428 |
| 2 | UR | 0.435 | 0.314 | 0.376 | 0.326 | 0.346 | 0.332 | 0.284 | 0.324 | 0.306 |
| | Infl. | 0.913 | 0.869 | 0.958 | 0.892 | 0.846 | 0.934 | 0.901 | 0.937 | 0.860 |
| | IR | 0.737 | 0.615 | 0.674 | 0.631 | 0.672 | 0.710 | 0.643 | 0.682 | 0.648 |
| 3 | UR | 0.550 | 0.408 | 0.480 | 0.432 | 0.466 | 0.461 | 0.392 | 0.455 | 0.428 |
| | Infl. | 1.017 | 0.965 | 1.077 | 0.988 | 0.938 | 1.034 | 1.039 | 1.049 | 0.996 |
| | IR | 0.922 | 0.767 | 0.840 | 0.817 | 0.893 | 0.913 | 0.846 | 0.883 | 0.852 |
| 4 | UR | 0.634 | 0.498 | 0.566 | 0.529 | 0.591 | 0.572 | 0.498 | 0.572 | 0.559 |
| | Infl. | 1.111 | 1.068 | 1.188 | 1.112 | 1.059 | 1.177 | 1.219 | 1.197 | 1.213 |
| | IR | 1.086 | 0.931 | 1.001 | 0.998 | 1.108 | 1.066 | 1.019 | 1.046 | 1.046 |
| 5 | UR | 0.699 | 0.575 | 0.633 | 0.619 | 0.718 | 0.654 | 0.588 | 0.661 | 0.697 |
| | Infl. | 1.193 | 1.168 | 1.279 | 1.253 | 1.216 | 1.210 | 1.318 | 1.276 | 1.438 |
| | IR | 1.233 | 1.100 | 1.153 | 1.181 | 1.323 | 1.186 | 1.167 | 1.176 | 1.246 |

This table shows the avCRPS for unemployment (UR), inflation (Infl.), and interest rate (IR) over five different horizons; from one-quarter ahead till five-quarters ahead. For additional information, see the note following Table 3.3.

after adding variables to the small monetary VAR model. In contrast to the average predictive densities, we can compare the density forecast performance between the small and large models using the continuous ranked probability score. Table 3.5 shows that in the short term, the large model does a better job in density forecasts for unemployment. In general, there is no clear increase in performance by adding more variables.

In sum, the infinite hidden Markov model with a time-varying covariance matrix shows for most forecast horizons and variables the best performance based on various evaluation measures, and is always close to the best performing model if it is not the best one itself. For all considered evaluation measures and forecast horizons, the infinite hidden Markov model outperforms the benchmark models in forecasting unemployment. We find that inflation and interest rate are for some horizons better predicted by the KK13 or P05 models based on point forecast evaluation metrics. However, based on the predictive densities the infinite hidden Markov model shows superior forecast performance. Finally, increasing the number of variables in the VAR does, in general, not lead to an increase in forecast performance for unemployment, inflation rate, and interest rate.

## 3.4 Conclusion

In this chapter we propose a new method to estimate time-varying parameters in a VAR model. To avoid the curse of dimensionality, we opt for a semi-parametric approach. The infinite hidden Markov model encourages estimation of a parsimonious model by clustering parameter values over time, without restricting the parameter space. To accommodate for persistence in macroeconomic data, we impose the Dirichlet process mixture on the transition probabilities in a hidden Markov-switching framework. Parameter values are assigned to a possibly infinite number of states, with a potentially increased probability of self-transition. Except from the degree $L$ weak limit approximation, which comes with negligible costs, the estimation algorithm of the model does not impose any restrictions on or (linear) approximations to the parameters.

The empirical application shows that the semi-parametric Bayesian framework is a promising alternative for parametric approaches to time-varying parameter VAR modelling. We identify both abrupt and smooth parameter changes in a structural analysis and find posterior evidence for time-varying volatility. A real-time forecasting exercise shows that over a collection of forecast evaluation criteria the infinite hidden Markov model often outperforms popular benchmark models, even in large VAR models consisting of ten variables.

## 3.A Parameter Restrictions

This appendix discusses the sample steps for three different restricted versions of model (3.1); a model in which either the coefficient matrix, covariance matrix, or long-term impact matrix are set to be constant over time. posterior results of the parameters in these restricted models are obtained with only slight modifications to the sampler in Subsection 3.2.2. In practice, we adjust Step 6 of the sampler and add an extra step to sample the fixed parameters outside the structure of the mixture model.

The model in (3.1) can be generalized to

$$y_t = B_t x_t + C\tilde{x}_t + \varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}(0, \Sigma_t), \qquad t = 1, \ldots, T,$$

from which follows the unrestricted model in (3.1) by setting the parameters in $C$ equal to zero. Imposing $\Sigma_t = \Sigma$ results in the model with a time-invariant covariance matrix. A model with time-invariant macroeconomic relations is defined by $x_t = 1$, $\tilde{x}_t = [y'_{t-1}, \ldots, y'_{t-l}]'$, and $\tilde{y}_t = y_t - B_t x_t$. Finally, the short-term model follows from $x_t = [1, y'_{t-1}]'$, $\tilde{x}_t = [\Delta y'_{t-1}, \ldots, \Delta y'_{t-l}]'$, and $\tilde{y}_t = y_t - B_t x_t$.

We sample $B_t$ as in step 6 of the sample algorithm but we take $(y_t - C\tilde{x}_t)$ for $y_t$. The same holds for $\Sigma_t$ when the covariance parameters are unrestricted. After the seventh step, when the state assignments in the mixture model are settled down for the current iteration, we sample the time-invariant parameter matrices.

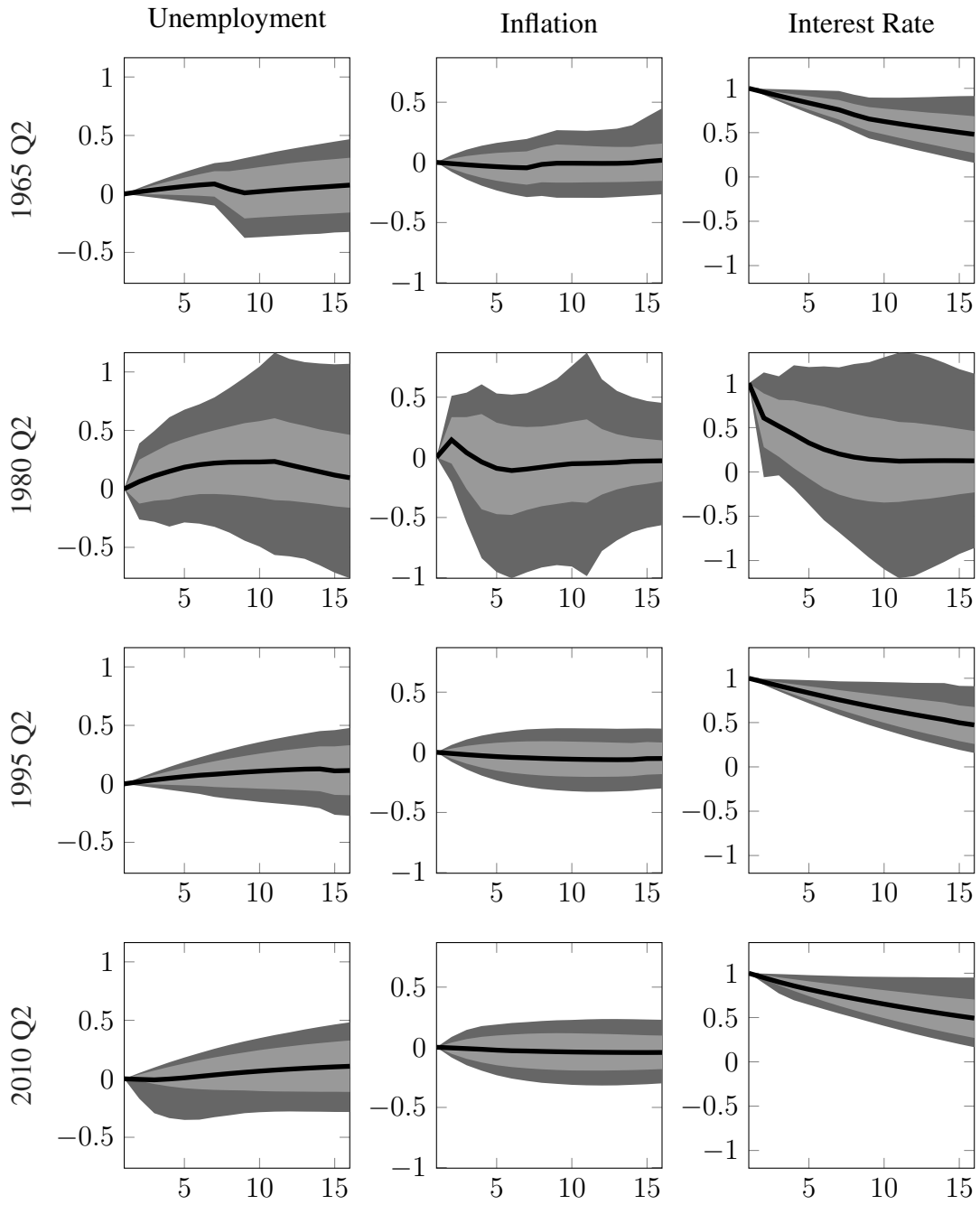
In models with a restricted covariance matrix we compute $\varepsilon = (\varepsilon'_1, \ldots, \varepsilon'_T)'$ where $\varepsilon_t = y'_t - x_t B'_t - \tilde{x}_t C'$ and add the sampling step

$$\bar{S} = S_\Sigma + \varepsilon'\varepsilon, \qquad \Sigma_j | y, \Theta \sim \mathcal{IW}(\nu_\Sigma + T, \bar{S}).$$

For the model with restricted coefficients we compute $\tilde{X} = (\Sigma_1^{-\frac{1}{2}} \tilde{X}'_1, \ldots, \Sigma_T^{-\frac{1}{2}} \tilde{X}'_T)'$, where $\tilde{X}_t = I_p \otimes \tilde{x}_t$, $t = 1, \ldots, T$, and $I_p$ is the identity matrix of dimension $p$, and $\tilde{Y} = (\tilde{y}'_1, \ldots, \tilde{y}'_T)$. Now we can perform the sampling steps

$$\bar{B} = (\tilde{X}'\tilde{X} + (I_p \otimes V_B)^{-1})^{-1}, \qquad \bar{b} = \bar{B}(\tilde{X}'\tilde{Y} + (I_p \otimes V_B)^{-1} \text{vec}(b_B)),$$
$$\text{vec}(B) | \tilde{y}, \Sigma, \Theta \sim \mathcal{N}(\text{vec}(\bar{b}), \bar{B}).$$

## 3.B Impulse Response Functions




This figure shows impulse response functions constructed from a model with time-invariant long term impact matrix. For additional information, see the note following Figure 3.6.

## 3.C Forecast Performance

We evaluate point forecasts using the root mean squared prediction error (RMSPE) and the mean absolute prediction error (MAPE). The RMSPE of the forecast produced by model $M$

for variable $i$ at horizon $h$ is

$$\text{RMSPE}_{ih}^{M} = \sqrt{\frac{1}{P}\sum_{t=T+1}^{T+P}\left(\hat{y}_{t+h}^{(i)}(M) - y_{t+h}^{(i)}\right)^2},$$

where $\hat{y}_{t+h}^{(i)}(M)$ is one of the $P$ point forecast of the $i$th variable $y_{t+h}^{(i)}$ made by model $M$. We set the point forecast equal to the mean of the predictive density. The MAPE is defined by

$$\text{MAPE}_{ih}^{M} = \frac{1}{P}\sum_{t=T+1}^{T+P}\left|\hat{y}_{t+h}^{(i)}(M) - y_{t+h}^{(i)}\right|,$$

with the point forecast equal to the median of the predictive density.

We evaluate the forecast performance of the whole predictive density with the average predictive densities (APD)

$$f^{M}(y_{t+h}) = \frac{1}{P}\sum_{t=T+1}^{T+P}\left(\frac{1}{S}\sum_{s=1}^{S}\mathcal{N}(B_{s_{T+h}^{(i)}}^{(i)}x_{T+h-1}^{(i)}, \Sigma_{s_{T+h}^{(i)}}^{(i)})\right),$$

where $S$ denotes the number of simulations.

The continuous ranked probability score (CRPS) is computed as

$$\text{CRPS}_t(y_{t+h}^{(i)}) = E_f|Y_{t+h}^{(i)} - y_{t+h}^{(i)}| - \frac{1}{2}E_f|Y_{t+h}^{(i)} - Y_{t+h}^{\prime(i)}|,$$

where $f$ is the predictive density function of model $M$ for prediction $y_{t+h}^{(i)}$, $E_f$ is the expectation operator over the function $f$, $|.|$ denotes the absolute value, and $Y_{t+h}^{(i)}$ and $Y_{t+h}^{\prime(i)}$ are independent random variables with sampling density $f$. The average CRPS across all forecasts is

$$\text{avCRPS}_{ih}^{M} = \frac{1}{P}\sum_{t=T+1}^{T+P}\text{CRPS}_t(y_{t+h}^{(i)}).$$

The avQS-C, avQS-R, and avQS-L evaluate the center, the right tail, and the left tail of the predictive density, respectively.

$$\text{avQS-C}_{ih}^{M} = \frac{1}{P}\sum_{t=T+1}^{T+P}\left(\frac{1}{99}\sum_{j=1}^{99}\alpha_j(1-\alpha_j)\text{QS}(\alpha_j, y_{t+h}^{(i)}, M)\right),$$
$$\text{avQS-R}_{ih}^{M} = \frac{1}{P}\sum_{t=T+1}^{T+P}\left(\frac{1}{99}\sum_{j=1}^{99}\alpha_j^2\text{QS}(\alpha_j, y_{t+h}^{(i)}, M)\right),$$
$$\text{avQS-L}_{ih}^{M} = \frac{1}{P}\sum_{t=T+1}^{T+P}\left(\frac{1}{99}\sum_{j=1}^{99}(1-\alpha_j)^2\text{QS}(\alpha_j, y_{t+h}^{(i)}, M)\right),$$

where $\alpha_j = j/100$ and

$$\text{QS}(\alpha, y_{t+h}^{(i)}, M) = (I\{y_{t+h}^{(i)} \leq Q_f^{\alpha}\} - \alpha)(Q_f^{\alpha} - y_{t+h}^{(i)}),$$

where $Q_f^{\alpha}$ represents quantile $\alpha$ of the predictive density function $f$ of model $M$ for prediction $y_{t+h}^{(i)}$.

# 3.D Additional Forecasting Results

**Table 3.6:** Forecasting Results MAPE

| | | small VAR | | | | | large VAR | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | IHM | | | KK13 | P05 | IHM | | | KK13 |
| hor. | var. | $\theta_{s_t}$ | $\Sigma_{s_t}$ | $\theta$ | $\theta_t \sim$ff | $\theta_t \sim$rw | $\theta_{s_t}$ | $\Sigma_{s_t}$ | $\theta$ | $\theta_t \sim$ff |
| 1 | UR | 0.377 | 0.270 | 0.312 | 0.276 | 0.295 | 0.291 | 0.251 | 0.264 | 0.271 |
| | Infl. | 1.159 | 1.117 | 1.168 | 1.115 | 1.078 | 1.106 | 1.083 | 1.117 | 1.050 |
| | IR | 0.670 | 0.537 | 0.599 | 0.537 | 0.543 | 0.645 | 0.590 | 0.594 | 0.564 |
| 2 | UR | 0.575 | 0.416 | 0.500 | 0.437 | 0.462 | 0.447 | 0.387 | 0.437 | 0.399 |
| | Infl. | 1.255 | 1.222 | 1.307 | 1.235 | 1.152 | 1.317 | 1.269 | 1.327 | 1.181 |
| | IR | 1.004 | 0.820 | 0.924 | 0.828 | 0.859 | 0.992 | 0.860 | 0.957 | 0.855 |
| 3 | UR | 0.728 | 0.526 | 0.647 | 0.567 | 0.600 | 0.638 | 0.527 | 0.632 | 0.528 |
| | Infl. | 1.404 | 1.352 | 1.467 | 1.327 | 1.243 | 1.415 | 1.410 | 1.435 | 1.266 |
| | IR | 1.256 | 1.046 | 1.156 | 1.088 | 1.138 | 1.290 | 1.158 | 1.256 | 1.137 |
| 4 | UR | 0.836 | 0.641 | 0.772 | 0.689 | 0.759 | 0.793 | 0.679 | 0.794 | 0.681 |
| | Infl. | 1.544 | 1.473 | 1.627 | 1.476 | 1.406 | 1.622 | 1.663 | 1.638 | 1.453 |
| | IR | 1.476 | 1.277 | 1.382 | 1.337 | 1.403 | 1.500 | 1.405 | 1.482 | 1.377 |
| 5 | UR | 0.923 | 0.742 | 0.860 | 0.788 | 0.923 | 0.895 | 0.790 | 0.893 | 0.801 |
| | Infl. | 1.666 | 1.596 | 1.757 | 1.628 | 1.574 | 1.760 | 1.785 | 1.747 | 1.606 |
| | IR | 1.685 | 1.514 | 1.598 | 1.590 | 1.660 | 1.695 | 1.625 | 1.658 | 1.592 |

This table shows the MAPE for unemployment (UR), inflation (Infl.), and interest rate (IR) over five different horizons; from one-quarter ahead till five-quarters ahead. For additional information, see the note following Table 3.3.

**Table 3.7:** Forecasting Results avQS-left

| | | small VAR | | | | | large VAR | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | IHM | | | KK13 | P05 | IHM | | | KK13 |
| hor. | var. | $\theta_{s_t}$ | $\Sigma_{s_t}$ | $\theta$ | $\theta_t \sim$ff | $\theta_t \sim$rw | $\theta_{s_t}$ | $\Sigma_{s_t}$ | $\theta$ | $\theta_t \sim$ff |
| | UR | 0.041 | 0.029 | 0.036 | 0.030 | 0.032 | 0.032 | 0.027 | 0.029 | 0.030 |
| 1 | Infl. | 0.121 | 0.117 | 0.124 | 0.118 | 0.115 | 0.121 | 0.119 | 0.121 | 0.116 |
| | IR | 0.079 | 0.064 | 0.072 | 0.066 | 0.069 | 0.075 | 0.069 | 0.070 | 0.069 |
| | UR | 0.059 | 0.044 | 0.053 | 0.046 | 0.050 | 0.049 | 0.042 | 0.048 | 0.047 |
| 2 | Infl. | 0.131 | 0.128 | 0.138 | 0.132 | 0.128 | 0.135 | 0.132 | 0.137 | 0.131 |
| | IR | 0.109 | 0.095 | 0.100 | 0.098 | 0.104 | 0.104 | 0.101 | 0.100 | 0.102 |
| | UR | 0.073 | 0.056 | 0.066 | 0.060 | 0.067 | 0.068 | 0.058 | 0.067 | 0.067 |
| 3 | Infl. | 0.143 | 0.139 | 0.152 | 0.140 | 0.141 | 0.151 | 0.153 | 0.153 | 0.151 |
| | IR | 0.133 | 0.116 | 0.121 | 0.124 | 0.138 | 0.131 | 0.128 | 0.126 | 0.132 |
| | UR | 0.084 | 0.068 | 0.077 | 0.074 | 0.085 | 0.083 | 0.074 | 0.084 | 0.089 |
| 4 | Infl. | 0.154 | 0.154 | 0.168 | 0.156 | 0.158 | 0.169 | 0.177 | 0.171 | 0.181 |
| | IR | 0.156 | 0.138 | 0.143 | 0.152 | 0.171 | 0.152 | 0.153 | 0.148 | 0.163 |
| | UR | 0.092 | 0.078 | 0.085 | 0.087 | 0.103 | 0.095 | 0.086 | 0.096 | 0.113 |
| 5 | Infl. | 0.164 | 0.169 | 0.180 | 0.176 | 0.180 | 0.179 | 0.189 | 0.179 | 0.216 |
| | IR | 0.176 | 0.164 | 0.164 | 0.180 | 0.205 | 0.170 | 0.175 | 0.165 | 0.195 |

This table shows the avQS-l for unemployment (UR), inflation (Infl.), and interest rate (IR) over five different horizons; from one-quarter ahead till five-quarters ahead. For additional information, see the note following Table 3.3.

**Table 3.8:** Forecasting Results avQS-center

| | | small VAR | | | | | large VAR | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | IHM | | | KK13 | P05 | IHM | | | KK13 |
| hor. | var. | $\theta_{s_t}$ | $\Sigma_{s_t}$ | $\theta$ | $\theta_t \sim$ff | $\theta_t \sim$rw | $\theta_{s_t}$ | $\Sigma_{s_t}$ | $\theta$ | $\theta_t \sim$ff |
| | UR | 0.028 | 0.019 | 0.023 | 0.020 | 0.021 | 0.021 | 0.018 | 0.019 | 0.019 |
| 1 | Infl. | 0.081 | 0.078 | 0.083 | 0.078 | 0.076 | 0.080 | 0.078 | 0.080 | 0.074 |
| | IR | 0.049 | 0.039 | 0.045 | 0.040 | 0.041 | 0.047 | 0.043 | 0.044 | 0.041 |
| | UR | 0.042 | 0.030 | 0.036 | 0.032 | 0.034 | 0.032 | 0.028 | 0.032 | 0.030 |
| 2 | Infl. | 0.089 | 0.085 | 0.093 | 0.087 | 0.082 | 0.092 | 0.089 | 0.092 | 0.084 |
| | IR | 0.071 | 0.060 | 0.065 | 0.061 | 0.064 | 0.069 | 0.062 | 0.067 | 0.063 |
| | UR | 0.053 | 0.039 | 0.046 | 0.042 | 0.045 | 0.045 | 0.038 | 0.044 | 0.041 |
| 3 | Infl. | 0.099 | 0.095 | 0.105 | 0.096 | 0.090 | 0.102 | 0.102 | 0.103 | 0.096 |
| | IR | 0.089 | 0.075 | 0.082 | 0.080 | 0.085 | 0.090 | 0.083 | 0.087 | 0.083 |
| | UR | 0.061 | 0.048 | 0.055 | 0.051 | 0.056 | 0.056 | 0.049 | 0.056 | 0.053 |
| 4 | Infl. | 0.109 | 0.104 | 0.116 | 0.108 | 0.101 | 0.116 | 0.119 | 0.117 | 0.115 |
| | IR | 0.105 | 0.091 | 0.098 | 0.097 | 0.105 | 0.105 | 0.100 | 0.103 | 0.101 |
| | UR | 0.067 | 0.055 | 0.061 | 0.059 | 0.068 | 0.064 | 0.057 | 0.064 | 0.065 |
| 5 | Infl. | 0.117 | 0.114 | 0.125 | 0.121 | 0.115 | 0.125 | 0.129 | 0.125 | 0.134 |
| | IR | 0.120 | 0.108 | 0.113 | 0.115 | 0.125 | 0.118 | 0.114 | 0.116 | 0.119 |

This table shows the avCS-c for unemployment (UR), inflation (Infl.), and interest rate (IR) over five different horizons; from one-quarter ahead till five-quarters ahead. For additional information, see the note following Table 3.3.

**Table 3.9:** Forecasting Results avQS-right

| | | small VAR | | | | | large VAR | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | IHM | | | KK13 | P05 | IHM | | | KK13 |
| hor. | var. | $\theta_{s_t}$ | $\Sigma_{s_t}$ | $\theta$ | $\theta_t \sim$ff | $\theta_t \sim$rw | $\theta_{s_t}$ | $\Sigma_{s_t}$ | $\theta$ | $\theta_t \sim$ff |
| | UR | 0.049 | 0.032 | 0.040 | 0.034 | 0.034 | 0.034 | 0.029 | 0.033 | 0.031 |
| 1 | Infl. | 0.132 | 0.127 | 0.137 | 0.125 | 0.125 | 0.131 | 0.128 | 0.133 | 0.116 |
| | IR | 0.081 | 0.063 | 0.075 | 0.065 | 0.067 | 0.078 | 0.069 | 0.074 | 0.065 |
| | UR | 0.076 | 0.053 | 0.064 | 0.055 | 0.057 | 0.054 | 0.046 | 0.053 | 0.048 |
| 2 | Infl. | 0.151 | 0.140 | 0.159 | 0.144 | 0.136 | 0.152 | 0.146 | 0.151 | 0.136 |
| | IR | 0.121 | 0.096 | 0.109 | 0.099 | 0.106 | 0.115 | 0.099 | 0.110 | 0.099 |
| | UR | 0.099 | 0.071 | 0.084 | 0.074 | 0.079 | 0.075 | 0.064 | 0.073 | 0.068 |
| 3 | Infl. | 0.172 | 0.159 | 0.182 | 0.167 | 0.154 | 0.168 | 0.168 | 0.171 | 0.160 |
| | IR | 0.153 | 0.121 | 0.139 | 0.129 | 0.143 | 0.150 | 0.134 | 0.146 | 0.132 |
| | UR | 0.115 | 0.088 | 0.100 | 0.091 | 0.100 | 0.094 | 0.081 | 0.093 | 0.087 |
| 4 | Infl. | 0.189 | 0.177 | 0.201 | 0.191 | 0.175 | 0.197 | 0.200 | 0.199 | 0.202 |
| | IR | 0.182 | 0.150 | 0.167 | 0.158 | 0.177 | 0.177 | 0.162 | 0.174 | 0.163 |
| | UR | 0.127 | 0.101 | 0.112 | 0.107 | 0.123 | 0.111 | 0.097 | 0.110 | 0.108 |
| 5 | Infl. | 0.205 | 0.193 | 0.217 | 0.215 | 0.203 | 0.214 | 0.219 | 0.215 | 0.243 |
| | IR | 0.208 | 0.176 | 0.193 | 0.187 | 0.213 | 0.201 | 0.185 | 0.196 | 0.196 |

This table shows the avCS-r for unemployment (UR), inflation (Infl.), and interest rate (IR) over five different horizons; from one-quarter ahead till five-quarters ahead. For additional information, see the note following Table 3.3.

# Chapter 4

# Forecasting Using Random Subspace Methods

*Joint work with Tom Boot*

## 4.1 Introduction

Due to the increase in available macroeconomic data, dimension reduction methods have become an indispensable tool for accurate forecasting. One well-known approach to reduce the dimension of the predictor set is to identify a small set of factors that drive most of the variation in the high-dimensional predictor set, as in Stock and Watson (2002, 2006) and Bai and Ng (2006, 2008). Whether one uses the original predictor set or the extracted factors, selection of the relevant predictors is commonly subject to substantial uncertainty. Consequently, employing model selection and shrinkage methods that estimate inclusion weights for the predictors increases the forecast variance (Ng, 2013).

A seemingly naive strategy is to forgo data-based shrinkage or selection, and assign random weights to the predictors. Although a priori there seems to be little reason to expect this approach to lead to accurate forecasts, empirical evidence suggests otherwise. For example, Elliott et al. (2013, 2015) find that averaging over forecasts constructed from many randomly selected subsets of fixed size substantially lowers the mean squared forecast error compared with data-driven alternatives. The theoretical justification of such randomized approaches is

not completely understood. We provide both theoretical and extensive empirical evidence for the intriguingly strong performance of random subspace methods.

We distinguish two different approaches to constructing a random subspace. The first method we consider is random subset regression, where a randomly chosen subset of predictors is used to estimate a low-dimensional approximation to the original model and construct a forecast. The forecasts from many such submodels are then combined in order to lower the mean squared forecast error (MSFE).

Instead of selecting a subset of available predictors, random projection regression forms a low-dimensional subspace by averaging over predictors using random weights drawn from a standard normal distribution. Although not required in the setup here, the justification for this method is usually derived from the Johnson-Lindenstrauss lemma (Johnson and Lindenstrauss, 1984), which has very recently inspired several applications in the econometric literature on discrete choice models by Chiong and Shum (2016), forecasting product sales by Schneider and Gupta (2016), and forecasting using large vector autoregressive models by Koop et al. (2016) based on the framework of Guhaniyogi and Dunson (2015).

There are many random sampling methods which are widely used in the statistical and machine learning literature but rather new to economics (Ng, 2015). Bagging or bootstrap aggregation also selects a subset of available predictors, but differs from random subset regression in that each submodel is subject to some form of model selection. Averaging over the submodels serves to smooth selection errors (Breiman, 1996; Bühlmann and Yu, 2002; Inoue and Kilian, 2008). Similar to random projections, Frieze et al. (2004) and Mahoney and Drineas (2009) construct a new set of predictors by using predictor weights. However, these weights are drawn from distributions that depend on the original set of predictors. Ma et al. (2015) discuss related sampling methods focusing on a large number of observations instead of a large set of predictors. Furthermore, the random subspace methods we consider in this chapter differ from alternatives by using random weights that are independent of the data, involve a single tuning parameter, are less time consuming, and are extremely simple to implement.

We derive expressions for the upper bound on the asymptotic MSFE for random subset and random projection regression and use these bounds to determine in which settings the methods are most effective. A direct comparison between the two random subspace methods can be made when the predictors are uncorrelated. This setting nevertheless brings out

the main features we observe in general settings studied in Monte Carlo experiments. The bounds elicit that random projection regression shares certain properties with ridge regression. It achieves a low MSFE when highly variable predictors are the ones that are most strongly related to the dependent variable. On the other hand, the bound for random subset regression only depends on the aggregate signal and not on the variance of the individual predictors. When the relevant predictors have a lower than average variance, the bound for random subset regression is lower compared to random projection regression.

For random subset regression, the construction of an upper bound on the asymptotic MSFE appears new. For random projection regression, bounds are only available for the in-sample mean squared error under fixed regressors by Maillard and Munos (2009), Kabán (2014) and Thanei et al. (2017). Our out-of-sample bound improves upon the existing results for the in-sample mean squared error.

The bounds are derived for forecasts that take the expected value over the random subspaces. In practice, we have to settle for a finite number of draws. We show that this has a negligible effect on the asymptotic MSFE when the number of draws scales linearly with the number of predictors, up to a logarithmic factor. This explains why Elliott et al. (2013) find no deterioration in performance when not all subsets are used, which would require a number of draws exponential in the number of predictors.

The theoretical findings are confirmed in a set of Monte Carlo experiments, which also compare the performance of the randomized methods to several well-known alternatives: principal component regression, based on Pearson (1901), partial least squares by Wold (1982), ridge regression by Hoerl and Kennard (1970) and the lasso by Tibshirani (1996). Both randomized methods offer superior forecast accuracy over principal component regression, even in some cases when the data generating process is specifically tailored to suit this method. The random subspace methods outperform the lasso unless there is a small number of very large non-zero coefficients. Ridge regression is outperformed for a majority of the settings where the coefficients are not very weak. When the data exhibits a factor structure, and factors associated with intermediate eigenvalues drive the dependent variable, random subset regression is the only method that outperforms the historical mean of the data.

We empirically test the theoretical and Monte Carlo findings using monthly macroeconomic series in the FRED-MD dataset, introduced by McCracken and Ng (2016). Random subset regression provides the lowest MSFE relative to the benchmark models for at least

66% of the 130 series, followed by random projection regression. For both random subspace methods, the accuracy is shown to be substantially less dependent on the dimension of the reduced subspace than it is in case of principal component regression. Moreover, the dimension of the subspace should be chosen relatively large ($\geq 20$). This stands in stark contrast to what is common for principal component regression, where one often uses a small number of factors, see for example Stock and Watson (2012). We show how the average weights of the predictors in the random subspaces provide insight in the main drivers of the forecasts of the random subspace methods.

The article is structured as follows. Section 4.2 introduces the random subspace methods. The theoretical results on the forecast performance of these methods are derived in Section 4.3. A Monte Carlo study in Section 4.4 highlights the performance of the techniques under different model specifications. Section 4.5 considers an extensive empirical application using monthly macroeconomic data. Section 4.6 concludes.

## 4.2 Methods

Consider the model

$$y_{t+1} = w_t'\beta_w + x_t'\beta_x + \varepsilon_{t+1}, \tag{4.1}$$

where $w_t$ is a $p_w \times 1$ vector of variables that are always included in the model, $x_t$ is a $p_x \times 1$ vector of variables which potentially contain information on $y_{t+1}$, and the forecast error is denoted by $\varepsilon_{t+1}$. The time index $t$ runs from $t = 0, \ldots, T$.

We assume that $E[\varepsilon_{t+1}|w_t, x_t] = 0$ and $E[\varepsilon_{t+1}^2|w_t, x_t] = \sigma^2$. Further assumptions on the sequence $\{w_t, x_t, \varepsilon_{t+1}\}$ will be given in Section 4.3. Under these assumptions, both $w_t$ and $x_t$ can contain lags of $y_{t+1}$ or they can consist of factors derived from an additional set of observed variables.

We study the properties of point forecasts $\hat{y}_{T+1}$ for $y_{T+1}$ when the number of available predictors $p$ is large and fixed, the predictors in $x_t$ are weakly related to $y_{t+1}$, and $T \to \infty$. The predictors $z_t = (w_t', x_t')'$, with $t = 0, \ldots, T-1$, are used in the estimation of the $p \times 1$ parameter vector $\beta = (\beta_w', \beta_x')'$, and $z_T = (w_T', x_T')'$ is only used for the construction of the forecast for $y_{T+1}$.

Estimating $\beta$ by ordinary least squares (OLS) yields the following forecast,

$$\hat{y}_{T+1}^{\text{OLS}} = z_T'\hat{\beta} = z_T'(Z'Z)^{-1}Z'y, \tag{4.2}$$

where $y = (y_1, \ldots, y_T)'$, $Z = (z_0, \ldots, z_{T-1})'$, and $\hat{\beta}$ is the OLS estimator.

Since $\varepsilon_{T+1}$ is unpredictable for any method, we set $\varepsilon_{T+1} = 0$ to save on notation. Then the asymptotic mean squared forecast error equals,

$$\mathrm{E}_\varepsilon \left[ \lim_{T\to\infty} T\mathrm{E}_{z_T} \left[ \left( y_{T+1} - z_T'\hat{\beta} \right)^2 \right] \right] = \sigma^2 p, \tag{4.3}$$

which is an increasing function in the number of parameter estimates $p$.

### 4.2.1 Random subspace methods

Since the MSFE under OLS estimates increases with the number of estimated coefficients, the forecast in (4.2) gets inaccurate when $x_t$ contains a large number of predictors. To prevent this, we reduce the dimensionality of the predictor set by multiplying $x_t$ with a $p_x \times k$ matrix $R$, where $k < p_x$, to obtain the approximating model

$$y_{t+1} = w_t'\beta_w + x_t'R\beta_{x,R} + u_{t+1}. \tag{4.4}$$

The construction of the matrix $R$ is often data-driven. Model selection methods based on information criteria effectively estimate $R$ as a selection matrix based on the available data. Principal component regression takes $R$ as the matrix of principal component loadings corresponding to the $k$ largest eigenvalues from the sample covariance matrix of the regressors $x_t$. The key to random subspace methods is to generate the elements of $R$ from a probability distribution that is independent of the data. We consider the following two choices for $R$, which yield random subset regression and random projection regression.

#### Random subset regression

In random subset regression (RS), the matrix $R$ is a random selection matrix that selects a random set of $k$ predictors out of the original $p_x$ available predictors. For example, if $p_x = 5$

and $k = 3$, a possible realization of $R$ is

$$\begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}. \tag{4.5}$$

More in general, define an index $l = 1, \dots k$ with $k$ the dimension of the subspace, and a scalar $c(l)$ such that $1 \leq c(l) \leq p_x$. Denote by $e_{c(l)}$ a $p_x$-dimensional vector with all zeros except for the $c(l)$-th entry that equals one, then random subset regression is based on random matrices of the form

$$\left[e_{c(1)}, \dots, e_{c(k)}\right], \quad e_{c(m)} \neq e_{c(n)} \text{ if } m \neq n. \tag{4.6}$$

**Random projection regression**

Instead of selecting a subset of predictors, we can also take weighted averages to construct a new set of predictors. Random projection regression (RP) chooses the weights at random from a normal distribution. In this case, each entry of $R$ is independent and identically distributed as

$$[R]_{ij} \sim N(0, 1), \quad 1 \leq i \leq p_x, \; 1 \leq j \leq k. \tag{4.7}$$

### 4.2.2 Forecasts from low-dimensional models

We rewrite the approximating model (4.4) as

$$y_{t+1} = z_t' S_R \beta_R + u_{t+1}, \text{ with } S_R = \begin{pmatrix} I_{p_w} & O \\ O & R \end{pmatrix}. \tag{4.8}$$

The least squares estimator of $\beta_R$ is denoted by $\hat{\beta}_R$ and given by

$$\hat{\beta}_R = (S_R' Z' Z S_R)^{-1} S_R' Z' y. \tag{4.9}$$

Using this estimate, we construct a forecast as

$$\hat{y}_{T+1,R} = z_T' S_R \hat{\beta}_R. \tag{4.10}$$

If $R$ is a random matrix, then intuitively, relying on a single realization is suboptimal and we can improve upon (4.10). By Jensen's inequality, we find that averaging over forecasts based on different realizations of $R$ will lower the expected asymptotic MSFE compared to an individual forecast,

$$\mathrm{E}_\varepsilon \left[ \lim_{T\to\infty} T\mathrm{E}_{z_T} \left[ \left( y_{T+1} - \mathrm{E}_R \left[ \hat{y}_{T+1,R} \right] \right)^2 \right] \right] \leq \mathrm{E}_R \left[ \mathrm{E}_\varepsilon \left[ \lim_{T\to\infty} T\mathrm{E}_{z_T} \left[ \left( y_{T+1} - \hat{y}_{T+1,R} \right)^2 \right] \right] \right],$$

where $\mathrm{E}_R$ denotes the expectation with respect to the random matrix $R$. Therefore, we forecast $y_{T+1}$ as

$$\hat{y}_{T+1} = \mathrm{E}_R \left[ \hat{y}_{T+1,R} \right]. \tag{4.11}$$

In practice, we need to replace the expectation with a finite sum. In Section 4.3.2, we show that this does not affect the mean squared forecast error as long as the number of draws of $R$ is of $O\left(\frac{p_x \log p_x}{k}\right)$. This also implies that for a sufficient number of draws, forecasters that use a different sequence of random matrices will obtain the same forecast accuracy.

## 4.3 Theoretical results

The results in this section are based on the linear regression model defined in (4.1) and the following additional assumptions on the regressors $z_t$ and error terms $\varepsilon_{t+1}$. Consider the time index $t = 0, \ldots, T$, and the parameter index $i = 1, \ldots, p$. Denote by $\Delta$ a finite constant independent of the dimensions $p$ and $T$.

**Assumption 4.1** $\{z_t', \varepsilon_{t+1}\}$ *is a strong mixing sequence of size* $a = -r/(r-2)$, $r > 2$.

**Assumption 4.2** $E[\varepsilon_{t+1}|z_{ti}] = 0$.

**Assumption 4.3** $E[\varepsilon_{t+1}^2|z_{ti}] = \sigma^2$.

**Assumption 4.4** $E|z_{ti}\varepsilon_{t+1}|^r \leq \Delta < \infty$.

**Assumption 4.5** $E[z_t z_t'] = \Sigma_z = \begin{pmatrix} \Sigma_w & \Sigma_{wx} \\ \Sigma_{xw} & \Sigma_x \end{pmatrix}$ *is positive definite.*

**Assumption 4.6** *$V_n = Var(T^{-1/2}Z'\varepsilon)$ is uniformly positive definite.*

**Assumption 4.7** *$E|z_{ti}^2|^{r/2+\delta} \leq \Delta < \infty$.*

Under these assumptions we derive theoretical results that apply to weakly dependent time series models. In particular, they allow both $w_t$ and $x_t$ to contain lagged values of the dependent variable.

The mixing size $a$ in Assumption 4.1 is defined as in White (1984), Definition 3.42. In addition to standard results on asymptotic normality, the strong mixing assumption allows us to establish independence between $z_T$ and the estimation error $\sqrt{T}(\hat{\beta} - \beta)$, as we show in Appendix 4.A.1. This independence is essential to the proof of our main theorem. The necessity for this independence has been noted in Hansen (2008), and appears to be implied in Equation (2.2) of Hirano and Wright (2017).

Together, Assumptions A4.1-A4.7, guarantee that

$$\frac{1}{\sqrt{T}}Z'\varepsilon \xrightarrow{(d)} N(0, \sigma^2\Sigma_z), \quad \operatorname*{plim}_{T\to\infty} \frac{1}{T}Z'Z = \sigma^2\Sigma_z, \tag{4.12}$$

see for example White (1984).

We make one additional assumption with regard to the strength of the predictors, which rules out the possibility to consistently estimate $\beta$ as $T \to \infty$.

**Assumption 4.8** *The parameter vector $\beta$ is local-to-zero, i.e.*

$$\beta_x = \frac{1}{\sqrt{T}}\beta_{x,0}, \tag{4.13}$$

*where $\beta_{x,0} = O(1)$.*

Under local-to-zero coefficients, the bias induced by using a low-dimensional subspace is finite, see Claeskens and Hjort (2008). When coefficients are stronger than in Assumption A4.8, the forecast based on OLS estimation in (4.2) using $p$ variables is asymptotically the optimal forecast.

The theoretical results also suit forecasting models that assume a factor structure in $x_t$, such as the diffusion index model (Stock and Watson, 2002). In this case, if the factors are only weakly related to the dependent variable as in Assumption A4.8, the diffusion index model can be treated along the same lines as (4.1) upon replacing $x_t$ with $p_f$ common factors

in $f_t$. It is common to treat $p_f$ as fixed and let $p_x$ grow with $T$. The forecast error distribution for this model is derived by Bai and Ng (2006). Their results show that if $p_x/T \to \infty$, estimation of the factors does not affect the forecast distribution. If $p_x/T = O(1)$, an additive term enters due to the estimation error in the factors. This term is not affected by the methods in this chapter, so that the MSFE only incurs an additional term independent of $R$.

### 4.3.1 MSFE for forecasts from low-dimensional models

Denote the asymptotic mean squared forecast error of (4.11) as

$$\rho(k) = E_\varepsilon \left[ \lim_{T\to\infty} T\mathrm{E}_{z_T} \left[ \left( z_T'\beta - z_T'\mathrm{E}_R \left[ S_R \hat{\beta}_R \right] \right)^2 \right] \right]. \tag{4.14}$$

The following theorem provides a bound on the asymptotic mean squared forecast error for matrices $R$ which can be deterministic or random.

**Theorem 4.1** *Let $R \in \mathbb{R}^{p_x \times k}$ be a matrix such that $E_R[RR'] = \frac{k}{p_x} I_{p_x}$. The asymptotic mean squared forecast error $\rho(k)$ in* (4.14) *under* (4.1) *satisfying Assumption A4.1-A4.8, is upper bounded by*

$$\rho(k) \leq \sigma^2(p_w + k) + \beta_{x,0}'\Sigma_x\beta_{x,0} - \beta_{x,0}\Sigma_x \left( \frac{p_x}{k} E_R \left[ RR'\Sigma_x RR' \right] \frac{p_x}{k} \right)^{-1} \Sigma_x\beta_{x,0}. \tag{4.15}$$

A proof is presented in Appendix 4.A.2. Theorem 4.1 holds for general matrices $R$ after suitable scaling.

The first term of (4.15) represents the variance of the estimates. This can be compared to the variance that is achieved by forecasting using OLS estimates for $\beta$, which is equal to $\sigma^2 p = \sigma^2(p_w + p_x)$. In empirical applications, we expect $p_w$ to be small, as $w_t$ usually only contains a constant and a small number of lags. The number of additional variables $p_x$ can however be large, and hence, the reduction in variance to $k$ can be substantial.

The remaining terms in (4.15) reflect the bias that arises by projecting $x_t$ to a low-dimensional subspace. If any signal is present, this bias is strictly smaller then the bias of the naive estimator that does not use any of the predictors, which equals $\beta_{x,0}'\Sigma_x\beta_{x,0}$.

Loosely speaking, the product $\frac{p_x}{k} RR'\Sigma_x RR' \frac{p_x}{k}$ first projects $\Sigma_x$ to a $k$-dimensional subspace by multiplying with $R$ from the left and the right, and then re-inflates by another

multiplication with $R$. If little information is lost in this procedure, then the expectation will be close to $\Sigma_x$, in which case the bias is small.

For both random subset regression and random projection regression, the bound in (4.15) can be evaluated explicitly. We start with random subset regression.

### MSFE bound for random subset regression

For the random selection matrices in (4.6) we have the following result.

**Lemma 4.1** *Let $R \in \mathbb{R}^{p_x \times k}$ be a random selection matrix and $\Sigma_x$ a positive definite matrix. Then, $E_R[RR'] = \frac{k}{p_x} I_{p_x}$, and*

$$E_R\left[RR'\Sigma_x RR'\right] = \frac{k}{p_x}\left(\frac{k-1}{p_x-1}\Sigma_x + \frac{p_x-k}{p_x-1}D_{\Sigma_x}\right), \tag{4.16}$$

*where $[D_{\Sigma_x}]_{ii} = [\Sigma_x]_{ii}$, and $[D_{\Sigma_x}]_{ij} = 0$ if $i \neq j$.*

A proof is provided in Appendix 4.A.3.

Using Lemma 4.1 in the bound from Theorem 4.1, we obtain the following bound on the MSFE for random subset regression.

**Corollary 4.1** *For random subset regression, the asymptotic mean squared forecast error $\rho(k)$ in* (4.14) *under* (4.1) *satisfying Assumption A4.1-A4.8, is upper bounded by*

$$\rho(k) \leq \sigma^2(p_w + k) + \beta_{x,0}'\Sigma_x\beta_{x,0} - \frac{k}{p_x}\beta_{x,0}'\Sigma_x\left(\frac{k-1}{p_x-1}\Sigma_x + \frac{p_x-k}{p_x-1}D_{\Sigma_x}\right)^{-1}\Sigma_x\beta_{x,0}.$$

The bound for random subset regression depends on a convex combination $u\Sigma_x + (1-u)D_{\Sigma_x}$, for $0 \leq u \leq 1$. All weight is put on $D_{\Sigma_x}$ when $k = 1$, which implies that all information on cross-correlations is lost in the low-dimensional subspace. When $k = p_x$, the bound reduces to the exact expression for OLS using $p$ predictors as in (4.3).

### MSFE bound for random projection regression

When $R$ is constructed as in (4.7), the columns are not exactly orthogonal. Potentially, the lack of orthogonality of $R$ results in an unnecessary loss of information compared to the use of a $p_x \times k$ matrix $Q$ with orthogonal columns. However, the following lemma states that no such loss occurs.

**Lemma 4.2** *Suppose $R$ is a $p_x \times k$ matrix of independent standard normal random variables, $Q = R(R'R)^{-1/2}$ a $p_x \times k$ matrix with orthogonal columns, and $P = (R'R)^{1/2}$ an invertible $k \times k$ matrix, then*

$$\rho(k) = E_\varepsilon \left[ \lim_{T\to\infty} T E_{z_T} \left[ \left( z_T'\beta - z_T' E_Q \left[ S_Q \hat{\beta}_Q \right] \right)^2 \right] \right]. \tag{4.17}$$

A proof is provided in Appendix 4.A.4.

By Lemma 4.2 we can replace $R$ in Theorem 4.1 by $Q$, even though we are using $R$ in the construction of the estimator. To complete the bound from Theorem 4.1, we then need the following.

**Lemma 4.3** *Let $R \in \mathbb{R}^{p_x \times k}$ be a matrix of independent standard normal entries, and define $Q = R(R'R)^{-1/2} \in \mathbb{R}^{p_x \times k}$. Furthermore, let $\Sigma_x$ be a positive definite matrix. Then, $E_Q[QQ'] = \frac{k}{p_x} I_{p_x}$, and*

$$E_Q[QQ'\Sigma_x QQ'] = \frac{k}{p_x} \left( \frac{p_x(k+1)-2}{(p_x+2)(p_x-1)} \Sigma_x + \frac{(p_x-k)p_x}{(p_x+2)(p_x-1)} \frac{\mathit{trace}(\Sigma_x)}{p_x} I_{p_x} \right).$$

A proof is provided in Appendix 4.A.5, which relies on somewhat tedious calculations of the fourth order moments of the elements of the matrix $Q$.

Using Lemma 4.3 in the bound from Theorem 4.1, we obtain a bound on the asymptotic mean squared forecast error for random projection regression.

**Corollary 4.2** *For random projection regression, the asymptotic mean squared forecast error $\rho(k)$ in* (4.14) *under* (4.1) *satisfying Assumption A4.1-A4.8, is upper bounded by*

$$\begin{aligned}\rho(k) \leq\ & \sigma^2(p_w + k) + \beta_{x,0}'\Sigma_x\beta_{x,0} \\ & - \frac{k}{p_x}\beta_{x,0}'\Sigma_x \left( \frac{p_x(k+1)-2}{(p_x+2)(p_x-1)} \Sigma_x + \frac{(p_x-k)p_x}{(p_x+2)(p_x-1)} \frac{\mathit{trace}(\Sigma_x)}{p_x} I_{p_x} \right)^{-1} \Sigma_x \beta_{x,0}.\end{aligned}$$

The bound for random projection regression depends on a convex combination $u\Sigma_x + (1-u)\frac{\text{trace}(\Sigma_x)}{p_x}$. When $k = 1$, nearly all weight is put on trace$(\Sigma_x)$, while when $k = p_x$, all weight is put on $\Sigma_x$ and the bound reduces to (4.3).

Maillard and Munos (2009) provide a bound on the in-sample mean squared error under fixed regressors for random projection regression, which was subsequently improved by

Kabán (2014). Thanei et al. (2017) arrive at a similar expression as in (4.15), and use the expressions in Kabán (2014) to evaluate the expectation. However, their bound is suboptimal. For example, it has the unattractive feature of not reducing to (4.3) when $k$ is set equal to $p_x$. The bound in Corollary 4.2 solves this problem, by noting that we can rely on the matrix $Q$, which has orthogonal columns, instead of $R$ in the calculations. Appendix 4.A.6 shows that the resulting bound is uniformly tighter than the currently available bounds.

**Comparison between the MSFE bounds of RS and RP**

Based on the difference between the expressions for the MSFE bounds for random subset and random projection regression in Corollary 4.1 and 4.2, we show that there exists no covariance matrix $\Sigma_x$ for which one of the methods offers a superior bound uniformly over all possible parameter vectors $\beta_{x,0}$.

The difference in the bounds is given by

$$\Delta = \beta_{x,0}\Sigma_x \left(M_{RP}^{-1} - M_{RS}^{-1}\right) \Sigma_x \beta_{x,0}, \tag{4.18}$$

where

$$M_{RP} = \frac{k}{p_x}\left(\frac{k-1}{p_x-1}\Sigma_x + \frac{p_x-k}{p_x-1}D_{\Sigma_x}\right)^{-1},$$
$$M_{RS} = \frac{k}{p_x}\left(\frac{p_x(k+1)-2}{(p_x+2)(p_x-1)}\Sigma_x + \frac{(p_x-k)p_x}{(p_x+2)(p_x-1)}\frac{\text{trace}(\Sigma_x)}{p_x}I_{p_x}\right)^{-1}.$$

If $\Delta > 0$, then the bound for random projection regression lies above the bound for random subset regression. Denote $A-B \succ 0$ if $A-B$ is a positive definite matrix. If $M_{RP}^{-1} - M_{RS}^{-1} \succ 0$, then $\Delta > 0$ uniformly over the choice of $\beta_{x,0}$. This occurs if and only if $M_{RP} - M_{RS} \prec 0$, where

$$M_{RP} - M_{RS} = \frac{p_x-k}{p_x-1}\left[\frac{2}{p_x+2}\left(\Sigma_x - D_{\Sigma_x}\right) + \frac{p_x}{p_x+2}\left(\frac{\text{trace}(\Sigma_x)}{p_x}I_{p_x} - D_{\Sigma_x}\right)\right].$$

Unless $\Sigma_x$ is a multiple of the identity matrix, subtracting $D_{\Sigma_x}$ yields an indefinite matrix. This is easily seen as the sum of the eigenvalues of $\Sigma_x - D_{\Sigma_x}$ equals the trace, which is identically equal to zero. Similarly, unless $D_{\Sigma_x}$ is a multiple of the identity matrix, the second term yields an indefinite matrix. Hence, there does not exist a covariance matrix $\Sigma_x$

for which $M_{RP} - M_{RS} \prec 0$, and hence where one of the methods outperforms the other uniformly over the choice of $\beta_{x,0}$.

However, we can distinguish cases in which the random subspace methods are expected to perform equally well or outperform each other when we take the relation between the covariance matrix of the regressors and the coefficients of the regressors into account. We consider a simplified setting based on (4.1) with $\Sigma_x$ a diagonal $p_x \times p_x$ matrix, for which the bounds in Corollary 4.1 and 4.2 simplify to

$$\begin{aligned} \rho(k)^{RS} &\leq \sigma^2(p_w + k) + \beta'_{x,0}\Sigma_x\beta_{x,0}\left(1 - \frac{k}{p_x}\right), \\ \rho(k)^{RP} &\leq \sigma^2(p_w + k) + \beta'_{x,0}\Sigma_x\left[I_{p_x} - \frac{k}{p_x}D(\Sigma_x)\right]\beta_{x,0}, \end{aligned} \tag{4.19}$$

respectively, where

$$[D(\Sigma_x)]_{ii} = \frac{\lambda_i}{u\lambda_i + (1-u)\bar{\lambda}}, \quad u = \frac{p_x(k+1) - 2}{(p_x+2)(p_x-1)}, \quad \bar{\lambda} = \frac{1}{p_x}\sum_{i=1}^{p_x}\lambda_i, \tag{4.20}$$

where $\lambda_1, \ldots, \lambda_{p_x}$ are the eigenvalues of $\Sigma_x$ in decreasing order.

For a well-conditioned covariance matrix, i.e. $\lambda_i \approx \bar{\lambda}$ which means that the eigenvalues of the covariance matrix are of the same size, we have $D(\Sigma_x) \approx I$. From (4.19) we infer that in this scenario, the methods are expected to perform equally well.

When the eigenvalues of the covariance matrix of the regressors are not of the same size, two things can happen. First, consider a typical principal component regression setting where the nonzero values of $\beta_{x,0}$ are associated with eigenvalues that are larger than the average eigenvalue. For random projection regression, $[D(\Sigma_x)]_{ii} > 1$ when $\lambda_i > \bar{\lambda}$. Therefore random projection will offer a superior bound compared to random subset regression in this case. In this sense, the behavior of random projection regression appears similar to that of ridge regression, in that it performs most shrinkage on small eigenvalues.

In contrast, it is also possible that the factor associated with the largest eigenvalue of the covariance matrix is not associated with the dependent variable. Random subset regression does not assume that large eigenvalues in $\Sigma_x$ are informative on the relative importance with respect to $y$. Since in the bound for random projection regression it holds that $[D(\Sigma_x)]_{ii} < 1$ if $\lambda_i < \bar{\lambda}$, random subset regression now offers a superior bound.

### Comparison between the MSFE of RS, RP, and OLS

Here we study the performance of the random subspace methods relative to OLS for different signal strength, in the same setting as the previous section.

Based on the MSFE bound, we find that for small signal strength, random subset regression outperforms OLS. Equating the exact MSFE of OLS in (4.2) to the bound for the MSFE of random subset regression results in the following condition,

$$\frac{\beta_{x,0}'\Sigma_x\beta_{x,0}}{\sigma^2 p_x} = 1, \tag{4.21}$$

which implies that random subset regression outperforms OLS when the average signal strength falls below 1.

The relative performance of random projection regression to OLS depends not only on the signal strength, but also on which coefficients in $\beta_{x,0}$ are non-zero. Therefore, condition (4.21) does not apply to random projection regression. If non-zero coefficients are related to larger than average eigenvalues, the bound is lower than the MSFE under OLS as long as $\frac{\beta_{x,0}'\Sigma_x\beta_{x,0}}{\sigma^2 p_x} < 1 + u$, for some $u > 0$. When non-zero coefficients are related to smaller than average eigenvalues, we obtain $\frac{\beta_{x,0}'\Sigma_x\beta_{x,0}}{\sigma^2 p_x} < 1 - u$, for $u > 0$.

### Quality MSFE bounds

To provide insight in the quality of the bounds obtained in Corollary 4.1 and 4.2, we consider a setting in which we obtain an expression for the exact MSFE. For random subset regression this is achieved when the regressors are independent. If in addition we assume the variances of the regressors to be equal, we also obtain an exact expression under random projection regression.

When $\Sigma_z = D_{\Sigma_z}$, we have that $\mathrm{E}_R[S_R\hat{\beta}_R]$ in (4.14) boils down to

$$\mathrm{E}_R[S_R\hat{\beta}_R] = \begin{pmatrix} (W'W)^{-1}W' \\ \mathrm{E}_R[R(R'X'XR)^{-1}R']X' \end{pmatrix} y, \tag{4.22}$$

where $W = (w_0, \ldots, w_{T-1})'$ and $X = (x_0, \ldots, x_{T-1})'$. When $\frac{1}{T}X'X$ converges to a diagonal matrix, we can explicitly evaluate the expectation for random subset regression,

$$\mathrm{E}_R[R(R'D_{\Sigma_x}R)^{-1}R'] = \frac{k}{p_x}D_{\Sigma_x}^{-1}, \tag{4.23}$$

where $R$ is a random permutation matrix. This follows from the fact that each diagonal element of $\Sigma_x$ is selected with probability $k/p_x$ in random subset regression, see Appendix 4.A.3.

We obtain the same result for random projection regression under independent predictors with equal variance, $\Sigma_x = cI_{p_x}$,

$$c^{-1}\mathrm{E}_R[R(R'R)^{-1}R'] = c^{-1}\mathrm{E}_Q[QQ'] = \frac{k}{p_x}c^{-1}I_{p_x}, \tag{4.24}$$

where the expression for the second moment follows from Lemma 4.3.

Subsequently, the exact MSFE for both random subspace methods is given by

$$\rho(k) = \sigma^2\left(p_w + k\frac{k}{p_x}\right) + \beta_{x,0}'\Sigma_x\beta_{x,0}\left(1 - \frac{k}{p_x}\right)^2. \tag{4.25}$$

In case of independent regressors, the bounds in Corollary 4.1 and 4.2 simplify to (4.19). Since we assume $\Sigma_x = c \cdot I_{p_x}$ for the bound of random projection regression, $[D(\Lambda)]_{ii} = 1$, and the bounds of the random subspace methods are identical.

Comparing the exact MSFE from (4.25) to the bounds in (4.19), we see that the bounds overestimate the variance by a factor $p_x/k$, and the bias by a factor $(1 - k/p_x)^{-1}$. The difference is maximized for $\frac{k}{p_x} = \frac{1}{2}$ in which case the bounds are conservative by at most a factor $\frac{1}{2}$.

As an alternative to the upper bound on the MSFE in Theorem 4.1, the MSFE can be bounded by bounding the eigenvalues of the expectation over the random matrix $R$. Using the eigenvalue inequalities in Appendix 4.A.7, we derive both a conservative upper and lower bound on the MSFE in Appendix 4.A.8. Since these bounds ignore the eigenvalue structure of the covariance matrix of the predictors, these bounds are in almost all cases uninformative. Furthermore, the bounds are identical for random subset regression and random projection regression. They therefore do not elicit the difference between the two methods.

### 4.3.2 Feasibility of the MSFE bounds

The bounds from the previous section are based on forecasts that depend on the expectation over the random matrix $R$. In practice, we need to approximate this expectation by using a finite number of draws of the matrix $R$. For the feasibility of the method in practice, it is important that the required number of draws is not too large. If one would have to draw all possible subsets of size $k$ from $p_x$ predictors, the number of required draws is exponential in $p_x$, limiting the practical use of the methods. The following theorem guarantees that in order to get close to the expectation, we only require a number of draws that is linear in $p_x$, up to logarithmic factors.

**Theorem 4.2** *Let $\hat{y}_{T+1,S} = \frac{1}{N}\sum_{i=1}^{N} \hat{y}_{T+1,R_i}$, with $\hat{y}_{T+1,R_i}$ as in* (4.10) *where $R_i$ is a realization of the random matrix $R$, and $\hat{y}_{T+1}$ as in* (4.11)*. Denote by $\rho_S(k)$ the asymptotic mean squared forecast error based on $\hat{y}_{T+1,S}$, and denote by $\rho(k)$ the asymptotic mean squared forecast error based on $\hat{y}_{T+1}$ as in* (4.14)*. Furthermore, let $N = O\left(\frac{p_x \log p_x}{k}\right)$. Then for an arbitrarily small constant $\epsilon$,*

$$\rho_S(k) = (1+\epsilon)\rho(k). \tag{4.26}$$

A proof is provided in Appendix 4.A.9.

This result shows the feasibility of random subset regression in practice. It also provides a theoretical justification of the results obtained in Elliott et al. (2013) and Elliott et al. (2015), where it was found that little prediction accuracy is lost by using a relatively small number of random subsets instead of all available subsets. Instead of drawing a number of subsets exponential in $p_x$, $N = \binom{p_x}{k} = O\left(\left[\frac{p_x}{k}\right]^k\right)$, which is the case for complete subset regression, we only require a number of draws linear in $p_x$.

## 4.4 Monte Carlo experiments

We examine the practical implications of the theoretical results in a Monte Carlo experiment. In a first set of experiments we show the effect of sparsity and signal strength on the MSFE, and a second set of experiments shows in which settings one of the random subspace methods is preferred over the other. The prediction accuracy of the random subspace methods is evaluated relative to several widely used alternative regularization techniques.

### 4.4.1 Monte Carlo set-up

The set-up we employ parallels Elliott et al. (2015). The data generating process takes the form

$$y_{t+1} = x_t'\beta_x + \varepsilon_{t+1}, \tag{4.27}$$

where $x_t$ is a $p_x \times 1$ vector with predictors, $\beta_x$ a $p_x \times 1$ coefficient vector, $\varepsilon_{t+1}$ an error term with $\varepsilon_{t+1} \sim N(0, \sigma_\varepsilon^2)$, and $t = 0, \ldots, T$. In each replication of the Monte Carlo simulations, predictors are generated by drawing $x_t \sim N(0, \Sigma_x)$, after which we standardize the predictor matrix. The covariance matrix of the predictors equals $\Sigma_x = \frac{1}{p_x} P'P$, where $P$ is a $p_x \times p_x$ matrix whose elements are independently and randomly drawn from a standard normal distribution. As argued by Elliott et al. (2015), this ensures that the eigenvalues of the covariance matrix are reasonably spaced.

The strength of the individual predictors is considered local-to-zero by setting $\beta_x = \sqrt{\sigma_\varepsilon^2/T} \cdot b\iota_s$ for a fixed constant $b$. The vector $\iota_s$ contains $s$ non-zero elements that are equal to one. We refer to $s$ as the sparsity of the coefficient vector. We vary the signal strength $b$ and the sparsity $s$ across different Monte Carlo experiments. In all experiments, the error term of the forecast period $\varepsilon_{T+1}$ is set to zero, as this only yields an additional noise term $\sigma^2$ which is incurred by all forecasting methods.

We employ two sets of experimental designs, which mimic the high-dimensional setting in the empirical application by choosing the number of predictors $p_x = 100$ and the sample size $T = 200$. Results are based on $M = 10{,}000$ replications of the data generating process (4.27).

In the first set of experiments, we vary the signal to noise ratio $b$ and the sparsity $s$ over the grids $b \in \{0.5, 1.0, 2.0\}$ and $s \in \{10, 50, 100\}$. This allows us to study the effect of sparsity and signal strength on the MSFE and the optimal subspace dimension.

The second set of experiments reflects scenarios where random subset and random projection regression are expected to differ based on the discussion in Section 4.3.1. In this case we replace $x_t$ in (4.27) by factors extracted from $x_t$, $t = 0, \ldots, T$, using principal component analysis. Denote by $f_i$ for $i = 1, \ldots, p_x$ the extracted factors sorted by the explained variation in the predictors. In the first three experiments, we associate nonzero coefficients with the 10 factors that explain most of the variation in the predictors. We refer to this setting as

the top factor setting. This setting is expected to suit random projection over random subset regression. In the remaining experiments, we associate the nonzero coefficients with factors $\{f_{46}, \ldots, f_{55}\}$, which are associated with intermediately sized eigenvalues. This setting is referred to as the intermediate factor setting and expected to suit random subset regression particularly well. In both the top and intermediate factor setting, the coefficient strength $b$ is again varied as $b \in \{0.5, 1.0, 2.0\}$.

We generate one-step-ahead forecasts by means of random projection and random subset regression using equation (4.4) in which we vary the subspace dimension over $k = \{1, \ldots, p_x\}$. The subspace methods, as well as the benchmark models discussed below, estimate (4.27) with the inclusion of an intercept that is not subject to the dimension reduction or shrinkage procedure. We average over $N = 1{,}000$ predictions of the random subspace methods to arrive at a one-step-ahead forecast. This is in line with the findings in Section 4.3.2 which suggest to use $O(p_x \log p_x) = O(100 \cdot \log 100) = O(460)$ draws.

**Benchmark models** We compare the performance of the random methods with principal component (PC) regression and partial least squares (PL) regression introduced by Wold (1982). Both methods approximate the data generating process (4.27) as

$$y_{t+1} = w_t' \beta_w + \sum_{i=1}^{k} f_{ti} \beta_{f,i} + \eta_t, \tag{4.28}$$

where $k \in \{1, \ldots, p_x\}$ and $w_t$ includes an intercept. The methods differ in their construction of the factors $f_{ti}$. Principal component regression is implemented by extracting the factors from the standardized predictors $x_t$ with $t = 0, \ldots, T$ using principal component analysis. This is a diffusion index model along the lines of Stock and Watson (2002). Partial least squares uses a two-step procedure to construct the factors, as described for example by Groen and Kapetanios (2016). We use the static approach as discussed by Fuentes et al. (2015), who find good forecast performance for a similar macroeconomic forecasting exercise as in Section 4.5, in which the factors are extracted by applying partial least squares between the target variable $y_{t+1}$ and the predictors $x_t$. We then estimate for both methods (4.28) and generate a forecast as $\hat{y}_{T+1} = w_T' \hat{\beta}_w + \sum_{i=1}^{k} f_{Ti} \hat{\beta}_{f,i}$. Note that the principal component regression model is correctly specified for the top factor setting in the second set of experiments.

In addition to comparing the random subspace methods to principal component regression and partial least squares, we include two widely used alternatives: ridge (RI) regression (Hoerl and Kennard, 1970) and the lasso (LA) (Tibshirani, 1996). We generate one-step-ahead forecasts using these methods by $\hat{y}_{T+1} = w_T'\hat{\beta}_w + x_T'\hat{\beta}_x$, with

$$(\hat{\beta}_w, \hat{\beta}_x) = \arg\min_{\beta_w, \beta_x} \left( \frac{1}{n} \sum_{t=0}^{T-1} (y_{t+1} - w_t'\beta_w - x_t'\beta_x)^2 + kP(\beta_x) \right). \tag{4.29}$$

The penalty term $P(\beta_x) = \sum_{i=1}^{p_x} \frac{1}{2}\beta_{x,i}^2$ in case of ridge regression and $P(\beta_x) = \sum_{i=1}^{p_x} |\beta_{x,i}|$ for the lasso. The penalty parameter $k$ controls the amount of shrinkage. In contrast to the previous subspace methods, the values of $k$ are not bounded to integers nor is there a natural grid. We consider forecasts based on equally spaced grids for $\ln k$ of 100 values; $\ln k \in \{-30, \ldots, 0\}$ for lasso and $\ln k \in \{-15, \ldots, 15\}$ for ridge regression. In general, we expect lasso to do well when the model contains a small number of large coefficients. Ridge regression, on the other hand, is expected to do well when we have many weak predictors.

**Evaluation criterion** We evaluate forecasts by reporting their MSFE relative to that of the prevailing mean model that takes $\bar{y}_{T+1} = \frac{1}{T}\sum_{t=0}^{T-1} y_{t+1}$. The mean squared forecast error is computed as

$$\textit{MSFE} = \frac{1}{M} \sum_{j=1}^{M} \left( y_{T+1}^{(j)} - \hat{y}_{T+1}^{(j)} \right)^2, \tag{4.30}$$

where $y_{T+1}^{(j)}$ is the realized value and $\hat{y}_{T+1}^{(j)}$ the predicted value in the $j$th replication of the Monte Carlo simulation. The number of replications $M$ is set equal to $M = 10{,}000$.

## 4.4.2 Simulation results

### Sparsity and signal strength

Table 4.1 shows the Monte Carlo simulation results for the first set of experiments for the value of $k$ that yields the lowest MSFE. Results for different values of $k$ are provided in Table 4.7 in Appendix 4.B. The predictive performance of each forecasting method is reported relative to the prevailing mean. Values below one indicate that the benchmark model is outperformed.

**Table 4.1:** Simulation results: MSFE optimal subspace dimension

| $b$ | RP | RS | PC | PL | RI | LA |
|---|---|---|---|---|---|---|
| | | | $s = 10$ | | | |
| 0.5 | 0.967 (2) | 0.966 (2) | 1.253 (1) | 9.592 (1) | 0.966 (-3.3) | 1.000 (-29.7) |
| 1.0 | 0.864 (8) | 0.865 (8) | 1.056 (1) | 3.099 (1) | 0.864 (-2.1) | 0.959 (-27.9) |
| 2.0 | 0.638 (21) | 0.637 (21) | 0.929 (7) | 0.961 (1) | 0.640 (-0.6) | 0.669 (-27.3) |
| | | | $s = 50$ | | | |
| 0.5 | 0.815 (10) | 0.815 (10) | 1.034 (1) | 2.377 (1) | 0.814 (-1.8) | 0.961 (-27.9) |
| 1.0 | 0.568 (25) | 0.569 (25) | 0.885 (12) | 0.805 (1) | 0.570 (-0.6) | 0.706 (-27.3) |
| 2.0 | 0.300 (46) | 0.301 (46) | 0.453 (43) | 0.374 (2) | 0.301 (0.6) | 0.366 (-26.4) |
| | | | $s = 100$ | | | |
| 0.5 | 0.710 (16) | 0.709 (16) | 0.980 (2) | 1.372 (1) | 0.710 (-1.2) | 0.877 (-27.6) |
| 1.0 | 0.422 (36) | 0.423 (35) | 0.663 (29) | 0.535 (1) | 0.423 (0.0) | 0.539 (-26.7) |
| 2.0 | 0.188 (56) | 0.189 (56) | 0.268 (59) | 0.227 (3) | 0.189 (1.2) | 0.242 (-26.1) |

Note: this table reports the MSFE relative to the benchmark of the prevailing mean, for the subspace dimension corresponding to the minimum MSFE which is given in parentheses.

We find that in general, a lower degree of sparsity results in a lower relative MSFE. Since the predictability increases in $s$, it is not surprising that a less sparse setting results in better forecast performance relative to the prevailing mean, which ignores all information in the predictors. Similarly, the prediction accuracy also clearly increases with increasing signal strength. The results for different values of $k$, reported in Table 4.7 in Appendix 4.B, show that increasing the subspace dimension in case of a weak signal worsens the performance, due to the increasing effect of the parameter estimation error when the predictive signal is small. This dependency on $k$ tends to decrease for large values of $s$ and $b$, where we observe smaller differences between the predictive performance over the different values of $k$.

Comparing the random subspace methods, we find that in these experiments, as expected, the predictive performance of random projection regression and random subset regression is almost the same. Table 4.1 shows that when choosing the optimal subspace dimension, these methods outperform both the prevailing mean as principal component regression and partial least squares for each setting. Lasso is not found to perform well. Only in the extremely sparse settings where $s = 10$ and $b$ increases, its performance tends towards the random subspace methods. Ridge regression yields similar prediction accuracy as the random subspace methods. For strong signals, the random subspace methods perform better, whereas for very weak signals ridge regression appears to have a slight edge.

**Table 4.2:** Simulation results: MSFE optimal subspace dimension - factor design

| $b$ | RP | RS | PC | PL | RI | LA |
|---|---|---|---|---|---|---|
| | Top factor setting | | | | | |
| 0.5 | 0.722 (10) | 0.955 (9) | 0.992 (2) | 2.466 (1) | 0.721 (-1.8) | 0.887 (-27.9) |
| 1.0 | 0.428 (21) | 0.842 (28) | 0.300 (10) | 0.495 (1) | 0.429 (-0.9) | 0.485 (-27.6) |
| 2.0 | 0.205 (33) | 0.580 (60) | 0.078 (10) | 0.139 (1) | 0.206 (0.0) | 0.150 (-27.3) |
| | Intermediate factor setting | | | | | |
| 0.5 | 1.013 (1) | 0.998 (1) | 1.501 (1) | 16.347 (1) | 1.000 (-14.7) | 1.000 (-29.7) |
| 1.0 | 1.003 (1) | 0.981 (4) | 1.176 (1) | 7.140 (1) | 1.000 (-7.5) | 1.000 (-29.1) |
| 2.0 | 1.001 (1) | 0.923 (16) | 1.060 (1) | 2.969 (1) | 1.000 (-14.7) | 1.000 (-29.7) |

Note: this table reports the MSFE relative to the benchmark of the prevailing mean, for the subspace dimension corresponding to the minimum MSFE which is given in parentheses.

Table 4.1 shows that the optimal subspace dimension increases with both the sparsity $s$ and the signal strength governed by $b$. Interestingly, random subset regression and random projection regression select, apart from one setting, exactly the same subspace dimension. The number of factors selected in principal component regression is lower for almost all settings. The results for partial least squares reflect that in settings with a small number of weak predictors, the factors cannot be constructed with sufficient accuracy. In these settings, more accurate forecasts are therefore obtained by ignoring the factors altogether. Note that where the parameter $k$ has an intuitive appeal in the dimension reduction methods, the values in the grid of $k$ for lasso and ridge regression methods lack interpretation.

### Experiments using a factor design

The small differences between random subset and random projection regression in the previous experiments stand in stark contrast with the findings on the factor structured experiments. The relative MSFE for the choice of $k$ that yields the lowest MSFE compared to the prevailing mean is reported in Table 4.2. Table 4.8 in in Appendix 4.B shows results for different values of $k$. We observe precisely what was anticipated based on the discussion in Section 4.3.1. In the top factor setting, where the nonzero coefficients are associated with the factors corresponding to the largest 10 eigenvalues, random projection regression outperforms random subset regression by a wide margin. For a weak signal, when $b = 0.5$, it even outperforms principal component regression, which is correctly specified in this set-up. When

$b = 2$, we are in a setting where we have a small number of large coefficients. As expected, this favors lasso, although not to the extend that it outperforms principal component regression. The findings are almost completely reversed in the intermediate factor setting, when the nonzero coefficients are associated with factors $f_{46}, \ldots, f_{55}$. Here we observe that random subset regression outperforms random projection. In fact, random subset regression is the only method that is able to extract an informative signal from the predictors and outperform the prevailing mean benchmark.

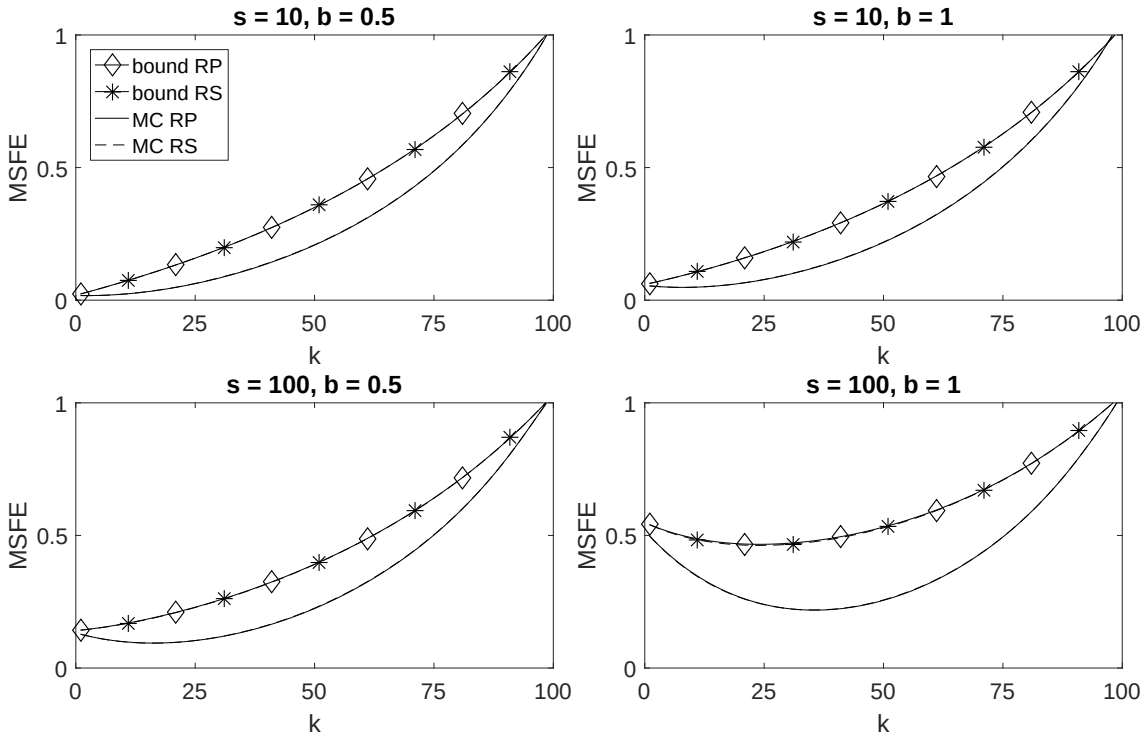
The difference in predictive performance is reflected in the optimal subspace dimension reported in parentheses in Table 4.2. For the top factor setting, when $b = \{1, 2\}$, we observe that the MSFE for random subset regression is minimized at substantially larger values than for random projection regression. This evidently increases the forecast error variance, and the added predictive content is apparently too small to outweigh this. Principal component regression, in turn, selects the correct number of factors when $b = \{1, 2\}$. In the intermediate factor setting, the dimension of random subset is again larger than for random projection, with an impressive difference when $b = 2$. Here, random projection is apparently not capable to pick up any signal and selects $k = 1$, while random subset regression uses a subspace dimension of $k = 16$. Lasso and ridge both choose such a strong penalization that they reduce to the prevailing mean benchmark for all choices of $b$.

### 4.4.3 Simulation results versus theoretical bounds

The qualitative correspondence between the simulation results and the theoretical results show that the bounds are useful to determine settings where the random subspace methods are expected to do well. In this section, we investigate how close the bounds are to the exact MSFE obtained in the Monte Carlo experiments.

Figure 4.1 shows the MSFE over different subspace dimensions of random projection and random subset regression, along with the theoretical upper bounds on the MSFE derived in Section 4.3.1, for the first set of experiments described above. As we found in Table 4.7 in Appendix 4.B, the values of the MSFE of the random subspace methods are almost identical to each other over the whole range of $k$. This also holds for the bounds. The bounds differ most from the exact MSFE from the Monte Carlo experiments for intermediate values of $k$ when there is a strong signal and no sparsity.

**Figure 4.1:** Simulation results: comparison with theoretical bounds



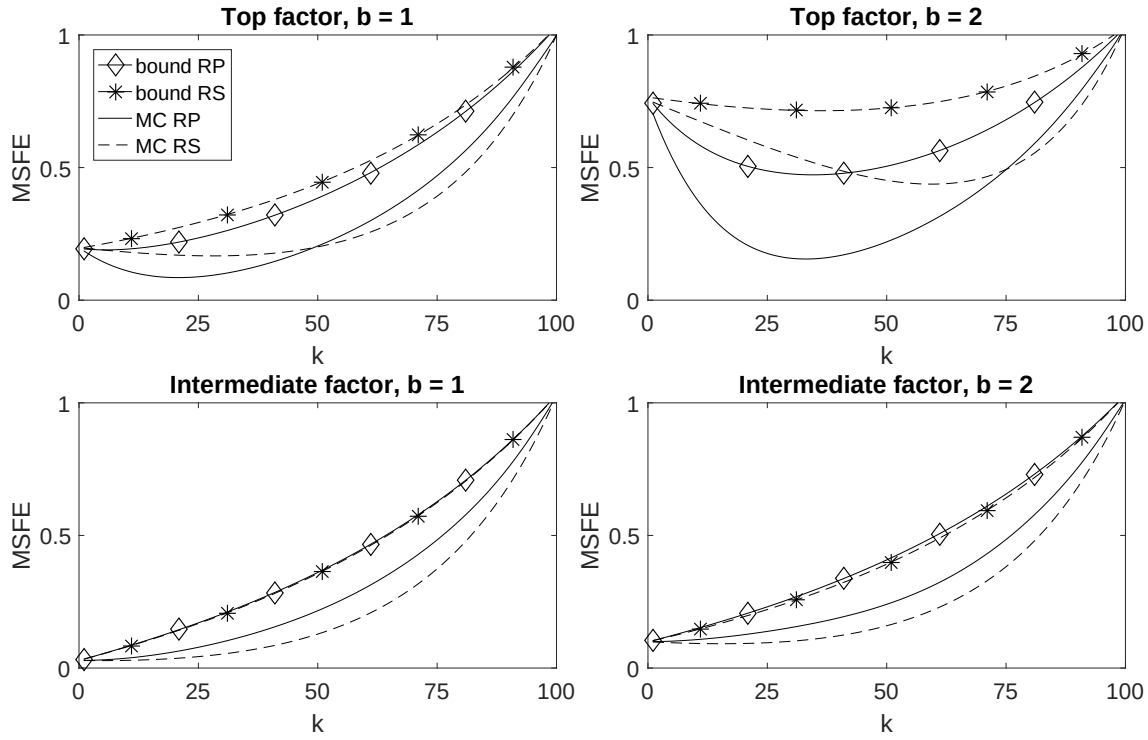

Note: this figure shows the MSFE for different values of the subspace dimension $k$, along with the theoretical upper bounds on the MSFE derived in Section 4.3.1 after a small sample size correction. The different lines correspond to the upper bound for random projections (bound RP, diamond marker), upper bound for random subsets (bound RS, asterisk marker), and the evaluation criteria for the dimension reduction methods random projections (MC RP, solid) and random subsets (MC RS, dashed). The four panels correspond to settings in which the sparsity $s$ alternates between 10 and 100, and the signal to noise ratio parameter $b$ between 0.5 and 1.

In Figure 4.2 we show the bounds for the factor settings. Here we see that the bounds correctly indicate which method is expected to yield better results in the settings under consideration. The upper panels, corresponding to the top factor structure, show the bound for random projection to be lower. The lower panels display the MSFE in the intermediate factor setting. We observe that both the bounds and the exact simulation results indicate that random subset regression is best suited in this case.

## 4.5 Empirical application

This section evaluates the forecast performance of the random subspace methods in a macroeconomic application.

**Figure 4.2:** Simulation results: comparison with theoretical bounds - factor design




Note: this figure shows the MSFE for different values of the subspace dimension $k$, along with the theoretical upper bounds on the MSFE derived in Section 4.3.1 for the top and intermediate factor settings. For additional information, see the note following Figure 4.1.

### 4.5.1 Data

We use the FRED-MD database consisting of 130 monthly macroeconomic and financial series running from January 1960 through December 2014. The data can be grouped in eight different categories: output and income (1), labor market (2), consumption and orders (3), orders and inventories (4), money and credit (5), interest rate and exchange rates (6), prices (7), and stock market (8). The data is available from the website of the Federal Reserve Bank of St. Louis, together with code for transforming the series to render them stationary and to remove severe outliers. The data and transformations are described in detail by McCracken and Ng (2016). After transformation, we find a small number of missing values, which are recursively replaced by the value in the previous time period of that variable. The FRED-MD can be seen as an updated version of the Stock and Watson (2005) dataset. For completeness, Section 4.5.5 also applies the methods to the original Stock and Watson (2005) data.

### 4.5.2 Forecasting framework

We generate forecasts for each of the 130 macroeconomic time series using the following equation

$$y_{t+1} = w_t'\beta_w + x_t'\beta_x + \varepsilon_{t+1},$$

where $w_t$ is a $p_w \times 1$ vector with predictors which are always included in the model and not subject to the dimension reduction methods, and $x_t$ a $p_x \times 1$ vector with possible predictors.

We follow Bai and Ng (2008) in considering up to six lags of the dependent variable and evaluating the forecast performance relative to an AR(4) model. The dependent variable $y_{t+1}$ is one of the macroeconomic time series, $w_t$ includes an intercept and the first four lags of the dependent variable $y_{t+1}$, and $x_t$ consists of the fifth and sixth lag of $y_{t+1}$, and all 129 remaining variables in the database. In Section 4.5.4, $x_t$ also includes the second up to the sixth lag of the 129 remaining variables in the database.

We apply dimension reduction to the predictors in $x_t$ using four different methods: random projection regression (RP), random subset regression (RS), principal component regression (PC), and partial least squares (PL). In addition, we compare the performance to lasso (LA) and ridge regression (RI) as described in Section 4.4.1. Predictive accuracy is measured by the MSFE defined in (4.30).

We standardize the predictors in each estimation window. In case of RP and RS we average over $N = 1{,}000$ forecasts to obtain one prediction. In some cases, random subset regression encounters substantial multicollinearity between the original predictors. Insofar this leads to estimation issues due to imprecise matrix inversion, these are discarded from the average. The models generate forecasts with subspace dimension $k$ running from 0 to 100 and, as in Elliott et al. (2013), we recursively select the optimal $k$ based on past predictive performance, using a burn-in period of 60 observations. Note that when $k = 0$, no additional predictors are included and we estimate an AR(4) model.

We use an expanding window to produce 420 forecasts, from January 1980 to December 2014. Due to the burn-in period, the initial estimation sample runs from January 1960 to December 1975 and contains 180 observations, from which we discard the first six observations to estimate the lags. This is larger than the initial estimation sample in, for instance, Bai and

**Table 4.3:** FRED-MD: percentage best forecast performance

| | | percentage loss | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | RP | RS | PC | PL | RI | LA | AR | All |
| percentage wins | RP | | 40.77 | 86.15 | 80.77 | 57.69 | 65.38 | 85.38 | 17.69 |
| | RS | 56.92 | | 89.23 | 81.54 | 66.92 | 70.77 | 83.85 | 40.00 |
| | PC | 11.54 | 8.46 | | 47.69 | 12.31 | 29.23 | 69.23 | 3.85 |
| | PL | 17.69 | 16.92 | 50.77 | | 21.54 | 30.00 | 63.85 | 7.69 |
| | RI | 42.31 | 33.08 | 87.69 | 78.46 | | 60.77 | 84.62 | 4.62 |
| | LA | 34.62 | 29.23 | 70.77 | 70.00 | 39.23 | | 80.77 | 18.46 |
| | AR | 14.62 | 16.15 | 30.77 | 32.31 | 15.38 | 19.23 | | 7.69 |

Note: this table shows the percentage wins in terms of lowest MSFE of the method listed in the rows over the method listed in the columns, and with respect to all other methods (last column). The percentages are calculated over forecasts for all 130 series in FRED-MD. Ties occur if only $k = 0$ is selected by both methods throughout the evaluation period, which is why losses and wins do not necessarily add up to 100.

Ng (2008), since the theory requires the number of variables $p_w + p_x = 136$ to be smaller than the sample size $T$.

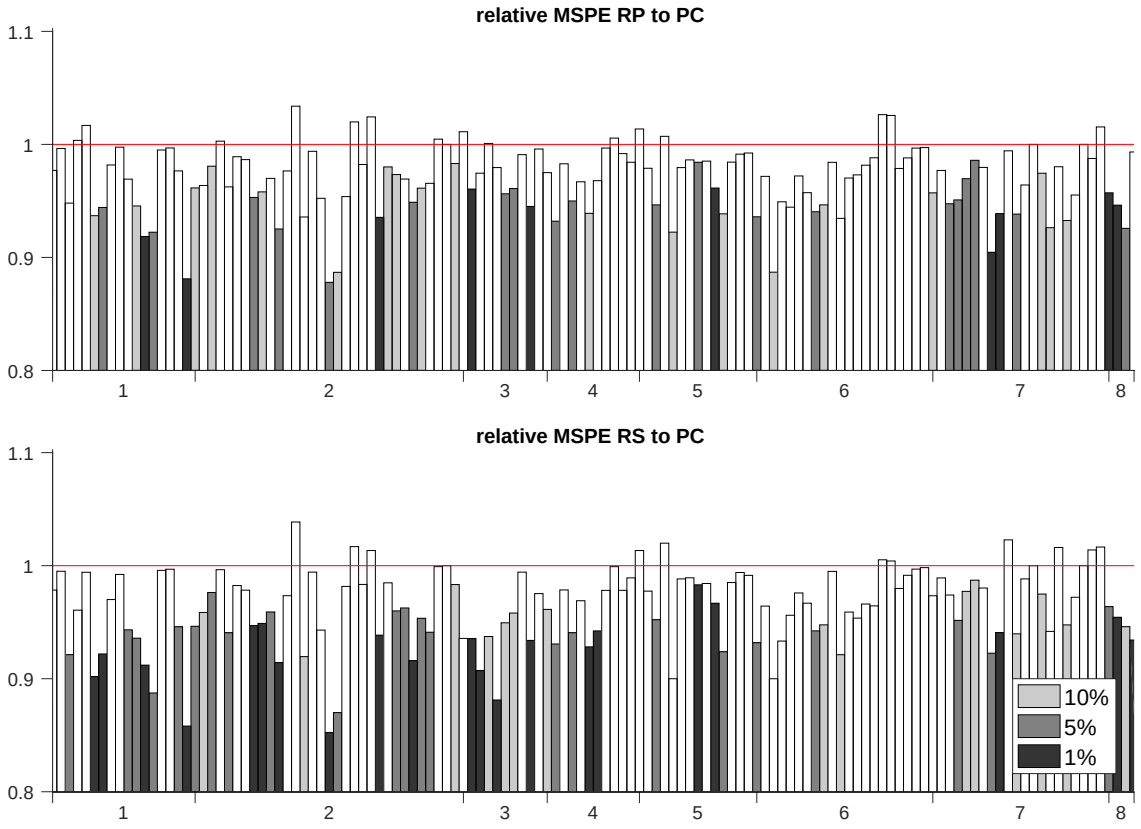
We report aggregate statistics over all 130 series, as well as detailed results for 4 major macroeconomic indicators out of the 130 series; industrial production index (INDP), unemployment rate (UNR), inflation (CPI), and the three-month Treasury Bill rate (3mTB). These series correspond to the FRED mnemonics INDPRO, UNRATE, CPIAUCSL, and TB3MS, respectively.

### 4.5.3 Empirical results

#### Aggregate statistics

We obtain series of forecasts for 130 macroeconomic variables generated by seven different methods. Table 4.3 shows the percentage wins of a method in terms of lowest MSFE compared to each of the other methods. The last column reports the percentage of the series for which a method outperforms all other methods. We find that random subset regression is more accurate than the other methods for 40% of the series. This is a substantial difference with random projections and lasso that win in approximately 18% of the cases. Principal component regression, partial least squares, ridge regression, and the AR(4) model score at most 8%.

**Figure 4.3:** FRED-MD: forecast accuracy relative to principal component regression




Note: this figure shows the MSFE of the forecasts for all series in the FRED-MD dataset produced by random projection regression (upper panel) and random subset regression (lower panel), scaled by the MSFE of principal component regression. Series are grouped in different macroeconomic indicators as described in McCracken and Ng (2016). Values below one prefer the method over principal components. Colors of the bars different from white indicate that the difference from one is significant at the 10% level (grey), 5% level (dark-grey), or 1% level (black), based on a two-sided Diebold-Mariano test.

If a model is the second most accurate on all series, this cannot be observed in the overall comparison. For this reason, we analyze the relative performance of the methods in a bivariate comparison. Table 4.3 shows again that random subset regression achieves the best results, outperforming the benchmark models for at least 66% of the series. Interestingly, a close competitor is random projection, which itself is also more accurate than all five benchmarks for a majority of the series. Out of the benchmark models, ridge regression appears closest to random subset regression, which is nevertheless outperformed for more than 66% of the series.

In addition to the ranking of the methods, we are also interested in the relative MSFE of the methods. To get an overview of the forecast performance of the random subspace methods sorted by category, Figure 4.3 shows relative forecast performance compared with

principal component regression, for all series available in the FRED-MD dataset. The MSFE is calculated for the subspace dimension as determined by past predictive performance. The upper panel shows the relative MSFE of random projection regression to principal component regression and the lower panel compares random subset to principal component regression. Values below one, indicate that the random method is preferred over the benchmark. As found in Table 4.3, the random subspace methods outperform the principal components in most of the cases. For random subset regression this happens in 89% of the cases, which is slightly lower for random projections with 86%. Figure 4.3 also shows the significance of the differences between the methods. The color of the bar indicates significance as determined by a Diebold and Mariano (1995) test. We see that for series where principal component regression is more accurate, the difference with the random methods is never significant, even at a 10% level. Random projection regression shows the largest improvements in forecast performance in category 7, including price indicators, and random subset regression in category 1 and 2, which contain output, income, and labor market.

Principal component regression is known for its good forecast performance in the presence of instabilities in the data (Rossi, 2013). However, the principal components are outperformed for almost all macroeconomic variables, indicating that random subspace methods are not disproportionately affected by these instabilities.

**A case study of four key macroeconomic indicators**

We look more closely into the forecast accuracy of the different methods for four key macroeconomic indicators: industrial production index (INDP), unemployment rate (UNR), inflation (CPI), and the three-month Treasury Bill rate (3TB). In Table 4.4 we show the MSFE relative to the AR(4) model for different values of the subspace dimension or penalty parameter $k$. The first row of each panel shows the relative MSFE corresponding to the recursively selected optimal value of $k$, denoted by $k_R$. The last column of each panel shows the average relative MSFE over all series.

Consistent with our previous findings, random subset regression performs best over all series when the optimal subspace dimension is selected. However, some differences are observed when analyzing the four individual series. For predicting inflation and the treasury bill rate, random projection yields a lower MSFE compared to random subset regression.

**Table 4.4:** FRED-MD: forecast accuracy relative to the AR(4)-model

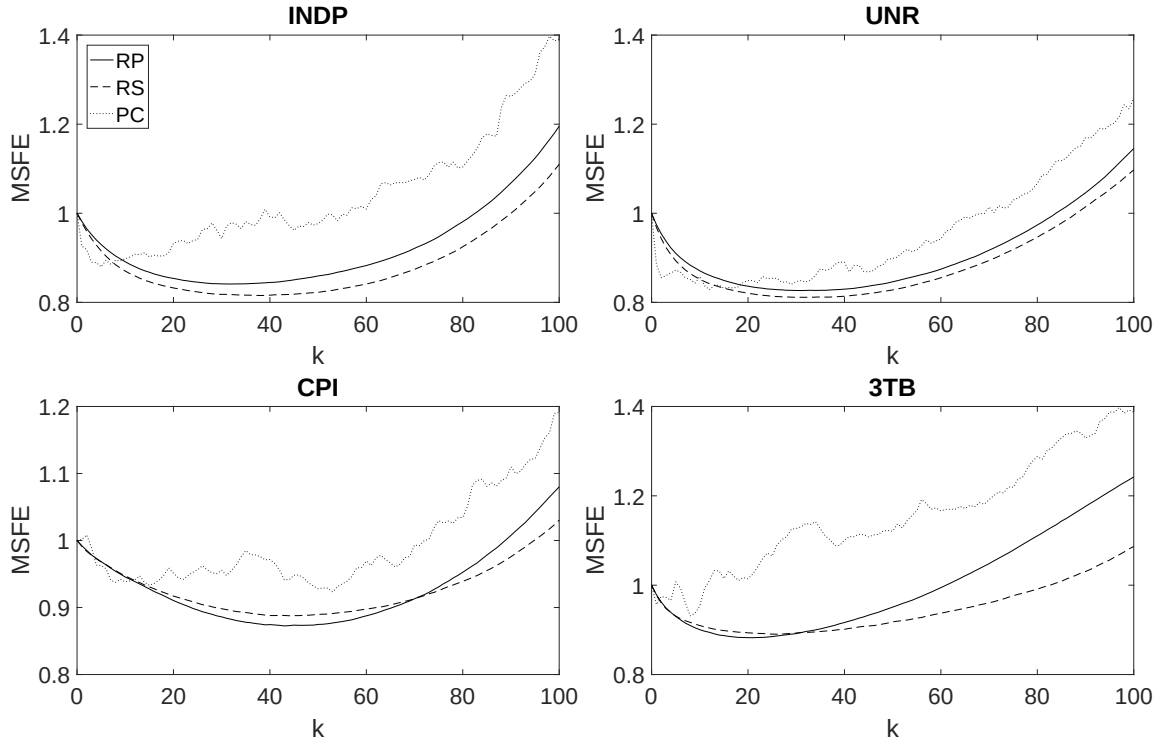
| | INDP | UNR | CPI | 3TB | Avg. | | INDP | UNR | CPI | 3TB | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $k$ | Random projection regression | | | | | $k$ | Random subset regression | | | | |
| $k_R$ | 0.843 | 0.842 | 0.870 | 0.892 | 0.929 | $k_R$ | 0.820 | 0.823 | 0.888 | 0.906 | 0.923 |
| 1 | 0.982 | 0.975 | 0.992 | 0.979 | 0.987 | 1 | 0.978 | 0.968 | 0.991 | 0.977 | 0.984 |
| 5 | 0.930 | 0.910 | 0.968 | 0.930 | 0.955 | 5 | 0.916 | 0.894 | 0.968 | 0.931 | 0.948 |
| 10 | 0.891 | 0.871 | 0.945 | 0.900 | 0.937 | 10 | 0.870 | 0.852 | 0.947 | 0.910 | 0.931 |
| 15 | 0.868 | 0.849 | 0.928 | 0.887 | 0.930 | 15 | 0.846 | 0.832 | 0.931 | 0.898 | 0.925 |
| 30 | 0.841 | 0.827 | 0.886 | 0.892 | 0.937 | 30 | 0.818 | 0.811 | 0.898 | 0.893 | 0.929 |
| 50 | 0.859 | 0.846 | 0.875 | 0.951 | 0.983 | 50 | 0.822 | 0.828 | 0.890 | 0.918 | 0.966 |
| 100 | 1.195 | 1.145 | 1.080 | 1.242 | 1.309 | 100 | 1.110 | 1.097 | 1.030 | 1.087 | 1.245 |
| $k$ | Principal component regression | | | | | $k$ | Partial least squares | | | | |
| $k_R$ | 0.890 | 0.875 | 0.962 | 1.006 | 0.959 | $k_R$ | 0.898 | 0.891 | 0.872 | 0.945 | 0.965 |
| 1 | 0.926 | 0.886 | 1.002 | 0.956 | 0.972 | 1 | 0.907 | 0.856 | 0.987 | 0.938 | 0.973 |
| 5 | 0.880 | 0.872 | 0.963 | 1.008 | 0.957 | 5 | 1.009 | 0.925 | 0.928 | 1.152 | 1.108 |
| 10 | 0.898 | 0.858 | 0.938 | 0.954 | 0.968 | 10 | 1.173 | 1.111 | 0.993 | 1.253 | 1.273 |
| 15 | 0.902 | 0.832 | 0.933 | 1.015 | 0.977 | 15 | 1.272 | 1.209 | 1.074 | 1.354 | 1.378 |
| 30 | 0.943 | 0.847 | 0.956 | 1.127 | 1.030 | 30 | 1.429 | 1.344 | 1.168 | 1.432 | 1.511 |
| 50 | 0.977 | 0.898 | 0.928 | 1.121 | 1.107 | 50 | 1.465 | 1.357 | 1.180 | 1.423 | 1.546 |
| 100 | 1.390 | 1.258 | 1.191 | 1.387 | 1.469 | 100 | 1.521 | 1.369 | 1.185 | 1.414 | 1.560 |
| $\ln k$ | Ridge regression | | | | | $\ln k$ | Lasso | | | | |
| $k_R$ | 0.844 | 0.842 | 0.901 | 0.900 | 0.930 | $k_R$ | 0.826 | 0.848 | 0.897 | 0.894 | 0.935 |
| -6 | 0.993 | 0.990 | 0.997 | 0.991 | 0.995 | -28 | 0.864 | 0.846 | 0.920 | 0.894 | 0.947 |
| -4 | 0.966 | 0.952 | 0.984 | 0.959 | 0.975 | -27 | 0.831 | 0.830 | 0.880 | 0.927 | 0.949 |
| -2 | 0.880 | 0.859 | 0.935 | 0.896 | 0.933 | -26 | 0.887 | 0.898 | 0.902 | 1.022 | 1.022 |
| 0 | 0.847 | 0.832 | 0.869 | 0.930 | 0.961 | -25 | 1.005 | 1.014 | 0.975 | 1.156 | 1.148 |
| 4 | 0.946 | 0.946 | 0.931 | 1.080 | 1.099 | -22 | 1.273 | 1.229 | 1.113 | 1.254 | 1.358 |
| 8 | 1.216 | 1.173 | 1.102 | 1.261 | 1.340 | -15 | 1.666 | 1.520 | 1.277 | 1.389 | 1.644 |
| 12 | 1.463 | 1.361 | 1.226 | 1.334 | 1.532 | -5 | 1.841 | 1.651 | 1.370 | 1.484 | 1.788 |

Note: this table shows the relative MSFE, which equals values below one when the particular method outperforms the benchmark AR(4) model, for different values of subspace dimension $k$ and the recursively selected optimal value of $k$ denoted by $k_R$. For ridge regression and lasso, the penalty parameter runs over a grid of values $k$. The relative MSFE is reported for the dependent variables industrial production (INDP), unemployment rate (UNR), inflation (CPI), three month treasury bill rate (3TB), and the average over all series.

Principal component regression is worse than the random methods in predicting all four series and substantially worse on average over all series. The same holds for partial least squares, with the exception of inflation, where it outperforms random subset, but not random projection regression.

With regard to the lasso and ridge regression benchmarks, the results show that on average, these methods are outperformed by both random subset and random projection re-

**Figure 4.4:** FRED-MD: forecast accuracy for different subspace dimensions



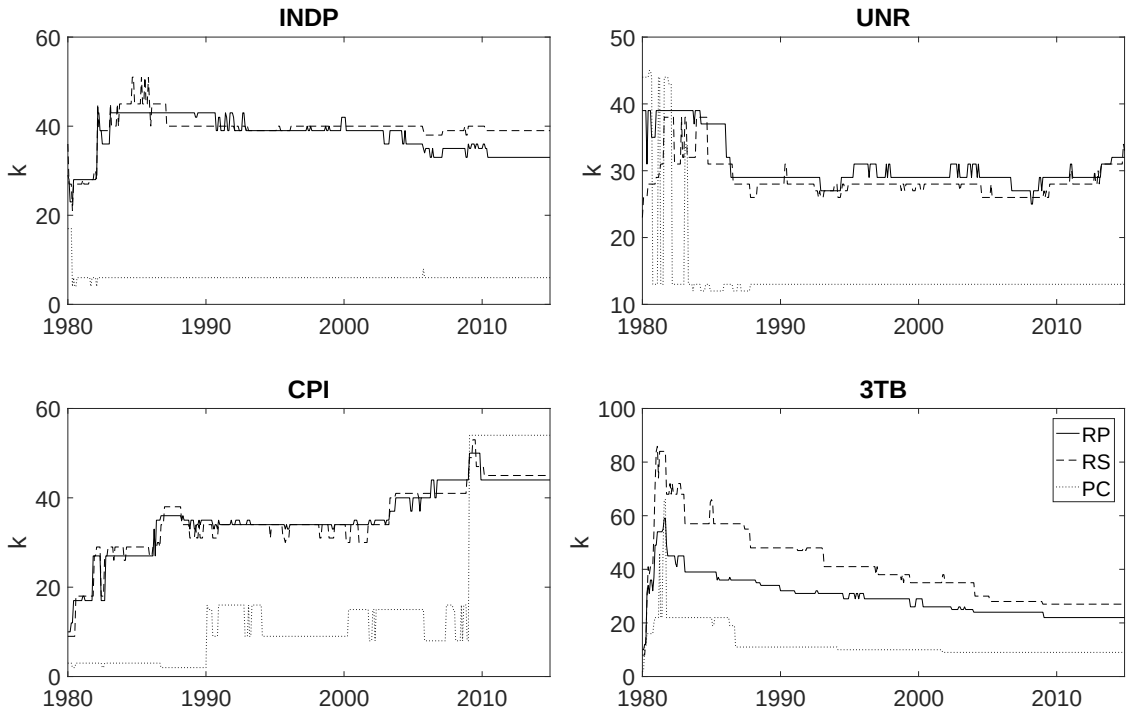

Note: this figure shows the relative MSFE for different values of the subspace dimension $k$. The different lines correspond to the evaluation criterium for the dimension reduction methods random projection (RP, solid), random subset (RS, dashed), and principal component regression (PC, dotted). The models at $k = 0$ corresponds to an autoregressive model of order four. The four panels correspond to four dependent variables, industrial production (INDP), unemployment rate (UNR), inflation (CPI), and three month treasury bill rate (3TB).

gression. Random projection regression has a slight edge on ridge regression, which is in line with our findings in Section 4.4. For the individual series reported here, the evidence is mixed. Random projection regression outperforms both ridge and lasso on these series, except for industrial production. Random subset regression is only outperformed by ridge or lasso when predicting the treasury bill rate.

Table 4.4 also shows the dependence of the MSFE on the value of $k$ if we were to pick the same $k$ throughout the forecasting period. Apart from the treasury bill rate, the random subspace methods outperform the AR(4) benchmark model for almost all subspace dimensions, even for very large values of $k$. Compared to principal component regression and partial least squares, we again see that the random methods select much larger values of $k$.

To visualize the dependence on $k$ for the different dimension reduction methods, Figure 4.4 shows the results for all subspace dimensions ranging from 0 to 100. The first thing to notice is the distinct development of the MSFE of forecasts generated by principal components

**Figure 4.5:** FRED-MD: recursive selection of subspace dimensions

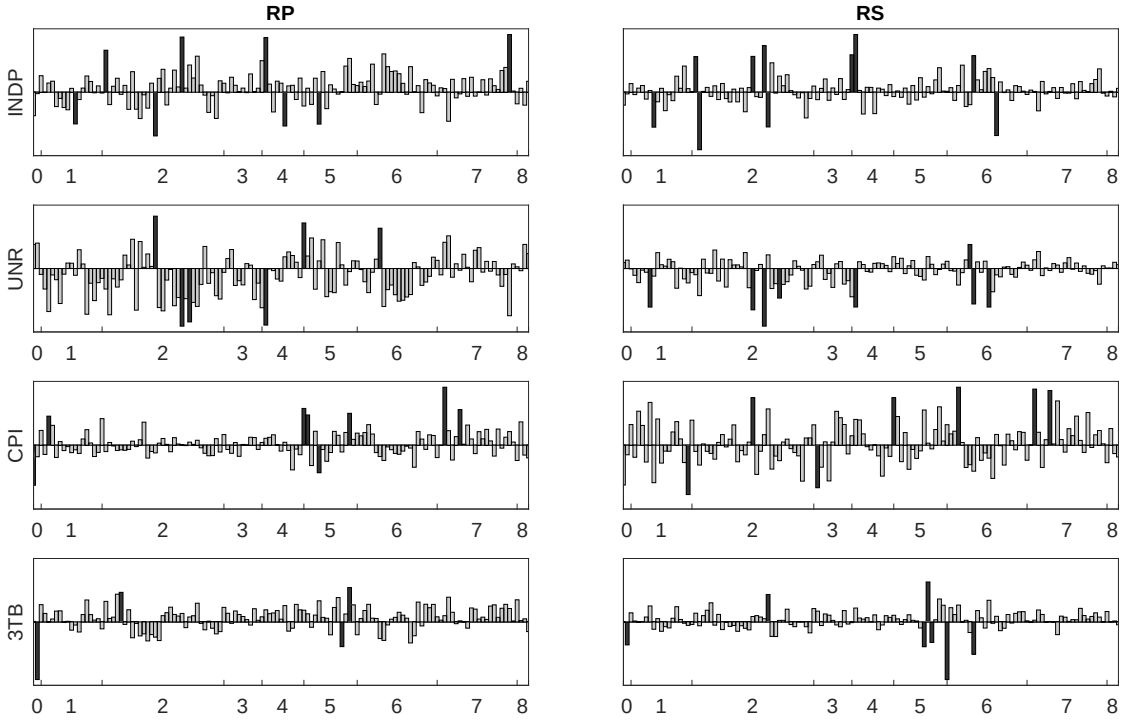



Note: this figure shows the selection of subset dimension $k$. The different lines correspond to the dimension reduction methods random projection (RP, solid), random subset (RS, dashed), and principal component regression (PC, dotted). At each point in time the subspace dimension is selected based on its past predictive performance up to that point in time. The four panels correspond to four dependent variables, industrial production (INDP), unemployment rate (UNR), inflation (CPI), and the three month treasury bill rate (3TB).

compared to the random subspace methods. The MSFE evolves smoothly over subspace dimensions for random projections and random subsets, where the MSFE of the principal components changes rather erratically.

Figure 4.4 shows that the random subspace methods reach their minimum for relatively large values of $k$. The selected value is substantially larger than the selected dimension when using principal component regression. The difference is especially clear for industrial production in the upper left panel, where principal components suggests to use six factors, while the random methods reach their minimum when using a subspace of dimension larger than 30. Apparently, the information in the additional random factors outweigh the increase in parameter uncertainty and contain more predictive content than higher order principal components. In general, the MSFE of the random subspace methods seems to be lower for most values of $k$.

**Figure 4.6:** FRED-MD: relative weight predictors in random subspace methods




Note: this figure shows the average coefficients of the predictors in $x_t$ in random projection regression (RP) in the left column and random subset regression (RS) in the right column, estimated by $\mathrm{E}_R\left[R\hat{\beta}_{x,R}\right]$ for the optimal subspace dimension in the last estimation sample. Series are grouped in different macroeconomic indicators as described in McCracken and Ng (2016) and the 'zero' group represents the lagged values of the dependent variable. The rows correspond to four dependent variables, industrial production (INDP), unemployment rate (UNR), inflation (CPI), and the three month treasury bill rate (3TB). Dark coloured bins indicate coefficients which differ two standard deviations from the average over all coefficients.

In practice, we do not know the optimal subspace dimension. Therefore, real-time forecasts are based on recursively selected values for $k$ based on past performance. Figure 4.5 shows the selection of the subspace dimension over time. In line with the ex-post optimal subspace dimension, the selected value of $k$ based on past predictive performance is smallest for principal component regression. The selected subspace dimension for random subset regression and random projection regression is very similar, but we do find quite some variation over time.

The left upper panel shows that for industrial production, the subspace dimension increases from approximately 30 to 40, where it is quite constant since the mid eighties. The dimension of random projection regression gradually declines back to 33 since the early 2000s. For the unemployment rate in the right upper panel, we observe that more factors seem to be selected since 2008 for both randomized methods, although this has not risen

above historically observed values. This is in contrast with the inflation series in the lower left panel. Since the early 2000s both random methods choose gradually larger subspaces, while principal components shows a single sharp increase in 2009. The right lower panel shows that for the treasury bill rate, as one might expect, the subspace dimension decreases over time, reaching its minimum after the onset of the global financial crisis. The historical low can be explained by the lack of predictive content in the data since the zero lower bound of the interest rate impedes most variation in the dependent variable.

Figure 4.6 provides insights in the relation between the predictors and the macroeconomic indicator of interest. We find that random projections and random subset regression estimate different values for the average coefficients. For instance, random projections assigns most weight to lagged values of the three month treasury bill rate to predict this variable, where random subsets mostly explains the one-step-ahead forecast by indicators for money and credit (5) and interest rate and exchange rates (6). The average coefficients also differ over the different series. Where industrial production and unemployment rate are related to variables from all indicator groups, inflation rate seems best explained by indicators for money and credit (5) and prices (7), especially for random projection regression.

### 4.5.4 Lagged predictors

Although the theoretical results in Section 4.3 assume $T > p$, we empirically find that the random subspace methods also outperform benchmark methods for $p > T$. Following Bai and Ng (2008) among others, we include lags of the predictors in the forecasting model. We extend $x_t$ with five lags of the variables in the database, such that we have six time periods for each macroeconomic indicator in the database in $x_t$. The first estimation sample contains 174 observations, while we have 781 regressors. We average over $N = 6{,}000$ forecasts to obtain one prediction in the random subspace methods.

The random subspace methods without including the extra lags of predictors show the best performance. Comparing the numbers in Table 4.5 to the relative MSFE for the optimal subspace dimension in Table 4.4, we find that random subset regression shows the overall best performance for industrial production and unemployment rate, and random projection regression for inflation and the treasury bill rate. Only principal component regression and

**Table 4.5:** FRED-MD: forecast accuracy with lagged predictors

| | RP | RS | PC | PL | RI | LA |
|---|---|---|---|---|---|---|
| INDP | 0.894 | 0.878 | 0.849 | 0.914 | 0.890 | 0.884 |
| UNR | 0.872 | 0.848 | 0.872 | 0.871 | 0.873 | 0.868 |
| CPI | 0.905 | 0.895 | 0.943 | 0.973 | 0.904 | 0.957 |
| 3TB | 0.958 | 0.978 | 1.158 | 1.047 | 0.976 | 0.971 |

Note: this table shows the relative MSFE generated by the optimal subspace dimension $k$ of different methods using six lags of the predictors in $x_t$, for the dependent variables, industrial production (INDP), unemployment rate (UNR), inflation (CPI), and the three month treasury bill rate (3TB).

partial least squares improve in some cases in forecast accuracy by including lagged predictors.

Table 4.5 shows no conclusive outcome for the relative forecast accuracy of the methods for the different macroeconomic indicators. Principal component regression is most accurate for industrial production, random subset regression for unemployment rate and inflation, and random projection for the treasury bill rate. Using random subspace methods in this high-dimensional setting increases the forecast performance for three out of the four macroeconomic indicators we consider.

### 4.5.5 Benchmark dataset

We perform the same analysis as discussed in 4.5.2 to the Stock and Watson (2005) data, which is used by many researchers to examine macroeconomic forecast accuracy of their methods (Stock and Watson, 2006; Bai and Ng, 2008; Fuentes et al., 2015). The 132 monthly time series run from January 1960 to December 2003. Because we consider six lags of $y_{t+1}$, the first estimation sample of ten years starts in June 1960. After the burn-in period, we generate forecasts from November 1973 to December 2003. Apart from the starting date, the design mimicks the empirical application in Bai and Ng (2008), where the first estimation sample starts in March 1960. Note that for the first 38 forecasts, the parameters are estimated in a setting where $p > T$.

Just as we found for the FRED-MD data, random subset regression performs best in terms of MSFE. Table 4.6 shows that random subset regression outperforms the other methods for industrial production and unemployment rate, and ranks second in terms of lowest MSFE for

**Table 4.6:** Stock and Watson (2005) data: forecast accuracy

| | RP | RS | PC | PL | RI | LA |
|---|---|---|---|---|---|---|
| INDP | 0.837 | 0.804 | 0.852 | 0.892 | 0.837 | 0.813 |
| UNR | 0.824 | 0.809 | 0.816 | 0.810 | 0.824 | 0.815 |
| CPI | 0.986 | 0.988 | 0.992 | 1.027 | 0.988 | 1.018 |
| 3TB | 0.903 | 0.900 | 0.936 | 0.893 | 0.906 | 0.935 |

Note: this table shows the relative MSFE generated by the optimal subspace dimension $k$ of different methods for the dependent variables, industrial production (INDP), unemployment rate (UNR), inflation (CPI), and the three month treasury bill rate (3TB).

inflation and treasury bill rate. Random projection regression is more accurate in predicting inflation, and partial least squares in predicting the three month treasury bill rate.

## 4.6 Conclusion

In this chapter we study two random subspace methods that offer a promising way of dimension reduction to construct accurate forecasts. The first method randomly selects many different subsets of the original variables to construct a forecast. The second method constructs predictors by randomly weighting the original predictors. Although counterintuitive at first, we provide a theoretical justification for these strategies by deriving bounds on their asymptotic mean squared forecast error. These bounds are highly informative on the scenarios where one can expect the two methods to work well and where one is to be preferred over the other.

The theoretical findings are confirmed in a Monte Carlo simulation, where in addition we show that the predictive accuracy increases for nearly all settings under consideration relative to several widely used benchmarks: principal component regression, partial least squares, lasso regularization and ridge regression. In the empirical application, random subset regression generates more accurate forecasts than the benchmarks for no less than 66% of the 130 macroeconomic indicators, and random projection regression outperforms the benchmarks in at least 57% of the series.

# 4.A Proofs

## 4.A.1 Independence between predictor and estimation error

We need the following independence result to derive properties on the forecast accuracy of the random subspace methods.

**Lemma 4.4** *For the regression model in* (4.1) *under Assumption A4.1-A4.7,* $z_T$ *is independent of* $\sqrt{T}(\hat{\beta} - \beta)$ *as* $T \to \infty$.

Proof: We have $T$ observations available for estimation of the parameter vector $\beta$. For some $\alpha > 0$, take $T_1 = (1 - T^{-\alpha})T$, such that $T_1/T = O(1)$, $(T - T_1)/T = o(1)$. We require $T - T_1 \to \infty$, such that $\alpha < 1$. The estimation error is given by

$$\sqrt{T}(\hat{\beta} - \beta) = \left(\frac{1}{T}\sum_{t=0}^{T-1} z_t z_t'\right)^{-1} \frac{1}{\sqrt{T}}\sum_{t=0}^{T-1} z_t \varepsilon_{t+1}. \tag{4.31}$$

We split $\frac{1}{\sqrt{T}}\sum_t z_t\varepsilon_{t+1}$ into a part that is independent of $z_T$ and one that is dependent of $z_T$, but negligible as $T \to \infty$.

$$\frac{1}{\sqrt{T}}\sum_{t=0}^{T-1} z_t\varepsilon_{t+1} = \sqrt{\frac{T_1}{T}}\frac{1}{\sqrt{T_1}}\sum_{t=0}^{T_1} z_t\varepsilon_{t+1} + \sqrt{\frac{T-T_1}{T}}\frac{1}{\sqrt{T-T_1}}\sum_{t=T_1+1}^{T-1} z_t\varepsilon_{t+1}. \tag{4.32}$$

By Assumption A4.4, $\text{var}(z_{ti}\varepsilon_{t+1}) = E[(z_{ti}\varepsilon_{t+1})^2] < \Delta < \infty$. By Chebyshev's inequality $P(|z_{it}\varepsilon_{t+1}| \geq T^{\frac{1}{4}}) \leq T^{-\frac{1}{2}}\Delta$. Using Bonferroni's inequality, we then have $P(\max_{t=T_1+1,\ldots,T-1}|z_{it}\varepsilon_{t+1}| \geq T^{\frac{1}{4}}) \leq T^{\frac{1}{2}-\alpha}\Delta$. For this to hold almost surely when $T \to \infty$, we require $\alpha > \frac{1}{2}$. Then,

$$\begin{aligned}\sqrt{\frac{T-T_1}{T}}\frac{1}{\sqrt{T-T_1}}\sum_{t=T_1+1}^{T-1} z_{it}\varepsilon_{t+1} &\leq \sqrt{\frac{T-T_1}{T}}\frac{1}{\sqrt{T-T_1}}\sum_{t=T_1+1}^{T-1}|z_{it}\varepsilon_{t+1}| \\ &\leq T^{-\frac{1}{2}+\frac{1}{4}+1-\alpha}.\end{aligned} \tag{4.33}$$

Choosing $\alpha > \frac{3}{4}$, we have that

$$\frac{1}{\sqrt{T}}\sum_{t=0}^{T-1} z_{it}\varepsilon_{t+1} = \sqrt{\frac{T_1}{T}}\frac{1}{\sqrt{T_1}}\sum_{t=0}^{T_1} z_{it}\varepsilon_{t+1} + o_p(1). \tag{4.34}$$

Since under Assumptions A4.1-A4.7 a central limit theorem yields $\frac{1}{\sqrt{T}}\sum_{T=1}^{T} z_{it}\varepsilon_{t+1} \sim N(0,\Sigma_z)$, the left-hand size is $O_p(1)$. This implies that the first term on the right-hand side is $O_p(1)$. Since $\{(z_t', \varepsilon_{t+1})\}$ is strong mixing by Assumption A4.1, and $T - T_1 \to \infty$ for $\alpha < 1$, we have that $z_T$ is independent of the first term of the right-hand side in the limit where $T \to \infty$. Then $z_T$ is also independent of the left-hand side when $T \to \infty$.

The same argument can be used to show that $z_T$ is asymptotically independent of $\frac{1}{T}\sum_{t=0}^{T-1} z_t z_t'$. This shows that as $T \to \infty$, $z_T$ is independent of $\sqrt{T}(\hat{\beta} - \beta)$. ■

## 4.A.2 Proof of Theorem 4.1

By Jensen's inequality, the asymptotic MSFE can be bounded as

$$\begin{aligned}\rho(k) &= \mathrm{E}_\varepsilon\left[\lim_{T\to\infty} T\mathrm{E}_{z_T}\left[\left(z_T'\beta - z_T'\mathrm{E}_R\left[S_R\hat{\beta}_R\right]\right)^2\right]\right] \\ &\leq \mathrm{E}_R\left[\mathrm{E}_\varepsilon\left[\lim_{T\to\infty} T\mathrm{E}_{z_T}\left[\left(z_T'\beta - z_T'S_R\hat{\beta}_R\right)^2\right]\right]\right].\end{aligned} \tag{4.35}$$

We define the expectation operator $E_{R,\varepsilon} = E_R[E_\varepsilon[.]]$ and rewrite the bound as

$$\begin{aligned}\rho(k) &\leq \mathrm{E}_{R,\varepsilon}\left[\lim_{T\to\infty} T\mathrm{E}_{z_T}\left[\mathrm{trace}\left\{z_T z_T'(\beta - S_R\hat{\beta}_R)(\beta - S_R\hat{\beta}_R)'\right\}\right]\right] \\ &= \mathrm{E}_{R,\varepsilon}\left[\lim_{T\to\infty} T\mathrm{trace}\left\{\mathrm{E}_{z_T}\left[z_T z_T'(\beta - A_R\hat{\beta})(\beta - A_R\hat{\beta})'\right]\right\}\right],\end{aligned} \tag{4.36}$$

where we use the linearity of the trace and define $A_R \equiv S_R(S_R'Z'ZS_R)^{-1}S_R'Z'Z$. We now invoke the asymptotic independence of $z_T$ and $\hat{\beta}$ established in Lemma 4.4 in Appendix 4.A.1 to evaluate the expectation with respect to $z_T$. Using that $\mathrm{E}[z_T z_T'] = \Sigma_z$, we then continue as

$$\begin{aligned}\rho(k) &\leq \mathrm{E}_{R,\varepsilon}\left[\lim_{T\to\infty} T(\beta - A_R\hat{\beta})'\Sigma_z(\beta - A_R\hat{\beta})\right] \\ &= \mathrm{E}_{R,\varepsilon}\left[\lim_{T\to\infty} (\beta - A_R\hat{\beta})'Z'Z(\beta - A_R\hat{\beta})R\right],\end{aligned} \tag{4.37}$$

where the second line follows from $\operatorname{plim}_{T\to\infty} \frac{1}{T} Z'Z = \Sigma_z$ in (4.12), and Slutsky's theorem. Since $A_R\hat{\beta} = S_R\hat{\beta}_R$, the bound can be rewritten to

$$\begin{aligned}\rho(k) &\leq \mathrm{E}_{R,\varepsilon}\left[\lim_{T\to\infty}(\beta - S_R\hat{\beta}_R)'Z'Z(\beta - S_R\hat{\beta}_R)\right] \\ &= \mathrm{E}_{R,\varepsilon}\left[\lim_{T\to\infty}(y - \varepsilon - ZS_R\hat{\beta}_R)'(y - \varepsilon - ZS_R\hat{\beta}_R)\right] = \\ &\mathrm{E}_{R,\varepsilon}\left[\lim_{T\to\infty}\left(\varepsilon'\varepsilon + (y - ZS_R\hat{\beta}_R)'(y - ZS_R\hat{\beta}_R) - 2\varepsilon'(y - ZS_R\hat{\beta}_R)\right)\right].\end{aligned} \tag{4.38}$$

To proceed, note that $\hat{\beta}_R = \arg\min_u (y - ZS_R u)'(y - ZS_R u)$. Therefore, it holds for an arbitrary $p \times 1$ vector $v$ that

$$\begin{aligned}\rho(k) &\leq \mathrm{E}_{R,\varepsilon}\left[\lim_{T\to\infty}\left(\varepsilon'\varepsilon + (y - ZS_R v)'(y - ZS_R v) - 2\varepsilon'(y - ZS_R\hat{\beta}_R)\right)\right] \\ &= \mathrm{E}_{R,\varepsilon}\left[\lim_{T\to\infty}\left((\beta - S_R v)'Z'Z(\beta - S_R v) + 2\varepsilon'(ZS_R\hat{\beta}_R - ZS_R v)\right)\right].\end{aligned} \tag{4.39}$$

Since we are free to choose $v$, we choose

$$v = \begin{pmatrix} \beta_w \\ \frac{1}{\sqrt{T}} R'u \end{pmatrix} + (S_R'Z'ZS_R)^{-1}S_R'Z'\varepsilon, \tag{4.40}$$

with $u$ a fixed $p_x \times 1$ vector. Using (4.12), $\frac{1}{\sigma^2}\varepsilon'ZS_R(S_R'Z'ZS_R)^{-1}S_R'Z'\varepsilon \xrightarrow{(d)} \chi^2(p_w + k)$. Substituting (4.40) into (4.39) and taking the expectation with respect to $\varepsilon$ conditional on $R$ gives

$$\rho(k) \leq \sigma^2(p_w + k) + \mathrm{E}_R\left[(\beta_{x,0} - RR'u)'\Sigma_x(\beta_{x,0} - RR'u)\right]. \tag{4.41}$$

The bound in (4.41) is valid for any choice of $u$. After taking the expectation with respect to $R$, we can therefore minimize the bound with respect to $u$. Together with the fact that $\mathrm{E}_R[RR'] = \frac{k}{p_x} I_{p_x}$, this yields

$$\rho(k) \leq \sigma^2(p_w + k) + \beta_{x,0}'\Sigma_x\beta_{x,0} - \beta_{x,0}'\Sigma_x\left(\frac{p_x}{k}\mathrm{E}_R\left[RR'\Sigma_x RR'\right]\frac{p_x}{k}\right)^{-1}\Sigma_x\beta_{x,0}. \tag{4.42}$$

■

### 4.A.3 Proof of Lemma 4.1

Note that $RR'$ is a $p_x \times p_x$ diagonal matrix with $k$ diagonal elements equal to 1, and the remaining elements equal to zero. This implies that

$$[RR'\Sigma_x RR']_{ij} = \begin{cases} [\Sigma_x]_{ij} & \text{if } [RR']_{ii}[RR']_{jj} = 1, \\ 0 & \text{if } [RR']_{ii}[RR']_{jj} = 0. \end{cases} \tag{4.43}$$

Because the non-zero entries are selected uniformly at random, $P\left([RR']_{ii} = 1\right) = \frac{k}{p_x}$ and $P\left([RR']_{ii}[RR']_{jj} = 1\right) = \frac{k}{p_x}\frac{k-1}{p_x-1}$ for $i \neq j$. This yields $\mathrm{E}_R[RR'] = \frac{k}{p_x} I_{p_x}$ and

$$\mathrm{E}\left[[RR'\Sigma_x RR']_{ii}\right] = \frac{k}{p_x}[\Sigma_x]_{ii}, \qquad \mathrm{E}\left[[RR'\Sigma_x RR']_{ij}\right] = \frac{k}{p_x}\frac{k-1}{p_x-1}[\Sigma_x]_{ij}. \tag{4.44}$$

We summarize this as

$$\begin{aligned} \mathrm{E}\left[RR'\Sigma_x RR'\right] &= \frac{k}{p_x}\frac{k-1}{p_x-1}\Sigma_x + \frac{k}{p_x}\left(1 - \frac{k-1}{p_x-1}\right) D_{\Sigma_x} \\ &= \frac{k}{p_x}\left(\frac{k-1}{p_x-1}\Sigma_x + \frac{p_x-k}{p_x-1} D_{\Sigma_x}\right), \end{aligned} \tag{4.45}$$

where $[D_{\Sigma_x}]_{ii} = [\Sigma_x]_{ii}$, and $[D_{\Sigma_x}]_{ij} = 0$ if $i \neq j$. ■

### 4.A.4 Proof of Lemma 4.2

Define $Q = R(R'R)^{-1/2}$ and $P = (R'R)^{1/2}$. Furthermore, define the matrices $W = (w_0, \ldots, w_{T-1})'$ and $X = (x_0, \ldots, x_{T-1})'$. We have

$$\begin{aligned} S_R\hat{\beta}_R &= \begin{pmatrix} I_{p_w} & O \\ O & R \end{pmatrix} \begin{pmatrix} W'W & W'XR \\ R'X'W & R'X'XR \end{pmatrix}^{-1} \begin{pmatrix} W' \\ R'X' \end{pmatrix} y \\ &= \begin{pmatrix} (W'W)^{-1}W' - (W'W)^{-1}W'XV_R X'M_W \\ V_R X'M_W \end{pmatrix} y, \end{aligned} \tag{4.46}$$

where $M_W = I - W(W'W)^{-1}W'$ and $V_R = R(R'X'M_W XR)^{-1}R'$. Using now that $R = QP$ with $P$ an $k \times k$ invertible matrix, we immediately see that $V_R = Q(Q'X'M_W XQ)^{-1}Q'$. Hence, $S_R\hat{\beta}_R = S_Q\hat{\beta}_Q$, which completes the proof. ■

### 4.A.5 Proof of Lemma 4.3

Consider a matrix $R$ with independent standard normal entries, and a matrix $Q = R(R'R)^{-1/2}$ with the following property.

**Lemma 4.5** *Let $R$ be a $p_x \times k$ matrix with independent standard normal entries. Consider the decomposition $R = QP$, where $Q(R) = R(R'R)^{-1/2}$ and $P(R) = (R'R)^{1/2}$. When we write $U \in \mathcal{O}(p)$ if $U$ is a $p \times p$ orthogonal matrix, we have*

1. $Q(R) \overset{(d)}{=} H_{p_x} Q(R)$ *for* $H_{p_x} \in \mathcal{O}(p_x)$.
2. $Q(R) \overset{(d)}{=} Q(R) H_k$ *for* $H_k \in \mathcal{O}(k)$.

Proof: (Part 1) We have

$$Q(H_{p_x} R) = H_{p_x} R(R' H_{p_x}' H_{p_x} R)^{-1/2} = H_{p_x} Q(R). \tag{4.47}$$

Also, $H_{p_x} R \overset{(d)}{=} R$. This can be seen from the fact that the matrix variate normal distribution only depends on $R$ through the trace of $R'R$. Then $Q(H_{p_x} R) \overset{(d)}{=} Q(R)$. Combining this with (4.47), we see that $H_{p_x} Q(R) \overset{(d)}{=} Q(R)$. (Part 2) Decompose $R'R = U \Lambda U'$, where $U \in \mathcal{O}(k)$. Note that $(H_k' U \Lambda U' H_k)^{1/2} = H_k' U \Lambda^{1/2} U' H_k$, and $(H_k' U \Lambda U' H_k)^{-1/2} = H_k' U \Lambda^{-1/2} U' H_k$. Now we have

$$Q(RH_k) = RH_k (H_k' R' R H_k)^{-1/2} = RH_k H_k' (R'R)^{-1/2} H_k = Q(R) H_k. \tag{4.48}$$

Also $RH_k \overset{(d)}{=} R$, by the same arguments as before. Then $Q(RH_k) \overset{(d)}{=} Q(R) \overset{(d)}{=} Q(R) H_k$. ■

We use Lemma 4.5 and the eigenvalue decomposition of $\Sigma_x = H \Lambda H'$, where $H \in \mathcal{O}(p_x)$, to rewrite

$$\begin{aligned} \mathrm{E}_Q[QQ' \Sigma_x QQ'] &= \mathrm{E}_Q[QQ' H \Lambda H' QQ'] \\ &= \mathrm{E}_Q[HH' QQ' H \Lambda H' QQ' HH'] = H \mathrm{E}_Q[QQ' \Lambda QQ'] H'. \end{aligned} \tag{4.49}$$

The elements of the matrix $M = QQ'\Lambda QQ'$, $m_{ii'}$, are a function of the eigenvalues of $\Sigma_x$, $\lambda_i$, and the elements of $Q$, $q_{ij}$, for $i, i' = 1, \ldots, p_x$ and $j = 1, \ldots, k$:

$$\begin{aligned} m_{ii} &= \lambda_i \left( \sum_{j=1}^{k} q_{ij}^4 + \sum_{j \neq j'} q_{ij}^2 q_{ij'}^2 \right) + \sum_{l \neq i} \lambda_l \left( \sum_{j=1}^{k} q_{ij}^2 q_{lj}^2 + \sum_{j \neq j'} q_{ij} q_{lj} q_{ij'} q_{lj'} \right), \\ m_{ii'} &= \lambda_i \left( \sum_{j=1}^{k} q_{ij}^3 q_{i'j} + \sum_{j \neq j'} q_{ij}^2 q_{ij'} q_{i'j'} \right) + \lambda_{i'} \left( \sum_{j=1}^{k} q_{i'j}^3 q_{ij} + \sum_{j \neq j'} q_{i'j}^2 q_{i'j'} q_{ij'} \right) \\ &\quad + \sum_{l \neq \{i,i'\}} \lambda_l \left( \sum_{j=1}^{k} q_{ij} q_{i'j} q_{lj}^2 + \sum_{j \neq j'} q_{ij} q_{i'j'} q_{lj} q_{lj'} \right). \end{aligned} \tag{4.50}$$

From (4.50) it follows that we need the (mixed) moments of $q_{ij}$ up to fourth order to evaluate $\mathrm{E}_Q[QQ'\Lambda QQ']$. These are provided in the following lemma.

**Lemma 4.6** *Suppose we have a $p_x \times k$ matrix $Q$ for which $Q'Q = I_k$ and the $i,j$-th entry of $Q$ is denoted by $q_{ij}$, where $i = 1, \ldots, p_x$, $j = 1, \ldots, k$, and $i \neq i', j \neq j'$. For any fixed $p_x \times p_x$ orthogonal matrix $H_{p_x}$ and $k \times k$ orthogonal matrix $H_k$ the matrix $Q$ satisfies the invariance property $H_{p_x} Q H_k \overset{(d)}{=} Q$. Then the non-zero (mixed) moments up to fourth-order are*

$$\begin{aligned} E\left[q_{ij}^2\right] &= \frac{1}{p_x}, \\ E\left[q_{ij}^4\right] &= \frac{3}{p_x(p_x+2)}, \\ E\left[q_{ij}^2 q_{ij'}^2\right] = E\left[q_{ij}^2 q_{i'j}^2\right] &= \frac{1}{p_x(p_x+2)}, \\ E\left[q_{ij}^2 q_{i'j'}^2\right] &= \frac{p_x+1}{p_x(p_x-1)(p_x+2)}, \\ E\left[q_{ij} q_{ij'} q_{i'j} q_{i'j'}\right] &= -\frac{1}{p_x(p_x-1)(p_x+2)}. \end{aligned} \tag{4.51}$$

Note that none of the non-zero (mixed) moments appear in the expression for $m_{ii'}$, such that $\mathrm{E}[m_{ii'}] = 0$.

Proof: We consider the orthogonal matrix $H$ with fixed indices $r$ and $r' \neq r$, and define the elements of $H$ as

$$h_{ij} = \begin{cases} 1 & \text{if } i = j, i \neq r, i \neq r', \\ \sin(\theta) & \text{if } i = j = r, \text{ or } i = j = r', \\ \cos(\theta) & \text{if } i = r, j = r', \\ -\cos\theta & \text{if } i = r', j = r, \\ 0 & \text{otherwise,} \end{cases} \tag{4.52}$$

where for $H_{p_x}$, $i, j = 1, \ldots, p_x$ and for $H_k$, $i, j = 1, \ldots, k$. $H_{p_x}$ sets $\theta = \theta_1$ and $H_k$ sets $\theta = \theta_2$. Throughout this proof, we use the notation that for any index $i' \neq i$. From the invariance property $H_{p_x} Q H_k \overset{(d)}{=} Q$ follows that the elements of $Q$ satisfy

$$\begin{aligned} q_{ij} \overset{(d)}{=} \sin(\theta_1)\sin(\theta_2) q_{ij} + \cos(\theta_1)\sin(\theta_2) q_{i'j} \\ - \sin(\theta_1)\cos(\theta_2) q_{ij'} - \cos(\theta_1)\cos(\theta_2) q_{i'j'}. \end{aligned} \tag{4.53}$$

**First moment** Choosing $\theta_1 = \theta_2 = \pi$, we get $q_{ij} \overset{(d)}{=} q_{i'j'}$. Similary, choosing $\theta_1 = 0$ and $\theta_2 = \frac{\pi}{2}$, we get $q_{ij} \overset{(d)}{=} q_{i'j}$. Proceeding in this manner, we conclude that the elements $q_{ij}$ are identically distributed. Furthermore, choosing $\theta_1 = \theta_2 = 0$, we see that $q_{ij} \overset{(d)}{=} -q_{i'j'}$. Since $\mathrm{E}[q_{ij}] = \mathrm{E}[q_{i'j'}] = -\mathrm{E}[q_{ij}]$, we have $\mathrm{E}[q_{ij}] = 0$.

**Second moment** We have $Q'Q = I_k$, which implies that $\sum_{i=1}^{p_x} q_{ij}^2 = 1$ for every $j$. Taking the expectations on both sides and noting that the elements of $Q$ are identically distributed, we have $\mathrm{E}[q_{ij}^2] = \frac{1}{p_x}$. We now proceed to the mixed moments. Take $\theta_2 = \pi/2$ and $\theta_1 = \theta$ in (4.53), such that $q_{ij} \overset{(d)}{=} \sin(\theta) q_{ij} + \cos(\theta) q_{i'j}$. Then $q_{ij}^2 \overset{(d)}{=} \sin^2(\theta) q_{ij}^2 + \cos^2(\theta) q_{i'j}^2 + 2\sin(\theta)\cos(\theta) q_{ij} q_{i'j}$. Since $\mathrm{E}[q_{ij}^2] = \mathrm{E}[q_{i'j}^2]$, $\mathrm{E}[q_{ij} q_{i'j}] = 0$. Similary, taking $\theta_1 = \pi/2$ and $\theta_2 = \theta$ yields $\mathrm{E}[q_{ij} q_{ij'}] = 0$. Considering then the case for general $\theta_1$ and $\theta_2$ and using the previously derived results, we find $\mathrm{E}[q_{ij} q_{i'j'}] = 0$. Summarizing,

$$\mathrm{E}[q_{ij}^2] = \frac{1}{p_x}, \quad \mathrm{E}[q_{ij} q_{i'j}] = 0, \quad \mathrm{E}[q_{ij} q_{ij'}] = 0, \quad \mathrm{E}[q_{ij} q_{i'j'}] = 0. \tag{4.54}$$

**Fourth moment** Setting $\theta_2 = \pi/2$ and $\theta_1 = \theta$ in (4.53) yields

$$\begin{aligned} q_{ij}^4 \overset{(d)}{=} \sin^4(\theta) q_{ij}^4 + \cos^4(\theta) q_{i'j}^4 + 6\sin^2(\theta)\cos^2(\theta) q_{ij}^2 q_{i'j}^2 \\ + 4\sin^3(\theta)\cos(\theta) q_{ij}^3 q_{i'j} + 4\sin(\theta)\cos^3(\theta) q_{ij} q_{i'j}^3. \end{aligned} \tag{4.55}$$

Since all the elements of $Q$ are identically distributed, $\mathrm{E}[q_{ij}^4] = \mathrm{E}[q_{i'j}^4]$, and we have

$$\begin{aligned} \mathrm{E}[q_{ij}^4] &= [\sin^4(\theta) + \cos^4(\theta)]\mathrm{E}[q_{ij}^4] + 6\sin^2(\theta)\cos^2(\theta)\mathrm{E}[q_{ij}^2 q_{i'j}^2] \\ &\quad + 4\sin^3\theta\cos(\theta)\mathrm{E}[q_{ij}^3 q_{i'j}] + 4\sin(\theta)\cos^3\theta\mathrm{E}[q_{ij} q_{i'j}^3] \\ &= \mathrm{E}[q_{ij}^4] + 2\sin^2(\theta)\cos^2(\theta)\left(3\mathrm{E}[q_{ij}^2 q_{i'j}^2] - \mathrm{E}[q_{ij}^4]\right) \\ &\quad + 4\sin^3(\theta)\cos(\theta)\mathrm{E}[q_{ij}^3 q_{i'j}] + 4\sin(\theta)\cos^3(\theta)\mathrm{E}[q_{ij} q_{i'j}^3] = \\ &\mathrm{E}[q_{ij}^4] + 2\sin^2(\theta)\cos^2(\theta)\left(3\mathrm{E}[q_{ij}^2 q_{i'j}^2] - \mathrm{E}[q_{ij}^4]\right) + 4\sin(\theta)\cos(\theta)\mathrm{E}[q_{ij}^3 q_{i'j}], \end{aligned} \tag{4.56}$$

where we use that $\mathrm{E}[q_{ij}^3 q_{i'j}] = \mathrm{E}[q_{ij} q_{i'j}^3]$. For the equality in (4.56) to hold, we require

$$\mathrm{E}[q_{ij}^4] = 3\mathrm{E}[q_{ij}^2 q_{i'j}^2], \qquad \mathrm{E}[q_{ij}^3 q_{i'j}] = 0. \tag{4.57}$$

We use that $Q'Q = I_k$. For any $j$,

$$1 = \sum_{i=1}^{p_x} q_{ij}^2 = \left(\sum_{i=1}^{p_x} q_{ij}^2\right)^2 = \sum_{i=1}^{p_x} q_{ij}^4 + \sum_{i \neq i'} q_{ij}^2 q_{i'j}^2. \tag{4.58}$$

Taking the expectation and using (4.57), we have that $1 = p_x \mathrm{E}[q_{ij}^4] + \frac{p_x(p_x-1)}{3}\mathrm{E}[q_{ij}^4]$, which yields $\mathrm{E}[q_{ij}^4] = \frac{3}{p_x(p_x+2)}$, and $\mathrm{E}[q_{ij}^2 q_{i'j}^2] = \frac{1}{p_x(p_x+2)}$. For $\theta_1 = \pi/2$ and $\theta_2 = \theta$, analogous calculations yield

$$\mathrm{E}[q_{ij}^2 q_{ij'}^2] = \frac{1}{p_x(p_x+2)}, \quad \mathrm{E}[q_{ij}^3 q_{ij'}] = 0. \tag{4.59}$$

To obtain the remaining fourth order moments, we consider general $\theta_1$ and $\theta_2$ in (4.53). Using previously derived expressions, we arrive after tedious calculations at

$$\begin{aligned} \mathrm{E}[q_{ij}^4] = \mathrm{E}[q_{ij}^4] - a(\theta_1,\theta_2)\mathrm{E}[q_{ij}^3 q_{i'j'}] + b(\theta_1,\theta_2)\mathrm{E}[3q_{ij}^2 q_{i'j'}^2 + 6q_{ij}q_{ij'}q_{i'j}q_{i'j'} - q_{ij}^4] \\ + c(\theta_1,\theta_2)\left\{\mathrm{E}[q_{ij}^2 q_{ij'}q_{i'j}] + 2\mathrm{E}[q_{ij}^2 q_{i'j'}q_{i'j}]d(\theta_1) - 2\mathrm{E}[q_{ij}^2 q_{ij'}q_{i'j'}]d(\theta_2)\right\}, \end{aligned} \tag{4.60}$$

where

$$\begin{aligned} a(\theta_1,\theta_2) &= 4\cos(\theta_1)\cos(\theta_2)\sin(\theta_1)\sin(\theta_2)(2\cos(\theta_1)^2\cos(\theta_2)^2-1),\\ b(\theta_1,\theta_2) &= 4\cos(\theta_1)^2\cos(\theta_2)^2\sin(\theta_1)^2\sin(\theta_2)^2,\\ c(\theta_1,\theta_2) &= -12\cos(\theta_1)\cos(\theta_2)\sin(\theta_1)\sin(\theta_2),\\ d(\theta) &= \sin(\theta)\cos(\theta). \end{aligned} \tag{4.61}$$

Again, since the expectations should be independent of $\theta_1$ and $\theta_2$, this implies that $\mathrm{E}[q_{ij}^3 q_{i'j'}] = \mathrm{E}[q_{ij}^2 q_{ij'} q_{i'j}] = \mathrm{E}[q_{ij}^2 q_{i'j} q_{i'j'}] = \mathrm{E}[q_{ij}^2 q_{ij'} q_{i'j'}] = 0$, and that

$$\mathrm{E}[3q_{ij}^2 q_{i'j'}^2 + 6q_{ij}q_{ij'}q_{i'j}q_{i'j'} - q_{ij}^4] = 0. \tag{4.62}$$

Since the off-diagonal elements of $Q'Q$ are equal to zero, we have for any $j' \neq j$,

$$0 = \sum_{i=1}^{p_x} q_{ij}q_{ij'} = \left(\sum_{i=1}^{p_x} q_{ij}q_{ij'}\right)^2 = \sum_{i=1}^{p_x} q_{ij}^2 q_{ij'}^2 + \sum_{i\neq i'} q_{ij}q_{ij'}q_{i'j}q_{i'j'}. \tag{4.63}$$

Taking the expectation and using (4.59), we get $\frac{1}{p_x+2} = -\sum_{i\neq i'} \mathrm{E}[q_{ij}q_{ij'}q_{i'j}q_{i'j'}]$. Since the expectation should not depend on our choice of $i, j, i', j'$ as long as $i \neq i'$ and $j \neq j'$, we have that $\mathrm{E}[q_{ij}q_{ij'}q_{i'j}q_{i'j'}] = -\frac{1}{p_x(p_x-1)(p_x+2)}$. Then from (4.62) we obtain $\mathrm{E}[q_{ij}^2 q_{i'j'}^2] = \frac{p_x+1}{p_x(p_x-1)(p_x+2)}$. There is one final identity that we need. We found that $\mathrm{E}\left[q_{ij}^2\right] = \frac{1}{p_x}$ from which follows that $\mathrm{E}_Q[QQ'] = \frac{k}{p_x} I_{p_x}$. Then also $\mathrm{E}_Q[QQ'QQ'] = \frac{k}{p_x} I_{p_x}$. For the off-diagonal elements

$$\begin{aligned} [QQ'QQ']_{mm'} = &\sum_{i=1}^{k} q_{mi}^3 q_{m'i} + \sum_{i\neq i'} q_{mi}^2 q_{mi'} q_{m'i'} + \sum_{i=1}^{k} q_{m'i}^3 q_{mi}\\ &+ \sum_{i\neq i'} q_{m'i}^2 q_{m'i'} q_{mi'} + \sum_{l\neq\{m,m'\}}^{p_x} \left(\sum_{i=1}^{k} q_{mi}q_{m'i}q_{li}^2 + \sum_{i\neq i'} q_{mi}q_{li}q_{m'i'}q_{li'}\right). \end{aligned} \tag{4.64}$$

We know that $E_Q[QQ'QQ']_{mm'} = 0$, and the only term on the right-hand side for which we have no expression is the final one. This implies that $\mathrm{E}[q_{mi}q_{m'i'}q_{li}q_{li'}] = 0$, which completes the calculation of the moments of $q_{ij}$ up to fourth order. ■

Since $Q'Q = I_k$ for $Q = R(R'R)^{-1/2}$, and Lemma 4.5 shows that this choice for $Q$ satisfies the invariance property, we can apply Lemma 4.6 to $Q$. Lemma 4.6 states that $\mathrm{E}\left[q_{ij}^2\right] = \frac{1}{p_x}$ from which follows that $\mathrm{E}_Q[QQ'] = \frac{k}{p_x} I_{p_x}$.

Substituting the moments in Lemma 4.6 in the expectation of (4.50), we have

$$\begin{aligned} m_{ii} &= \frac{k}{p_x}\left(\frac{2+k}{p_x+2}\lambda_i + \frac{p_x-k}{(p_x+2)(p_x-1)}\sum_{l\neq i}\lambda_l\right) \\ &= \frac{k}{p_x}\left(\frac{p_x(k+1)-2}{(p_x+2)(p_x-1)}\lambda_i + \frac{p_x-k}{(p_x+2)(p_x-1)}\sum_{l=1}^{p_x}\lambda_l\right). \end{aligned} \tag{4.65}$$

Substituting this expression in (4.49), we arrive at

$$E_Q[QQ'\Sigma_x QQ'] = \frac{k}{p_x}\left(\frac{p_x(k+1)-2}{(p_x+2)(p_x-1)}\Sigma_x + \frac{(p_x-k)p_x}{(p_x+2)(p_x-1)}\frac{\mathrm{tr}(\Sigma_x)}{p_x} I_{p_x}\right). \tag{4.66}$$

■

## 4.A.6 Uniform improvement MSFE bound RP

Define $R$ to be a $p_x \times k$ matrix with independent normal entries. We set the variance equal to $1/p_x$ to ensure that $\mathrm{E}[RR'] = \frac{k}{p_x} I_p$. Take $Q = R(R'R)^{-1/2}$ a random orthogonal matrix. To show that the use of $Q$ in Theorem 4.1 yields a uniform improvement over using $R$, we need to show that $\Delta = \mathrm{E}[RR'\Sigma_x RR'] - \mathrm{E}[QQ'\Sigma_x QQ'] \succ 0$. From Kabán (2014), Lemma 2, we have that

$$\mathrm{E}[RR'\Sigma_x RR'] = \frac{k}{p_x}\left[\frac{k+1}{p_x}\Sigma_x + \frac{\mathrm{tr}(\Sigma_x)}{p_x} I_{p_x}\right]. \tag{4.67}$$

Then

$$\begin{aligned} \Delta &= \frac{k}{p_x}\left[\left(\frac{k+1}{p_x} - \frac{p_x(k+1)-2}{(p_x+2)(p_x-1)}\right)\Sigma_x + \left(1 - \frac{(p_x-k)p_x}{(p_x+2)(p_x-1)}\right)\frac{\mathrm{tr}(\Sigma_x)}{p_x} I_{p_x}\right] \\ &= \frac{k}{p_x}\left[\frac{p_x(k+1)-2+2(p_x-k)}{p_x(p_x+2)(p_x-1)}\Sigma_x + \frac{p_x(k+1)-2}{(p_x+2)(p_x-1)}\frac{\mathrm{tr}(\Sigma)}{p_x} I_{p_x}\right]. \end{aligned}$$

For the first term, $p_x(k+1) \geq 2$, with equality only when $p_x = k = 1$. Also, $p_x - k \geq 0$. For the second term, again $p_x(k+1) \geq 2$. We see that when $p_x > 1$, $\Delta = a\Sigma_x + bI_{p_x}$ with $a, b > 0$. Since $\Sigma_x$ is positive definite, this implies $\Delta$ is positive definite. ■

### 4.A.7 Eigenvalue bounds

**Lemma 4.7** *Let $R$ be a $p_x \times k$ random selection or random projection matrix, $\Sigma$ a $p_x \times p_x$ positive definite matrix and $V_R = \Sigma^{1/2}R(R'\Sigma R)^{-1}R'\Sigma^{1/2}$. Then*

$$\frac{\lambda_{\min}(\Sigma)}{\lambda_{\max}(\Sigma)}\frac{k}{p_x}\frac{1}{\eta} \leq \lambda_{\min}(\boldsymbol{E}_R[V_R]) \leq \lambda_{\max}(\boldsymbol{E}_R[V_R]) \leq \frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)}\frac{k}{p_x}\eta, \tag{4.68}$$

*where $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote respectively the minimum and maximum eigenvalue of $A$, $\eta = 1$ for $R$ a random projection matrix, and $\eta = 2$ for $R$ a random selection matrix.*

We provide separate proofs for random projections and random subsets.

**Random projections** Since both $\Sigma$ and $\mathrm{E}_R[R(R'\Sigma R)^{-1}R']$ are positive definite,

$$\begin{aligned}\lambda_{\min}(\Sigma)\lambda_{\min}(\mathrm{E}_R[R(R'\Sigma R)^{-1}R']) \leq& \lambda_{\min}(\mathrm{E}_R[V_R]) \\ \leq& \lambda_{\max}(\Sigma)\lambda_{\max}(\mathrm{E}_R[R(R'\Sigma R)^{-1}R']).\end{aligned} \tag{4.69}$$

As discussed in Section 4.3.1, we can replace $R$ by $Q = R(R'R)^{-1/2}$ and $\mathrm{E}_R[V_R] = \mathrm{E}_Q[V_Q]$. Furthermore, we use the singular value decomposition of $\Sigma = U\Lambda U'$, with $U \in \mathcal{O}(p_x)$, and apply Lemma 4.5 in Appendix 4.A.5 which says that $UQ \overset{(d)}{=} Q$. Then

$$\begin{aligned}\lambda_{\max}(\mathrm{E}_Q\left[Q(Q'\Sigma Q)^{-1}Q'\right]) &= \lambda_{\max}(U\mathrm{E}_Q\left[Q(Q'\Lambda Q)^{-1}Q'\right]U') \\ &= \lambda_{\max}(\mathrm{E}_Q\left[Q(Q'\Lambda Q)^{-1}Q'\right]).\end{aligned} \tag{4.70}$$

Now we apply the following lemma to $\mathrm{E}_Q\left[Q(Q'\Lambda Q)^{-1}Q'\right]$:

**Lemma 4.8** *Suppose we have a $p \times p$ matrix $A$. If $\Omega A\Omega = A$ for any $p \times p$ diagonal matrix $\Omega$ with elements randomly drawn from $\{-1, 1\}$, the matrix $A$ is diagonal.*

Proof: Since $\Omega A\Omega = A$, the elements of $A$ satisfy $a_{ij} = \omega_{ii}\omega_{jj}a_{ij}$. Since this holds for any $\Omega$, there always is an $\Omega$ such that $\omega_{ii} = -\omega_{jj}$, in which case $a_{ij} = 0$. ■

Pick $\Omega$ as in Lemma 4.8, then,

$$\begin{aligned}\Omega\mathrm{E}_Q\left[Q(Q'\Lambda Q)^{-1}Q'\right]\Omega &= \mathrm{E}_Q\left[\Omega Q(Q'\Omega\Omega\Lambda\Omega\Omega Q)^{-1}Q'\Omega\right] \\ &= \mathrm{E}_Q\left[\Omega Q(Q'\Omega\Lambda\Omega Q)^{-1}Q'\Omega\right] \\ &= \mathrm{E}_Q\left[Q(Q'\Lambda Q)^{-1}Q'\right],\end{aligned} \tag{4.71}$$

where we use that $\Omega$ is an orthogonal matrix, and hence $\Omega Q \overset{(d)}{=} Q$. This proves the diagonality of $\mathrm{E}_Q[Q(Q'\Lambda Q)^{-1}Q']$. We upper bound the eigenvalues of this matrix as

$$\begin{aligned}
\mathrm{E}_Q[q_i'(Q'\Lambda Q)^{-1}q_i] &= \mathrm{E}_Q[q_i'(Q'Q)^{-1/2}((Q'Q)^{-1/2}Q'\Lambda Q(Q'Q)^{-1/2})^{-1}(Q'Q)^{-1/2}q_i] \\
&\leq \mathrm{E}_Q[\lambda_{\max}([(Q'Q)^{-1/2}Q'\Lambda Q(Q'Q)^{-1/2}]^{-1})q_i'(Q'Q)^{-1}q_i] \\
&= \mathrm{E}_Q[(\lambda_{\min}[(Q'Q)^{-1/2}Q'\Lambda Q(Q'Q)^{-1/2}])^{-1}q_i'(Q'Q)^{-1}q_i] \\
&\leq \frac{1}{\lambda_{\min}(\Lambda)}\mathrm{E}_Q[q_i'(Q'Q)^{-1}q_i] = \frac{1}{\lambda_{\min}(\Sigma)}\frac{k}{p_x},
\end{aligned}$$

where the introduction of $(Q'Q)^{-1/2} = I_{p_x}$ emphasizes that we can use the Poincaré separation lemma to obtain the fourth line. Using (4.69), this gives the bound

$$\lambda_{\max}(E_Q[\Sigma^{1/2}Q(Q'\Sigma Q)^{-1}Q'\Sigma^{1/2}]) \leq \frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)}\frac{k}{p_x}. \tag{4.72}$$

The proof for the lower bound on the minimum eigenvalue follows analogously. ■

**Random subsets** We first establish a lower bound on $\lambda_{\min}(\mathrm{E}_R\left[R(R'\Sigma R)^{-1}R'\right])$. Define a $p_x \times p_x$ random permutation matrix $P_1 = [R_1, R_2, \ldots, R_m]$, with $m = \frac{p_x}{k}$. Take the $p_x \times (p_x + r)$ matrix $P = [P_1, P_2]$, where $P_2$ is a $p_x \times r$ random selection matrix such that $\tilde{m} = (p_x + r)/k$ is an integer and $r < p_x$. Furthermore, define a $p_x \times p_x$ random matrix $S = D_{\tilde{m}} \otimes (\iota_k \iota_k')$, where $D_{\tilde{m}}$ is a random $\tilde{m} \times \tilde{m}$ matrix where each diagonal element is equal to 1 with probability $1/\tilde{m}$ and a draw of $D$ has only one nonzero element on its diagonal. The $\otimes$ denotes the Kronecker product and $\iota_k$ is a $k \times 1$ vector of ones. Note that $\mathrm{E}[S] = \frac{1}{\tilde{m}}B$, where $B = I_{\tilde{m}} \otimes (\iota_k \iota_k')$ is a $p_x \times p_x$ matrix. Then,

$$R(R'\Sigma R)^{-1}R' \overset{(d)}{=} P[S \circ (B \circ P'\Sigma P)^{-1}]P', \tag{4.73}$$

where $\circ$ denotes the Hadamard product, and hence

$$\begin{aligned}
\mathrm{E}_R[R(R'\Sigma R)^{-1}R'] =& \mathrm{E}_{P,S}[P[S \circ (B \circ P'\Sigma P)^{-1}]P'] \\
=& \mathrm{E}_P[P\mathrm{E}_S[S \circ (B \circ P'\Sigma P)^{-1}|P]P'].
\end{aligned} \tag{4.74}$$

For the minimum eigenvalue of $\mathrm{E}_R\left[R(R'\Sigma R)^{-1}R'\right]$ now follows

$$\begin{aligned}\lambda_{\min}(\mathrm{E}_R[R(R'\Sigma R)^{-1}R']) &\geq \mathrm{E}_P[\lambda_{\min}(P\mathrm{E}_S[S\circ(B\circ P'\Sigma P)^{-1}|P]P')]\\ &\geq \mathrm{E}_P[\lambda_{\min}(\mathrm{E}_S[S\circ(B\circ P'\Sigma P)^{-1}|P])]\\ &\geq \frac{1}{2}\frac{k}{p_x}\mathrm{E}[\lambda_{\min}((B\circ P'\Sigma P)^{-1})]\\ &\geq \frac{1}{2}\frac{k}{p_x}\lambda_{\min}(\Sigma^{-1}) = \frac{1}{2}\frac{k}{p_x}\frac{1}{\lambda_{\max}(\Sigma)},\end{aligned} \tag{4.75}$$

where in the first line we use that the minimal eigenvalue is a concave function. For the second inequality we use that for any matrix $PAP'$, we have $\lambda_{\min}(PAP') = \min_v \frac{v'PAP'v}{v'v}$, with $\frac{v'PAP'v}{v'v} = \frac{\tilde{v}A'\tilde{v}}{\tilde{v}'\tilde{v}-v'P_2P_2'v} \geq \frac{\tilde{v}A'\tilde{v}}{\tilde{v}'\tilde{v}}$. Then $\lambda_{\min}(PAP') \geq \lambda_{\min}(A)$. Next, we use that $E[S] = \frac{k}{p_x+r}B \geq \frac{1}{2}\frac{k}{p_x}B$. Finally, on the fourth line, we use that $B\circ P'\Sigma P$ is block diagonal, so that its eigenvalues are bounded by the eigenvalues of the blocks. The blocks itself are inverses of $k\times k$ principal submatrices of $\Sigma$, with their eigenvalues bounded by the eigenvalues of $\Sigma^{-1}$.

We derive an upper bound on $\lambda_{\max}(\mathrm{E}_R\left[R(R'\Sigma R)^{-1}R'\right])$ in a similar way. Define a $p_x\times p_x$ random permutation matrix $P_1 = [P, P_2]$. We take $P$ to be a $p_x\times(p_x-r)$ random selection matrix such that $(p_x-r)/k$ is an integer. Note that $r < \frac{1}{2}p_x$. We now repeat the argument above. For any matrix $PAP'$ we have $\lambda_{\max}(PAP') = \max_v \frac{v'PAP'v}{v'v}$, with $\frac{v'PAP'v}{v'v} = \frac{\tilde{v}A'\tilde{v}}{\tilde{v}'\tilde{v}+v'P_2P_2'v} \leq \frac{\tilde{v}A'\tilde{v}}{\tilde{v}'\tilde{v}}$. Then $\lambda_{\max}(PAP') \leq \lambda_{\max}(A)$. Moreover, $\mathrm{E}[S] = \frac{k}{p_x-r}B \leq \frac{2k}{p_x}B$. This results in

$$\lambda_{\max}(\mathrm{E}[R(R'\Sigma R)^{-1}R'] \leq 2\frac{1}{\lambda_{\max}(\Sigma)}\frac{k}{p_x}. \tag{4.76}$$

Combining the bounds on the eigenvalues of $\mathrm{E}_R\left[R(R'\Sigma R)^{-1}R'\right]$ with (4.69) completes the proof. ■

## 4.A.8 Lower bound on MSFE

We rewrite $\rho(k)$ in(4.14) using a bias-variance decomposition and Lemma 4.4 in Appendix 4.A.1,

$$\rho(k) = \mathrm{E}_\varepsilon[Y]'\Sigma_z\mathrm{E}_\varepsilon[Y] + \mathrm{E}_\varepsilon[(Y-\mathrm{E}_\varepsilon[Y])'\Sigma_z(Y-\mathrm{E}_\varepsilon[Y])], \tag{4.77}$$

where we introduce $\sqrt{T}(\beta - \mathrm{E}_R[S_R\hat{\beta}_R]) \xrightarrow{(d)} Y$ to shorten notation. We separately bound the bias and variance term in (4.77).

Using (4.46) from Appendix 4.A.4, we rewrite the bias term to

$$\begin{aligned}\mathrm{E}_\varepsilon[Y]'\Sigma_z\mathrm{E}_\varepsilon[Y] = \lim_{T\to\infty} T^{-1}\beta_0'Z'V'Z'ZVZ\beta_0 = \beta_{w,0}'\Sigma_w\beta_{w,0} + \beta_{x,0}'\Sigma_{xw}\beta_{w,0} \\ +\beta_{w,0}'\Sigma_{wx}\beta_{x,0} + \beta_{x,0}'\Sigma_{xw}\Sigma_w^{-1}\Sigma_{wx}\beta_{x,0} + \beta_{x,0}'\Sigma V_R\Sigma V_R\Sigma\beta_{x,0},\end{aligned} \tag{4.78}$$

where $\Sigma = \mathrm{plim}_{T\to\infty}\frac{1}{T}X'M_wX$, and

$$\begin{aligned}V &= \begin{pmatrix} (W'W)^{-1}W' - (W'W)^{-1}W'XV_RX'M_w \\ V_RX'M_w \end{pmatrix} \\ V_R &= \mathrm{E}_R[R(R'\Sigma R)^{-1}R'], \quad M_w = I_T - P_w, \quad P_w = W(W'W)^{-1}W'.\end{aligned} \tag{4.79}$$

The last term in (4.78) can be lower bounded by $\beta_{x,0}\Sigma'\beta_{x,0}\lambda_{\min}(\Sigma^{1/2}V_R\Sigma^{1/2})^2$, and upper bounded by the same expression with the minimum eigenvalues replaced by maximum eigenvalues.

For the variance, we have

$$\begin{aligned}\mathrm{E}_\varepsilon[(Y-\mathrm{E}_\varepsilon[Y])'\Sigma_z(Y-\mathrm{E}_\varepsilon[Y])] =& \mathrm{E}[\lim_{T\to\infty}\varepsilon'V'Z'ZV\varepsilon] \\ =& \mathrm{E}[\lim_{T\to\infty}\varepsilon'(P_w + T^{-1}M_wXV_R\Sigma V_RX'M_w)\varepsilon] \\ =& \sigma^2p_w + \mathrm{E}[\lim_{T\to\infty}T^{-1}\varepsilon'M_wXV_R\Sigma V_RX'M_w\varepsilon],\end{aligned} \tag{4.80}$$

where we use that $\varepsilon'P_w\varepsilon \xrightarrow{(d)} \sigma^2\chi^2(p_w)$. Since $T^{-1}\varepsilon'M_wX\Sigma^{-1}X'M_w\varepsilon \xrightarrow{(d)} \sigma^2\chi^2(p_x)$, the last term in (4.80) can be lower bounded by $\sigma^2p_x\lambda_{\min}(\Sigma^{1/2}V_R\Sigma^{1/2})^2$.

Using the bounds on $\lambda_{\min}(\Sigma^{1/2}V_R\Sigma^{1/2})$ in Lemma 4.7 in Appendix 4.A.7 together with the expressions for the bias and variance terms in (4.78) and (4.80), we have the following lower bound on the MSFE

$$\begin{aligned}\rho(k) \geq& \beta_{w,0}'\Sigma_w\beta_{w,0} + \beta_{x,0}'\Sigma_{xw}\beta_{w,0} + \beta_{w,0}'\Sigma_{wx}\beta_{x,0} + \beta_{x,0}'\Sigma_{xw}\Sigma_w^{-1}\Sigma_{wx}\beta_{x,0} + \\ & (\beta_{x,0}'\Sigma\beta_{x,0})\frac{\lambda_{\min}(\Sigma)^2}{\lambda_{\max}(\Sigma)^2}\frac{k^2}{p_x^2}\frac{1}{\eta^2} + \sigma^2\left(p_w + \frac{\lambda_{\min}(\Sigma)^2}{\lambda_{\max}(\Sigma)^2}\frac{1}{\eta^2}\frac{k^2}{p_x}\right),\end{aligned} \tag{4.81}$$

which completes the proof. ■

Although in many settings weaker than the bound in Theorem 4.1, we also directly obtain an upper bound on the MSFE:

$$\begin{aligned}\rho(k) \leq & \beta'_{w,0}\Sigma_w\beta_{w,0} + \beta'_{x,0}\Sigma_{xw}\beta_{w,0} + \beta'_{w,0}\Sigma_{wx}\beta_{x,0} + \beta'_{x,0}\Sigma_{xw}\Sigma_w^{-1}\Sigma_{wx}\beta_{x,0} + \\ & (\beta'_{x,0}\Sigma\beta_{x,0})\frac{\lambda_{\max}(\Sigma)^2}{\lambda_{\min}(\Sigma)^2}\frac{k^2}{p_x^2}\eta^2 + \sigma^2\left(p_w + \frac{\lambda_{\max}(\Sigma)^2}{\lambda_{\min}(\Sigma)^2}\frac{k^2}{p_x}\eta^2\right).\end{aligned} \tag{4.82}$$

## 4.A.9 Proof of Theorem 4.2

First, we use Lemma 4.4 in Appendix 4.A.1 to write $\rho_S(k)$ as

$$\begin{aligned}\rho_S(k) &= \mathrm{E}_\varepsilon\left[\lim_{T\to\infty} T\mathrm{E}_{z_T}\left[\left(z'_T\beta - z'_T\frac{1}{N}\sum_{i=1}^{N} S_{R_i}\hat{\beta}_{R_i}\right)^2\right]\right] \\ &= \mathrm{E}_\varepsilon\left[\lim_{T\to\infty} T\left(\beta - \frac{1}{N}\sum_{i=1}^{N} S_{R_i}\hat{\beta}_{R_i}\right)'\Sigma_z\left(\beta - \frac{1}{N}\sum_{i=1}^{N} S_{R_i}\hat{\beta}_{R_i}\right)\right].\end{aligned} \tag{4.83}$$

Define the $p \times 1$ vector $d$ such that,

$$\frac{1}{N}\sum_{i=1}^{N} S_{R_i}\hat{\beta}_{R_i} = \mathrm{E}[S_R\hat{\beta}_R] + \frac{1}{\sqrt{T}}\Sigma_z^{-1/2}\tilde{\epsilon}d. \tag{4.84}$$

Substituting (4.84) into (4.83) yields

$$\rho_S(k) = \rho(k) + \tilde{\epsilon}^2\mathrm{E}_\varepsilon[\lim_{T\to\infty} d'd] - 2\tilde{\epsilon}\mathrm{E}_\varepsilon[\lim_{T\to\infty}\sqrt{T}d'\Sigma_z^{1/2}(\beta - \mathrm{E}_R[S_R\hat{\beta}_R])], \tag{4.85}$$

where $\rho(k) = \mathrm{E}_\varepsilon\left[\lim_{T\to\infty} T(\beta - \mathrm{E}_R[S_R\hat{\beta}_R])'\Sigma_z(\beta - \mathrm{E}_R[S_R\hat{\beta}_R]\right]$ follows again from Lemma 4.4 in Appendix 4.A.1. We upper bound the last term in (4.85) as

$$\begin{aligned}&|2\mathrm{E}_\varepsilon[\lim_{T\to\infty}\sqrt{T}d'\Sigma_z^{1/2}(\beta - \mathrm{E}_R[S_R\hat{\beta}_R])]| \\ &\quad\leq 2\mathrm{E}_\varepsilon\left[\lim_{T\to\infty}\sqrt{Td'd(\beta - \mathrm{E}_R[S_R\hat{\beta}_R])'\Sigma_z(\beta - \mathrm{E}_R[S_R\hat{\beta}_R])}\right] \\ &\quad\leq \mathrm{E}_\varepsilon\left[\lim_{T\to\infty} d'd\right] + \rho(k),\end{aligned} \tag{4.86}$$

where we use the Cauchy-Schwarz inequality in the first line, and $a^2+b^2 > 2ab$ in the second line. Combining (4.85) and (4.86) results in a bound on $\rho_S(k)$;

$$\rho_S(k) \leq (1+\tilde{\epsilon})\rho(k) + (\tilde{\epsilon}+\tilde{\epsilon}^2)\mathrm{E}_\varepsilon[\lim_{T\to\infty} d'd] = (1+\tilde{\epsilon})\left(1 + \frac{\tilde{\epsilon}\mathrm{E}_\varepsilon[\lim_{T\to\infty} d'd]}{\rho(k)}\right)\rho(k).$$

For $\rho_S(k) = (1+\epsilon)\rho(k)$ to hold, we need $\tilde{\epsilon}\mathrm{E}_\varepsilon[\lim_{T\to\infty} d'd]$ to be smaller than the lower bound on $\rho(k)$ which we derive in Appendix 4.A.8.

It suffices to show that

$$\mathrm{E}[\lim_{T\to\infty} d'd] \leq \sigma^2 \left(\frac{\lambda_{\min}(\Sigma)}{\lambda_{\max}(\Sigma)}\right)^2 \frac{k^2}{p_x}. \tag{4.87}$$

We construct an upper bound on $\mathrm{E}[\lim_{T\to\infty} d'd]$ and show that this bound satisfies the bound in (4.87). By definition

$$\begin{aligned} d =& \sqrt{T}\Sigma_z^{1/2}\left(\frac{1}{N}\sum_{i=1}^{N} S_{R_i}\hat{\beta}_{R_i} - \mathrm{E}_R[S_R\hat{\beta}_R]\right) \\ =& \sqrt{T}\Sigma_z^{1/2}\begin{bmatrix} -(W'W)^{-1}W'X\Delta X'M_W \\ \Delta X'M_W \end{bmatrix} y = \sqrt{T}\Sigma_z^{1/2}V_\Delta y, \end{aligned} \tag{4.88}$$

where $\Delta = \frac{1}{N}\sum_{i=1}^{N} R_i(R_i'\Sigma R_i)^{-1}R_i' - \mathrm{E}_R[R(R'\Sigma R)^{-1}R']$ with $\Sigma = X'M_WX$. Then

$$\begin{aligned} \mathrm{E}_\varepsilon[\lim_{T\to\infty} d'd] &= \mathrm{E}_\varepsilon[\lim_{T\to\infty} y'V_\Delta'Z'ZV_\Delta y] = \mathrm{E}_\varepsilon[\lim_{T\to\infty} y'M_wX\Delta\Sigma\Delta X'M_wy] \\ &\leq \lambda_{\max}(\Sigma^{1/2}\Delta\Sigma^{1/2})^2\mathrm{E}_\varepsilon[\lim_{T\to\infty}(Z\beta+\varepsilon)'M_wX\Sigma^{-1}X'M_w(Z\beta+\varepsilon)] \\ &= \lambda_{\max}(\Sigma^{1/2}\Delta\Sigma^{1/2})^2(\beta_0'\Sigma_z\beta_0 + \mathrm{E}_\varepsilon[\lim_{T\to\infty}\varepsilon'M_wX\Sigma^{-1}X'M_w\varepsilon]) \\ &\leq \lambda_{\max}(\Sigma^{1/2}\Delta\Sigma^{1/2})^2(\beta_0'\beta_0\lambda_{\max}(\Sigma_z) + \sigma^2p_x) \\ &\leq c\lambda_{\max}(\Sigma^{1/2}\Delta\Sigma^{1/2})^2\sigma^2p_x, \end{aligned}$$

since $\varepsilon'M_WX(\Sigma)^{-1}X'M_W\varepsilon \xrightarrow{(d)} \sigma^2\chi^2(p_x)$, and $c > 0$ is a constant independent of $p_x$. To satisfy (4.87), we require $\lambda_{\max}(\Sigma^{1/2}\Delta\Sigma^{1/2}) \leq c\frac{k}{p_x}$. We apply the following lemma:

**Lemma 4.9 (Ahlswede and Winter (2002), Theorem 19)** *Let $X_i$ be a $p_x \times p_x$ independent symmetric positive definite matrix with $\lambda_{\max}(X_i) \leq 1$ almost surely and $i = 1, \ldots, N$. Let*

*$S_N = \sum_{i=1}^N X_i$ and $\Omega = \sum_{i=1}^N \lambda_{\max}(E[X_i])$, then for all $\epsilon \in (0,1)$*

$$P\left(\lambda_{\max}(S_N - E[S_N]) \geq \epsilon\Omega\right) \leq 2p\exp(-\epsilon^2\Omega/4). \tag{4.89}$$

This lemma is a non-trivial generalization of a Chernoff bound for sums of independent random variables. For an expository proof, see Section 2 of Wigderson and Xiao (2008). The main technical obstacle is that the proof for scalar random variables relies on the fact that scalars are commutative. To circumvent this, the Golden-Thompson inequality (Golden, 1965; Thompson, 1965) is used.

We define $X_i = \Sigma^{1/2}R_i(R_i'\Sigma R_i)^{-1}R_i'\Sigma^{1/2}$. Since $X_i$ is a projection matrix we have $\lambda_{\max}(X_i) = 1$. We apply Lemma 4.9 and set

$$\Omega = N\lambda_{\max}(\mathrm{E}_R\left[\Sigma^{1/2}R(R'\Sigma R)^{-1}R'\Sigma^{1/2}\right]), \tag{4.90}$$

$$e = \epsilon\lambda_{\max}(\mathrm{E}_R\left[\Sigma^{1/2}R(R'\Sigma R)^{-1}R'\Sigma^{1/2}\right]). \tag{4.91}$$

Then plugging in (4.90) and (4.91) into Lemma 4.9, we obtain

$$\begin{aligned}&\mathrm{P}\left(\lambda_{\max}(\Sigma^{1/2}\Delta\Sigma^{1/2}) \geq \epsilon\lambda_{\max}(\mathrm{E}_R\left[\Sigma^{1/2}R(R'\Sigma R)^{-1}R'\Sigma^{1/2}\right])\right)\\ &\leq 2p_x\exp\left(-\frac{\epsilon^2}{4}N\lambda_{\max}(\mathrm{E}_R\left[\Sigma^{1/2}R(R'\Sigma R)^{-1}R'\Sigma^{1/2}\right])\right).\end{aligned} \tag{4.92}$$

For $\lambda_{\max}(\Sigma^{1/2}\Delta\Sigma^{1/2}) \leq c\frac{k}{p_x}$ to hold, we need $\epsilon\lambda_{\max}(\mathrm{E}_R\left[\Sigma^{1/2}R(R'\Sigma R)^{-1}R'\Sigma^{1/2}\right]) \leq \frac{k}{p_x}$ which is guaranteed by Lemma 4.7 in Appendix 4.A.7. Moreover, the right-hand side of (4.92) needs to be close to zero, which requires for some $\delta \in (0,1)$ that

$$2p_x\exp\left(-\frac{\epsilon^2}{4}N\lambda_{\max}(\mathrm{E}_R\left[\Sigma^{1/2}R(R'\Sigma R)^{-1}R'\Sigma^{1/2}\right])\right) \leq \delta. \tag{4.93}$$

This implies that we need to choose the number of samples

$$N \geq \frac{4}{\epsilon^2\lambda_{\max}(\mathrm{E}_R\left[\Sigma^{1/2}R(R'\Sigma R)^{-1}R'\Sigma^{1/2}\right])}\log\left(\frac{2p_x}{\delta}\right). \tag{4.94}$$

We lower bound $\lambda_{\max}(\mathrm{E}_R\left[\Sigma^{1/2}R(R'\Sigma R)^{-1}R'\Sigma^{1/2}\right])$ by using the lower bound on the minimum eigenvalue in Lemma 4.7 in Appendix 4.A.7, for both random projection and random permutation matrices . We substitute the bound into (4.94). The result is that for both random

permutation matrices and random projection matrices, we need

$$N = O\left(\frac{p_x \log p_x}{k}\right), \tag{4.95}$$

draws. ■

## 4.B Monte Carlo experiments

**Table 4.7:** Monte Carlo simulation: relative MSFE

| | | Random projections - $k$ | | | | Random subsets - $k$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| $s$ | $b$ | 1 | 10 | 25 | 50 | 1 | 10 | 25 | 50 |
| 10 | 0.5 | 0.978 | 1.278 | 3.504 | 11.684 | 0.976 | 1.286 | 3.543 | 11.740 |
| | 0.1 | 0.967 | 0.872 | 1.389 | 3.909 | 0.967 | 0.874 | 1.399 | 3.927 |
| | 2.0 | 0.964 | 0.732 | 0.646 | 1.127 | 0.963 | 0.729 | 0.646 | 1.130 |
| 50 | 0.5 | 0.965 | 0.815 | 1.133 | 3.045 | 0.964 | 0.815 | 1.140 | 3.065 |
| | 0.1 | 0.962 | 0.712 | 0.568 | 0.885 | 0.962 | 0.710 | 0.569 | 0.890 |
| | 2 | 0.962 | 0.684 | 0.415 | 0.304 | 0.961 | 0.681 | 0.413 | 0.306 |
| 100 | 0.5 | 0.963 | 0.750 | 0.781 | 1.694 | 0.962 | 0.748 | 0.783 | 1.705 |
| | 0.1 | 0.962 | 0.693 | 0.463 | 0.493 | 0.962 | 0.690 | 0.462 | 0.496 |
| | 2.0 | 0.961 | 0.675 | 0.379 | 0.194 | 0.961 | 0.670 | 0.376 | 0.194 |
| | | Principal components - $k$ | | | | Partial least squares - $k$ | | | |
| $s$ | $b$ | 1 | 10 | 25 | 50 | 1 | 10 | 25 | 50 |
| 10 | 0.5 | 1.253 | 3.736 | 8.780 | 19.402 | 9.592 | 40.512 | 48.769 | 51.515 |
| | 0.1 | 1.056 | 1.665 | 3.073 | 6.297 | 3.099 | 13.265 | 15.882 | 16.731 |
| | 2.0 | 0.972 | 0.950 | 1.152 | 1.828 | 0.961 | 3.472 | 4.186 | 4.425 |
| 50 | 0.5 | 1.034 | 1.424 | 2.422 | 4.979 | 2.377 | 10.253 | 12.409 | 13.107 |
| | 0.1 | 0.972 | 0.900 | 0.962 | 1.428 | 0.805 | 2.693 | 3.248 | 3.415 |
| | 2.0 | 0.966 | 0.739 | 0.537 | 0.457 | 0.432 | 0.685 | 0.856 | 0.907 |
| 100 | 0.5 | 0.983 | 1.095 | 1.529 | 2.742 | 1.372 | 5.506 | 6.647 | 7.002 |
| | 0.1 | 0.968 | 0.778 | 0.677 | 0.775 | 0.535 | 1.364 | 1.683 | 1.765 |
| | 2.0 | 0.958 | 0.685 | 0.440 | 0.276 | 0.356 | 0.329 | 0.409 | 0.431 |
| | | Ridge regression - $\ln k$ | | | | Lasso - $\ln k$ | | | |
| $s$ | $b$ | -6 | -4 | -2 | 0 | -28 | -27 | -26 | -25 |
| 10 | 0.5 | 0.995 | 0.971 | 1.108 | 4.821 | 1.073 | 3.684 | 11.159 | 25.064 |
| | 0.1 | 0.993 | 0.953 | 0.864 | 1.787 | 0.959 | 1.553 | 3.793 | 8.342 |
| | 2.0 | 0.992 | 0.943 | 0.768 | 0.705 | 0.796 | 0.708 | 1.187 | 2.280 |
| 50 | 0.5 | 0.993 | 0.949 | 0.826 | 1.428 | 0.961 | 1.378 | 3.073 | 6.522 |
| | 0.1 | 0.992 | 0.940 | 0.750 | 0.594 | 0.844 | 0.713 | 1.004 | 1.789 |
| | 2.0 | 0.990 | 0.925 | 0.686 | 0.326 | 0.617 | 0.393 | 0.377 | 0.519 |
| 100 | 0.5 | 0.993 | 0.944 | 0.781 | 0.915 | 0.921 | 1.011 | 1.827 | 3.600 |
| | 0.1 | 0.991 | 0.934 | 0.721 | 0.423 | 0.767 | 0.547 | 0.610 | 0.950 |
| | 2.0 | 0.988 | 0.907 | 0.633 | 0.241 | 0.506 | 0.299 | 0.242 | 0.280 |

Note: this table shows the MSFE relative to the prevailing mean, for random projection regression, random subset regression, principal component regression, partial least squares, ridge regression, and lasso under the data generating process (4.27) based on 10,000 replications, for increasing values of the subspace dimension $k$. The coefficient size varies over $b = \{0.5, 1.0, 2.0\}$, and $s = \{10, 50, 100\}$ out of $p = 100$ coefficients are non-zero.

**Table 4.8:** Monte Carlo simulation: relative MSFE under a factor design

| | | Random projections - $k$ | | | | Random subsets - $k$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| $s$ | $b$ | 1 | 10 | 25 | 50 | 1 | 10 | 25 | 50 |
| Top | 0.5 | 0.944 | 0.722 | 1.243 | 3.931 | 0.991 | 0.955 | 1.136 | 2.591 |
| | 0.1 | 0.937 | 0.558 | 0.444 | 1.029 | 0.990 | 0.915 | 0.844 | 1.013 |
| | 2.0 | 0.935 | 0.513 | 0.233 | 0.291 | 0.990 | 0.902 | 0.764 | 0.598 |
| Int. | 0.5 | 1.013 | 1.841 | 5.724 | 19.064 | 0.998 | 1.199 | 2.735 | 10.897 |
| | 0.1 | 1.003 | 1.305 | 2.739 | 7.565 | 0.992 | 1.013 | 1.507 | 4.481 |
| | 2.0 | 1.001 | 1.075 | 1.390 | 2.418 | 0.991 | 0.934 | 0.961 | 1.604 |
| | | Principal components - $k$ | | | | Partial least squares - $k$ | | | |
| $s$ | $b$ | 1 | 10 | 25 | 50 | 1 | 10 | 25 | 50 |
| Top | 0.5 | 0.996 | 1.097 | 2.905 | 6.486 | 2.466 | 13.526 | 16.322 | 17.152 |
| | 0.1 | 0.917 | 0.300 | 0.749 | 1.685 | 0.495 | 3.461 | 4.260 | 4.470 |
| | 2.0 | 0.886 | 0.078 | 0.202 | 0.448 | 0.139 | 0.947 | 1.156 | 1.206 |
| Int. | 0.5 | 1.501 | 6.065 | 14.467 | 31.146 | 16.347 | 65.901 | 77.865 | 82.446 |
| | 0.1 | 1.176 | 2.948 | 6.438 | 12.808 | 7.140 | 24.905 | 29.846 | 31.725 |
| | 2.0 | 1.060 | 1.639 | 2.770 | 4.172 | 2.969 | 7.333 | 8.545 | 9.048 |
| | | Ridge regression - $\ln k$ | | | | Lasso - $\ln k$ | | | |
| $s$ | $b$ | -6 | -4 | -2 | 0 | -28 | -27 | -26 | -25 |
| Top | 0.5 | 0.989 | 0.918 | 0.734 | 1.675 | 0.887 | 1.729 | 4.143 | 9.125 |
| | 0.1 | 0.987 | 0.903 | 0.614 | 0.527 | 0.539 | 0.577 | 1.142 | 2.399 |
| | 2.0 | 0.984 | 0.880 | 0.531 | 0.206 | 0.194 | 0.166 | 0.312 | 0.661 |
| Int. | 0.5 | 1.001 | 1.023 | 1.486 | 7.887 | 1.796 | 7.268 | 19.791 | 43.692 |
| | 0.1 | 1.000 | 1.007 | 1.178 | 3.556 | 1.335 | 3.400 | 7.835 | 16.719 |
| | 2.0 | 1.000 | 1.003 | 1.049 | 1.577 | 1.114 | 1.512 | 2.543 | 4.954 |

Note: this table shows the MSFE relative to the prevailing mean, for random projection regression, random subset regression, principal component regression, partial least squares, ridge regression, and lasso in the Monte Carlo simulations when the underlying model has a factor structure. In the experiments referred to with 'High', we associate nonzero coefficients with the 10 factors that explain most of the variation in the predictors. In the remaining experiments referred to with 'Int.' we associate the nonzero coefficients with intermediate factors $\{f_{46}, \ldots, f_{55}\}$. For additional information, see the note following Table 4.7.

# Chapter 5

# Inference In High-Dimensional Linear Regression Models

*Joint work with Tom Boot*

## 5.1 Introduction

Different scientific fields are currently confronted with data sets where the number of explanatory variables approaches, or even exceeds the number of available observations. This is routinely observed in research using genetic data, but also occurs in economics, where cross-sectional datasets on economic growth such as Barro and Lee (1993) are bounded by the number of countries. The resulting rank deficiency of the empirical covariance matrix calls for new methods to obtain valid standard errors.

Estimation of high-dimensional models has been intensively studied in recent years. Well-known estimators include ridge regression (Hoerl and Kennard, 1970), lasso (Tibshirani, 1996), adaptive lasso (Zou, 2006), the Dantzig selector (Candes and Tao, 2007), and penalized likelihood methods by (Fan et al., 2004). The adequacy of these estimators is often argued through oracle inequalities established among others by Bickel et al. (2009); Candes and Tao (2007); Meinshausen and Yu (2009); Van de Geer (2008). An overview of theoretical results for the lasso is provided by Bühlmann and Van De Geer (2011). As the distribution of aforementioned estimators seems intractable, the construction of standard errors and valid confidence intervals remains a challenging problem.

We develop an asymptotically unbiased estimator for the full high-dimensional parameter vector in a linear regression model where the number of variables $p$ greatly exceeds the number of observations $n$. The estimator is accompanied by a closed form expression for the covariance matrix of the estimated parameters which is free of tuning parameters. This enables the construction of uniformly valid confidence intervals, hypothesis testing, and efficient adjustments for multiple testing. Standard errors are shown to decrease at the familiar $n^{-1/2}$ rate.

The estimator uses a diagonally scaled Moore-Penrose pseudoinverse to obtain parameter estimates, and implements a bias correction based on the lasso. The scaled Moore-Penrose pseudoinverse approximates the inverse of the singular high-dimensional covariance matrix of the regressors, and the lasso corrects for the bias resulting from this approximation. The remaining bias can be factorized into a term which reflects the accuracy of the pseudoinverse, and a term measuring the lasso estimation error. The product of these two components is of lower order compared to the variance of the estimator, yielding an asymptotically unbiased estimator. The proof relies on several extensions of the results of Fan and Lv (2008) and Wang and Leng (2015), who use the Moore-Penrose pseudoinverse to set up a variable screening technique.

Using the Moore-Penrose pseudoinverse is especially effective when the number of variables is much larger than the number of observations. If $p$ is relatively close to $n$, regularization of the inverse can reduce the standard errors while the bias remains negligible. This motivates an extension to two regularized variants of the Moore-Penrose pseudoinverse; random least squares and ridge regularization. For a suitable choice of the regularization parameters, these estimators yield smaller standard errors while maintaining the same theoretical validity.

Random least squares projects the columns of the regressor matrix onto a low-dimensional subspace by post-multiplying with a matrix with independently standard normally distributed elements. Repeatedly applying this procedure yields an estimate of the full parameter vector. Mean squared error properties of this estimator are studied by Maillard and Munos (2009) based on the lemma by Johnson and Lindenstrauss (1984), and refined by Kabán (2014). We show that random least squares results in a form of generalized ridge regularization on the empirical covariance matrix. The regularization strength is inversely related to the projection dimension, which should be chosen close to the sample size.

The second regularization method we consider is ridge regularization. In order to show that the bias of the estimator remains sufficiently small, we exploit the relation of the ridge regularized estimator to the Moore-Penrose inverse when the regularization strength is small. This extends the results of Bühlmann et al. (2013) to a setting with random design and possible non-gaussian errors.

The results depend on a sparsity assumption with regard to the high-dimensional parameter vector, and a mild restriction on the distributional class of the regressor matrix. We assume the sparsity of the parameter vector to be of the same order as in recent studies on high-dimensional inference by Zhang and Zhang (2014), van de Geer et al. (2014) and Javanmard and Montanari (2014). Furthermore, we require the rows of the regressor matrix to be generated from the class of elliptical distributions. This class includes the multivariate normal, power exponential and Student's t-distribution. We allow for correlation between and within the regressors, and for both gaussian and non-gaussian regression errors.

Our approach builds upon Bühlmann et al. (2013), Zhang and Zhang (2014), van de Geer et al. (2014), and Javanmard and Montanari (2014). The estimator proposed by Bühlmann et al. (2013) shares important properties with the ridge regularized estimator discussed above. However, the analysis there considers fixed regressors and gaussian errors, and the results on hypothesis testing appear conservative. We also propose a different diagonal scaling factor, ensuring that the estimator is asymptotically unbiased. Under an additional sparsity assumption on the elements of the inverse covariance matrix, Zhang and Zhang (2014) and van de Geer et al. (2014) use the lasso for each column of the regressor matrix to estimate the inverse covariance matrix. As an alternative, Javanmard and Montanari (2014) rely on direct numerical optimization to find an accurate approximate inverse. These methods lead to standard errors which depend on one or more additional regularization parameters that potentially influence the results.

We consider situations where interest lies in performing inference on the full high-dimensional parameter vector. Alternatively, one can focus on a low-dimensional subvector of the high-dimensional parameter vector. A sequence of papers (Belloni et al., 2013, 2010; Chernozhukov et al., 2015) introduces a multistage procedure that uses the lasso to select control variables in such a way that variable selection errors do not affect the distribution of the estimates of interest. This approach is effective when both the number of control variables related to the dependent variable, as well as the number of control variables related to

the variables of interest, are limited. Strengthening this assumption such that every variable is correlated with only a small number of the remaining variables, Lan et al. (2016) provide a method to construct confidence intervals for the full parameter vector.

In this chapter, we do not limit inference to a low-dimensional subvector of the parameter vector. The proposed method does not require the assumption that only a small number of control variables is related to the variables of interest. This relaxation might come at the cost of a potential power loss, although this is not reflected in the convergence rate of the estimator.

We confirm our theoretical results with a set of Monte Carlo experiments. We vary the specification of the covariance matrix, the amount of sparsity of the parameter vector, and the signal strength. In line with the theoretical results, we find that even in small samples where the number of regressors is twice the number of observations, coverage rates are close to the nominal rate of 95%. Random least squares and ridge regression yield narrower confidence intervals compared to using a Moore-Penrose pseudoinverse, but this comes at the expense of a slight downward bias. Coverage rates are substantially closer to the nominal rate compared to existing methods.

To compare our findings to existing results, we consider the empirical application of Bühlmann et al. (2014), van de Geer et al. (2014) and Javanmard and Montanari (2014). In this application, the riboflavin production in Bacillus subtilis is explained by gene expression levels. We find two genes to be significant that are different from previous findings, while one significant gene has been found by Javanmard and Montanari (2014) as well.

The outline of the chapter is as follows. Section 5.2 introduces the estimation approach and the proposed estimators. The theoretical properties of the Moore-Penrose pseudoinverse, random least squares, and ridge regression are presented in Section 5.3. Section 5.4 illustrates these results through Monte Carlo simulations and Section 5.5 applies the methods on the riboflavin data. Section 5.6 concludes.

**Notation** We use the following notation throughout the chapter: For any $n \times 1$ vector $a = (a_1, \ldots, a_n)'$, the $l_q$-norm is defined as $||a||_q := (\sum_{i=1}^n |a_i|^q)^{1/q}$ for $q > 0$ and $||a||_0$ denotes the number of nonzero elements of $a$. The maximum norm is written as $||a||_\infty = \max(|a_1|, \ldots, |a_n|)$. For a $p \times n$ matrix $A$, the $l_q$-norm is defined as $||A||_q := \sup_{x, ||x||_q = 1} \{||Ax||_q\}$ and the maximum norm is written as $||A||_{\max} = \max_{i=1,\ldots,n, j=1,\ldots,p} |A_{ij}|$. The $n \times n$ identity matrix is denoted by $I_n$. The vector $e_i$ has

its $i$-th entry equal to 1 and zeros everywhere else. For the regressor matrix $X$, we index the rows with the subscript $i = 1, \ldots, n$ and the columns with the subscript $j = 1, \ldots, p$. If $U$ is a $p \times p$ orthogonal matrix, we write $U \in \mathcal{O}(p)$. When two random variables $X$ and $Y$ follow the same distribution, this is denoted as $X \overset{(d)}{=} Y$.

## 5.2 High-dimensional linear regression

Consider the data generating process

$$y = X\beta + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 I_n), \tag{5.1}$$

where $y$ is an $n \times 1$ response vector, $X$ an $n \times p$ regressor matrix, $\beta = (\beta_1, \ldots, \beta_p)'$ a $p \times 1$ vector of unknown regressor coefficients, and $\varepsilon$ an $n \times 1$ vector of errors which are independent and normally distributed with variance $\sigma^2$. The empirical covariance matrix of $X$ is denoted by $\hat{\Sigma} = \frac{1}{n} X'X$. We will show how the normality assumption on the errors can be relaxed.

### 5.2.1 Approximate inverse and bias correction

Define $M$ as a $p \times n$ matrix for which $MX$ is close to the $p \times p$ identity matrix $I_p$, in a sense that will be made precise below. We refer to $M$ as an approximate inverse for $X$.

We start by considering estimators for $\beta$ of the form

$$\begin{aligned} \hat{\beta} &= My \\ &= MX\beta + M\varepsilon \\ &= \beta + (MX - I_p)\,\beta + M\varepsilon. \end{aligned} \tag{5.2}$$

The second term of (5.2) represents a bias which depends on the accuracy of the approximate inverse $M$. When $p \leq n$, ordinary least squares yields unbiased estimates by choosing $M = (X'X)^{-1}X'$. When $p > n$, the matrix $X'X$ is singular, and we have to resort to an expression for $M$ for which the bias is not equal to zero.

Suppose we have an accurate initial estimator $\hat{\beta}^{\text{init}}$, then we can reduce the bias in (5.2) by applying a correction

$$\begin{aligned}\hat{\beta}^c &= My - (MX - I_p)\,\hat{\beta}^{\text{init}} \\ &= \beta + (MX - I_p)\left(\beta - \hat{\beta}^{\text{init}}\right) + M\varepsilon.\end{aligned} \tag{5.3}$$

For the initial estimator $\hat{\beta}^{\text{init}}$ we take the lasso estimator of Tibshirani (1996). Alternative initial estimators can be used, as long as they satisfy a sufficiently tight accuracy bound on the $l_1$ norm of $\beta - \hat{\beta}^{\text{init}}$.

The goal of this chapter is to introduce choices of $M$ such that the bias of the estimator $\hat{\beta}^c$ is of lower order than the variance. Anticipating the usual $\sqrt{n}$ rate of convergence, we rescale the estimator in (5.3) as

$$\begin{aligned}\sqrt{n}\left(\hat{\beta}^c - \beta\right) &= \Delta + Z \\ \Delta &= \sqrt{n}\,(MX - I_p)\left(\beta - \hat{\beta}^{\text{init}}\right) \\ Z &= \sqrt{n}M\varepsilon\end{aligned} \tag{5.4}$$

The term $\Delta$ reflects the bias of the corrected estimator. To ensure asymptotic unbiasedness, $\Delta$ should be of lower order than the noise term $Z$. We propose specifications for the approximate inverse $M$ for which $Z|X \sim N(0, \sigma^2\Omega)$ with $\Omega = nMM'$ and the variance $\Omega_{jj} = O_p(1)$. This shows that the standard errors of the estimator $\hat{\beta}^c$ decrease at the familiar $n^{-1/2}$ rate.

In order for the bias to vanish compared to the variance term, given that $\Omega_{jj} = O_p(1)$, we now need $||\Delta||_\infty = o_p(1)$. Under a sparsity assumption on $\beta$, we show that this is indeed the case, which implies that $\hat{\beta}^c$ is an asymptotically unbiased estimator. Combined with a closed-form expression for the covariance matrix $\Omega$, confidence intervals can be constructed for the $j$-th parameter as

$$\left[\hat{\beta}^c_j - z_{\alpha/2}\sqrt{\sigma^2 m_j' m_j},\quad \hat{\beta}^c_j + z_{\alpha/2}\sqrt{\sigma^2 m_j' m_j}\right], \tag{5.5}$$

where $m_j'$ is the $j$th row of $M$ and $z_{\alpha/2}$ is the $\alpha/2$ critical value for the standard normal distribution. We discuss estimation of $\sigma$ in Section 5.2.3.

The estimator defined in (5.3) occurs in a different form in Zhang and Zhang (2014), van de Geer et al. (2014) and Javanmard and Montanari (2014), who consider $\hat{\beta}^c = \hat{\beta}^{\text{lasso}} + \frac{1}{n}\bar{M}X'(y - X\hat{\beta}^{\text{lasso}})$. This leads to an interpretation of $\hat{\beta}^c$ as a 'desparsified' version of the lasso estimator. An alternative to the standard lasso estimator is put forward by Caner and Kock (2014). The matrix $\bar{M}$ serves as an approximate inverse to the empirical covariance matrix $\frac{1}{n}X'X$, which is found by a series of lasso regressions in Zhang and Zhang (2014) and van de Geer et al. (2014), or direct numerical optimization in Javanmard and Montanari (2014). As a consequence of the complex estimation procedures, standard errors are not available in closed form, and their validity depends on the appropriate selection of one or more tuning parameters.

### 5.2.2 Choosing the approximate inverse $M$

This section proposes specifications of $M$ for which the bias $||\Delta||_\infty$ in (5.4) is small. We ensure that the diagonal terms of $MX - I_p$ are identically equal to zero by introducing a $p \times p$ diagonal matrix $D$, with diagonal elements $d_j$, and taking

$$M = D\tilde{M}, \qquad d_j = (\tilde{m}_j' x_j)^{-1}, \tag{5.6}$$

with $\tilde{m}_j'$ the $j$-th row of $\tilde{M}$. We first choose $M$ in the form defined in (5.6), with $\tilde{M}$ specified as the Moore-Penrose pseudoinverse of $X$. Subsequently, we consider regularized alternatives obtained by random least squares and ridge regression.

#### The Moore-Penrose pseudoinverse

A tuning parameter free choice for $\tilde{M}$ in (5.6) is the Moore-Penrose pseudoinverse. When $p \leq n$, and the columns of $X$ are linearly independent, $\tilde{M} = (X'X)^{-1}X'$. In the high-dimensional setting where $p > n$, the matrix $X$ has linearly dependent columns by default. In this case the pseudoinverse equals $X'(XX')^{-1}$, and

$$M^{\text{MPI}} = D^{\text{MPI}}X'(XX')^{-1}. \tag{5.7}$$

The diagonal elements $d_j^{\text{MPI}}$ of the diagonal scaling matrix $D^{\text{MPI}}$ equal

$$d_j^{\text{MPI}} = \left[x_j'(XX')^{-1}x_j\right]^{-1}. \tag{5.8}$$

This provides a closed-form expression for the approximate inverse. In addition, since the bias term of the estimator is of lower order compared to the variance, the covariance of $\hat{\beta}^c$ is available in closed form as well,

$$V(\hat{\beta}^c) = \sigma^2 D^{\text{MPI}} X'(XX')^{-2} X D^{\text{MPI}}. \tag{5.9}$$

**Regularizing the Moore-Penrose pseudoinverse**

The accuracy of the Moore-Penrose pseudoinverse depends on the concentration of the eigenvalues of the matrix $XX'$, which can be weak when $p$ is close to $n$. Regularizing the approximate inverse can improve in accuracy, with smaller standard errors as a result. This section introduces two regularization techniques, for which Section 5.3 shows the appropriate choice for the regularization strength.

**Random Least Squares** This method is based on projecting the high-dimensional regressor matrix $X$ onto a $k < n$ dimensional subspace by post-multiplying with a $p \times k$ matrix $R$ with independently standard normally distributed elements,

$$R_{jl} \sim N(0,1), \quad j = 1, \dots p, \quad l = 1, \dots, k. \tag{5.10}$$

The multiplication yields a low-dimensional analogue to (5.1),

$$y = XR\gamma_R + u. \tag{5.11}$$

Least squares estimation of $\gamma_R$ is straightforward as

$$\hat{\gamma}_R = (R'X'XR)^{-1}R'X'y, \tag{5.12}$$

from which an estimator for $\beta$ can be constructed by $\hat{\beta}_R = R\hat{\gamma}_R$. Since $R$ is random, Jensen's inequality can be used to show that the accuracy of this estimator can be improved

by averaging over different realizations of $R$. We then arrive at the following estimator of $\beta$,

$$\hat{\beta}_{\bar{R}} = \mathrm{E}_R[R\hat{\gamma}_R] = \mathrm{E}_R[R(R'X'XR)^{-1}R']X'y. \tag{5.13}$$

From equation (5.13), we recognize that random least squares yields an approximate inverse covariance matrix of $X$. Defining $\tilde{M} = \mathrm{E}_R[R(R'X'XR)^{-1}R']X'$ in (5.6) yields

$$M^{\text{RLS}} = D^{\text{RLS}}\mathrm{E}_R\left[R(R'X'XR)^{-1}R'\right]X', \tag{5.14}$$

with

$$d_j^{\text{RLS}} = \left\{\mathrm{E}_R[r_j'(R'X'XR)^{-1}R']X'x_j\right\}^{-1}. \tag{5.15}$$

**Ridge regression** An alternative regularization strategy is to use a ridge adjustment,

$$M^{\text{RID}} = D^{\text{RID}}(X'X + \gamma I_p)^{-1}X', \tag{5.16}$$

where $\gamma$ denotes the ridge penalty and the elements of the diagonal scaling matrix $D^{\text{RID}}$ equal

$$d_j^{\text{RID}} = \left(v_j'X'x_j\right)^{-1}, \tag{5.17}$$

with $v_j$ the $j$-th row of $(X'X + \gamma I_p)^{-1}$.

The regularization in (5.16) can be related to the Moore-Penrose pseudoinverse, since the latter is defined as

$$\begin{aligned} X'(XX')^{-1} &= \lim_{\gamma\to 0}\left(X'X + \gamma I_p\right)^{-1}X' \\ &= \lim_{\gamma\to 0}X'\left(XX' + \gamma I_n\right)^{-1}. \end{aligned} \tag{5.18}$$

which can be shown using the singular value decomposition of $X$ as in Albert (1972).

### 5.2.3 Estimation of the noise level

A consistent estimator of the noise level $\sigma^2$ is crucial to construct valid confidence intervals. Existing methods, such as van de Geer et al. (2014) and Javanmard and Montanari (2014)

rely on the scaled lasso developed by Sun and Zhang (2012), for which holds that $\left|\frac{\hat{\sigma}}{\sigma} - 1\right| = o_p(1)$ under Assumption A5.1 and Assumption A5.2 discussed in Section 5.3.1.

However, in the Monte Carlo simulations in Section 5.4, and in line with findings by Reid et al. (2016), we find the scaled lasso to be unreliable in many settings. An alternative is to use

$$\hat{\sigma}^2_{\text{lasso}} = \frac{1}{n - \hat{s}} \hat{\varepsilon}' \hat{\varepsilon}, \tag{5.19}$$

with $\hat{s}$ the number of non-zero coefficients retained by the lasso, and $\hat{\varepsilon}$ the $n \times 1$ vector of lasso regression errors. Corresponding to the results in Reid et al. (2016), we find that this leads to more robust estimation of the noise level.

## 5.3 Theoretical results

This section provides the main results of the chapter. Proofs for the theorems in this section are given in Appendix 5.B.

### 5.3.1 Assumptions

Performing inference in a linear regression model with more variables than observations requires additional assumptions over its low-dimensional counterpart. Our assumptions parallel Fan and Lv (2008) and Wang and Leng (2015). We will provide a discussion below.

**Assumption 5.1** *The sparsity $s_0 = ||\beta||_0$ satisfies $s_0 = o\left(\frac{\sqrt{n}}{\log p}\right)$.*

**Assumption 5.2** *The regressor matrix $X$ is generated from an elliptical distribution, i.e.*

$$X = \Sigma_1^{1/2} Z \Sigma_2^{1/2} = \Sigma_1^{1/2} V S U' \Sigma_2^{1/2}, \tag{5.20}$$

*where the $n \times n$ population covariance matrix $\Sigma_1$ and the $p \times p$ population covariance matrix $\Sigma_2$ determine the dependence between the rows and columns of $X$, respectively. The elements of the $n \times p$ matrix $Z$ are generated independently from a spherically symmetric distribution, $V \in \mathcal{O}(n)$, $S$ is an $n \times p$ matrix of singular values, and $U \in \mathcal{O}(p)$.*

*Furthermore,*

$$P\left(\lambda_{\max}(p^{-1} Z Z') \geq c_Z, \quad \lambda_{\min}(p^{-1} Z Z') \leq c_Z^{-1}\right) \leq e^{-C_Z n}, \tag{5.21}$$

*where $\lambda_{\max}(.)$ and $\lambda_{\min}(.)$ are the largest and smallest eigenvalues of a matrix respectively, and $c_Z, C_Z$ are positive constants.*

**Assumption 5.3** *For both the population covariance matrices $\Sigma_1$ and $\Sigma_2$, the eigenvalues are bounded by a constant, i.e. for $i = 1, 2$,*

$$0 < c_{i,1} \leq \lambda_{\min}(\Sigma_i) \leq \lambda_{\max}(\Sigma_i) \leq c_{i,2} < \infty. \tag{5.22}$$

Assumption A5.1 imposes a sparsity constraint which restricts the number of non-zero coefficients in $\beta$ by $s_0 = ||\beta||_0$. For lasso consistency, it is required that $s_0^2 = o\left(n/\log p\right)$. As noted in van de Geer et al. (2014) and Javanmard and Montanari (2014), a slightly stronger assumption is needed when constructing confidence intervals.

In recent work, for example by Chernozhukov et al. (2015), assumption A5.1 is relaxed to allow for approximate sparsity, arguably a more realistic assumption in practical applications. This restricts only the number of large non-zero coefficients, and allows the remaining coefficients to be sufficiently small. Since our results only depend on the $l_1$ norm of the lasso estimation error, which does not change under approximate sparsity, they remain valid under approximate sparsity.

Assumption A5.2 requires that the regressors are generated from an elliptical distribution. The class of elliptical distributions includes the multivariate normal distribution, but also allows for heavier tailed distributions such as the power exponential distribution and the multivariate $t$ distribution (Serfling, 2006; Dasgupta et al., 2012). This class precludes $X$ to consist of binomial variables. However, our results rely on the distribution of the elements of $X'(XX')^{-1}X$, which consist of sums of binomial variables. It is possible that one can use the convergence of these sums towards a normal distribution to extend the results towards binomial regressors.

The matrices $\Sigma_1$ and $\Sigma_2$ in Assumption A5.2 allows for dependence between the rows and the columns of $X$, respectively. Assumption A5.3 states that the eigenvalues of these population covariance matrices are finite and independent of the dimensions $n$ and $p$. This assumption can be relaxed by replacing $c_{i,2}$ with $c_{i,2}n^{\alpha}$. The standard errors then decrease at the rate of $1/\sqrt{n^{1-\alpha}}$ instead of $1/\sqrt{n}$.

### 5.3.2 Asymptotic unbiasedness and normality using the Moore-Penrose pseudoinverse

To prove that $\hat{\beta}^c$ in (5.3) based on the Moore-Penrose pseudoinverse is an asymptotically unbiased estimator, we show that with high probability the bias term in (5.4) is small and of lower order than the noise. Moreover, the construction of confidence intervals as in (5.5) requires $Z|X$ to follow a normal distribution. Efficiency of the estimator is ensured by showing that the standard errors decrease at the usual $n^{-1/2}$ rate.

The first requirement follows from bounding the bias term of the estimator in (5.4) by a norm inequality,

$$||\Delta||_\infty \leq \sqrt{n}\,||MX - I_p||_{\max}\,||\beta - \hat{\beta}^{\text{init}}||_1, \tag{5.23}$$

which is an element-wise bound on $MX - I_p$ together with an $l_1$ accuracy bound on $\beta - \hat{\beta}^{\text{init}}$.

The following lemma bounds on the first term in probability.

**Lemma 5.1** *Suppose Assumption A5.2 and A5.3 hold. Define* $M^{MPI} = D^{MPI}X'(XX')^{-1}$ *with* $D^{MPI}$ *a diagonal matrix with elements* $d_j^{MPI} = (x_j'(XX')^{-1}x_j)^{-1}$*, then we have*

$$P\left(\left|\left|M^{MPI}X - I_p\right|\right|_{\max} \geq a\sqrt{\frac{\log p}{n}}\right) = O(p^{-\tilde{c}}), \tag{5.24}$$

*with* $\tilde{c} = \frac{c}{2c_s}a^2 - 2$ *where* $a, c, c_s > 0$.

A proof is presented in Appendix 5.B.1. Note that the diagonal elements of $M^{\text{MPI}}X - I_p$ are identically zero, due to the diagonal scaling with $D^{\text{MPI}}$. Lemma 5.1 is therefore a statement on the off-diagonal elements of $M^{\text{MPI}}X - I_p$.

Next we show that the $l_1$ norm of the initial estimation error, in the second term in the bound for $||\Delta||_\infty$ in (5.23), is bounded with high probability. As the initial estimator we use the lasso estimator by Tibshirani (1996), which is defined as

$$\hat{\beta}^{\text{lasso}} = \arg\min_b \left[\frac{1}{n}(y - Xb)'(y - Xb) + \lambda||b||_1\right]. \tag{5.25}$$

The following bound applies to the $l_1$-error of the lasso estimator.

**Lemma 5.2** *Suppose Assumption A5.1 and Assumption A5.2 hold. Consider the lasso estimator* (5.25) *with* $\lambda \geq 8\sigma\sqrt{\frac{\log p}{n}}$*, then with probability exceeding* $1 - 2p^{-1}$ *we have*

$$\left|\left|\beta - \hat{\beta}^{lasso}\right|\right|_1 = O_p\left(s_0\sqrt{\frac{\log p}{n}}\right). \tag{5.26}$$

A proof is presented in Appendix 5.B.2. As shown in Bühlmann and Van De Geer (2011), this bound applies under a so-called compatibility condition on $X$. The proof amounts to showing that the compatibility condition is indeed satisfied under Assumption A5.1 and Assumption A5.2.

Combining Assumption A5.1, Lemma 5.1, and Lemma 5.2, we see that the bias can be bounded by

$$||\Delta||_\infty = O_p\left(s_0\frac{\log p}{\sqrt{n}}\right) = o_p(1). \tag{5.27}$$

In order for the estimator to be asymptotically unbiased, it is necessary that the bias in (5.27) is of lower order than the noise term of the estimator, given by $Z$ in (5.4). The following lemma states that this is indeed the case.

**Lemma 5.3** *Suppose Assumption A5.2 and A5.3 hold. For* $j = 1, \ldots, p$ *we have*

$$\begin{aligned} Z_j &= \sqrt{n}d_j^{MPI}x_j'(XX')^{-1}\varepsilon, \\ Z_j|X &\sim N(0, \sigma^2\Omega_j), \\ ||\Omega_{jj}||_2 &= O_p(1), \end{aligned} \tag{5.28}$$

*where* $\Omega_{jj} = nm_j'm_j$ *with* $m_j'$ *the* $j$*-th row of* $M^{MPI} = D^{MPI}X'(XX')^{-1}$ *and* $D^{MPI}$ *a diagonal matrix with* $d_j^{MPI} = [x_j'(XX')^{-1}x_j]^{-1}$.

A proof is presented in Appendix 5.B.3. Appendix 5.B.4 shows that under additional assumptions this result also holds for independent and identically distributed errors $\varepsilon_i$.

Combining Lemma 5.3 with (5.27) yields the central theorem of this chapter.

**Theorem 5.1** *Suppose A5.1-A5.3 hold. Let* $\hat{\beta}^c = My - (MX - I_p)\,\hat{\beta}^{init}$*, with* $\hat{\beta}^{init}$ *such that* $||\hat{\beta}^{init} - \beta||_1 = O_p\left(s_0\sqrt{\log(p)/n}\right)$*, and take* $M$ *as*

$$M^{MPI} = D^{MPI}X'(XX')^{-1},$$

*where $D^{MPI}$ is a diagonal matrix with elements $d_j^{MPI} = \left[x_j'(XX')^{-1}x_j\right]^{-1}$. Then,*

$$\sqrt{n}(\hat{\beta}^c - \beta) = Z + o_p(1),$$
$$Z|X \sim N\left(0, \sigma^2\Omega\right),$$

*where $\Omega = nM^{MPI}M^{MPI'}$ and $\Omega_{jj} = O_p(1)$.*

This theorem shows that the estimator $\hat{\beta}^c$ in (5.3) is asymptotically unbiased with covariance matrix $\Omega$, and standard errors that decrease at the usual $n^{-1/2}$ rate. Theorem 5.1 allows for the construction of confidence intervals that are uniformly valid over $j$. Uniformity is guaranteed since the bound on the lasso estimator given in Lemma 5.2 holds uniformly over all sets $S_0$ of size $s_0 = o(\sqrt{n}/\log p)$, see van de Geer et al. (2014) for a discussion.

Since the resulting covariance matrix of the estimator is available in closed form, efficient multiple testing procedures as in Bühlmann et al. (2013) can be employed, together with joint tests on estimated coefficients, as well as confidence intervals around predictions for future values of the dependent variable.

### 5.3.3 Regularized approximate inverse

When the number of variables is of the same order as the number of observations, the concentration of the eigenvalues in Assumption A5.2 might not be very tight. In this case, regularization of the pseudoinverse can increase the accuracy. We therefore analyze two regularization approaches.

**Random least squares** The key to the behavior of the regularized covariance matrix in repeated least squares, is the projection dimension $k$. The following lemma parallels Lemma 5.1 and Lemma 5.3 for an appropriate choice of the projection dimension.

**Lemma 5.4** *Define $M^{RLS} = D^{RLS}E_R\left[R(R'X'XR)^{-1}R'\right]X'$ where $D^{RLS}$ is a diagonal matrix with diagonal elements $d_j^{RLS} = \left\{E_R[r_j'(R'X'XR)^{-1}R']X'x_j\right\}^{-1}$, and $R$ a $p \times k$ matrix with normally and independently distributed entries. Choose the projection dimension $k$ as*

$$k = \left(1 - c_\kappa\sqrt{(\log p)/n}\right)(n-1), \tag{5.29}$$

*where $c_k$ is a positive constant.*

*Then we have*

$$P\left(||M^{RLS}X - I_p||_{\max} \geq a\sqrt{\frac{\log p}{n}}\right) = O\left(p^{-\tilde{c}}\right), \tag{5.30}$$

*with $\tilde{c}$ as in Lemma 5.1 with $a$ replaced by $\tilde{a} < a$. Furthermore, for $Z = \sqrt{n}d_j^{RLS}E[r_j(R'X'XR)^{-1}R']X'\varepsilon$, we have*

$$\begin{aligned} Z|X &\sim N(0, \sigma^2\Omega^{RLS}), \\ \Omega^{RLS} &= nM^{RLS}M^{RLS'}, \\ \Omega_{jj}^{RLS} &= O_p(1). \end{aligned} \tag{5.31}$$

The proof of Lemma 5.4 given in Appendix 5.B.5 relies on showing that when $k$ is sufficiently close to $n$, the regularized inverse approximates the Moore-Penrose inverse. The results from Section 5.3.2 can then be used to show that regularizing using random least squares does not adversely affect the bias. The proof of Lemma 5.4 also elicits that random least squares is equivalent to a generalized form of ridge regression, where the regularization strength is dependent on the eigenvalues of the regressor matrix $X$. Details on the constant $c_k$ are provided in the proof.

**Ridge regularization** Because of the relation between the Moore-Penrose pseudoinverse and ridge regularized covariance matrices displayed in (5.18), intuition suggests that for a sufficiently small penalty parameter $\lambda$, the results under a Moore-Penrose inverse carry over to a ridge adjusted estimator. The following lemma formalizes this intuition.

**Lemma 5.5** *Define $M^{RID} = D^{RID}(X'X + \gamma I_p)^{-1}X'$, with the elements of the diagonal scaling matrix $D^{RID}$ equal to $d_j^{RID} = \left(e_j'(X'X + \gamma I_p)^{-1}X'x_j\right)^{-1}$. If the ridge penalty parameter satisfies $\gamma \leq c_\gamma p\sqrt{\frac{\log p}{n}}$, where $c_\gamma$ is a positive constant, then we have*

$$P\left(||M^{RID}X - I_p||_{\infty} \geq a\sqrt{\frac{\log p}{n}}\right) = O\left(p^{-\tilde{c}}\right), \tag{5.32}$$

*with $\tilde{a}$ and $\tilde{c}$ as in Lemma 5.4.*

*Furthermore, for* $Z = \sqrt{n} d_j^{RID}(X'X + \gamma I_p)^{-1}X'\varepsilon$*, we have*

$$\begin{aligned} Z|X &\sim N(0, \sigma^2 \Omega^{RID}), \\ \Omega^{RID} &= nM^{RID}M^{RID'}, \\ \Omega_{jj}^{RID} &= O_p(1). \end{aligned} \tag{5.33}$$

A proof is provided in Appendix 5.B.6, which also gives a more detailed description of the constant $c_\gamma$.

**Inference using a regularized approximate inverse** Using Lemma 5.4 and Lemma 5.5, we arrive at the following theorem for the regularized estimators.

**Theorem 5.2** *Suppose A5.1-A5.3 hold. Let* $\hat{\beta}^c = My - (MX - I_p)\hat{\beta}^{init}$*, with* $\hat{\beta}^{init}$ *such that* $||\hat{\beta}^{init} - \beta||_1 = O_p\left(s_0\sqrt{\log(p)/n}\right)$*, and take* $M$ *as either* $M^{RLS} = D^{RLS}E_R\left[R(R'X'XR)^{-1}R'\right]X'$ *or* $M^{RID} = D^{RID}(X'X + \gamma^* I_p)^{-1}X'$*, where the elements of the diagonal matrices* $D$ *are defined in Lemma 5.4 and Lemma 5.5,* $R$ *is a* $p \times k^*$ *matrix with independent standard normal entries,* $k^* = k$ *as in Lemma 5.4, and* $\gamma^* = \gamma$ *as in Lemma 5.5. Then,*

$$\begin{aligned} \sqrt{n}(\hat{\beta}^c - \beta) &= Z + o_p(1), \\ Z|X &\sim N\left(0, \sigma^2\Omega\right), \\ \Omega &= nMM', \\ \Omega_{jj} &= O_p(1). \end{aligned}$$

This theorem follows directly from Lemma 5.4 and Lemma 5.5. It confirms that when $k$ is close to $n$ and $\gamma$ is sufficiently small, the estimator in (5.3) is asymptotically unbiased with covariance matrix $\Omega$, and standard errors that decrease at the usual $n^{-1/2}$ rate.

The reason one would opt for the regularized variants despite the additional tuning parameters is provided by the following theorem. Here we compare the variance of $Z$ in equation (5.4) for the different estimators.

**Theorem 5.3** *Denote the variance of the estimator $\hat{\beta}_j^c$ under a diagonal scaling matrix $D$ by $\Omega_{jj}(D)$. For the choice of $k$ as in Lemma 5.4, or $\gamma$ as in Lemma 5.5, we have*

$$\Omega_{jj}(D)^{RLS} - \Omega_{jj}(D)^{MPI} \leq 0, \qquad \Omega_{jj}(D)^{RID} - \Omega_{jj}(D)^{MPI} \leq 0. \tag{5.34}$$

The proof is given in Appendix 5.B.7.

Note that Theorem 5.3 requires the regularized estimator and the estimator based on the Moore-Penrose pseudoinverse to use the same diagonal scaling matrix. Using $D^{\text{MPI}}$ for the Moore-Penrose inverse, $D^{\text{RLS}}$ for the repeated least squares estimator, and $D^{\text{RID}}$ for the ridge regularized inverse, does not yield an ordering in terms of power. However, in all cases we have encountered, the inequality in Theorem 5.3 is satisfied when using the diagonal matrix specific to the estimator under consideration. This is also evident from the Monte Carlo results in Section 5.4.

### 5.3.4 Consistency

Although our focus in this chapter is on the construction of confidence intervals, the estimator $\hat{\beta}^c$ can be shown to be consistent when we restrict the growth rate of the number of variables relative to the number of observations.

**Assumption 5.4** *The number of variables grows near exponentially with the number of observations, i.e.*

$$\frac{\log p}{n} = o(1). \tag{5.35}$$

Since $Z_i$ is (asymptotically) normal, we have that $\max_{i=1,\ldots,j} |Z_i| = O_p(\sqrt{\log p})$. Since $\hat{\beta}^c = \beta + \frac{1}{\sqrt{n}}(\Delta + Z)$, Assumption A5.4 then guarantees that $\lim_{n\to\infty} \hat{\beta}^c = \beta$.

If one is only interested in consistency, then Assumption A5.2 can potentially be relaxed. In that case the bias is not required to be of lower order compared to the variance.

## 5.4 Monte Carlo Experiments

This section examines the finite sample behaviour of the proposed estimators in a Monte Carlo experiment.

### 5.4.1 Monte Carlo set-up

**Data generating process** The data generating process takes the form

$$y = X\beta + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 I_n), \tag{5.36}$$

where $y$ is an $n \times 1$ vector, $X$ an $n \times p$ regressor matrix, and $\beta$ a $p \times 1$ vector of unknown regressor coefficients. The rows of $X$ are fixed i.i.d. realizations from $\mathcal{N}_p(0, \Sigma)$. We specify two different covariance matrices $\Sigma$:

$$\text{Equicorrelated: } \Sigma_{jk} = 0.8, \quad \forall j \neq k, \quad \Sigma_{jj} = 1 \quad \forall j, \tag{5.37}$$

$$\text{Toeplitz: } \Sigma_{jk} = 0.9^{|j-k|}, \quad \forall j, k. \tag{5.38}$$

The strength of the individual predictors is considered local-to-zero by setting $\beta = \sqrt{\sigma_\varepsilon^2/n} \cdot b\iota_s$ for a fixed constant $b$. The vector $\iota_s$ contains $s$ randomly chosen non-zero elements that are equal to one. We vary signal strength $b$, sparsity $s$, and covariance matrix $\Sigma$ across different Monte Carlo experiments.

We set the number of predictors $p = 200$ and the sample size $n = 100$. In each replication the predictors in $X$ and the coefficients in $\beta$ are generated. We report average results for nonzero coefficients and zero coefficients, based on 1000 replications of the data generating process in (5.36).

**Estimation** We use (5.3) to estimate the coefficients by the Moore-Penrose pseudoinverse, random least squares, and ridge estimator. The lasso estimator uses a penalty term that minimizes the mean squared error under tenfold cross-validation. The random least squares estimator averages over $N = 1000$ realizations of the regularized covariance matrix and projects onto a subspace dimension with $k = 90$. The ridge regression based estimator sets its penalty parameter as $\gamma = 1$, following Bühlmann et al. (2013).

The proposed estimators are compared to three existing methods for constructing confidence intervals in high-dimensional regression for all coefficients. The method of van de Geer et al. (2014) (GBRD) serves as the first benchmark, in which $M$ is constructed by performing lasso for each column in $X$ on the remaining columns in $X$. For each lasso estimation the penalty parameter is selected by tenfold cross-validation. The method of Zhang

and Zhang (2014) is equivalent to this method for linear regression problems considered here. Second, Javanmard and Montanari (2014) (JM) construct $M$ by solving a convex program. We set the tuning parameter $\mu = 2\sqrt{n^{-1}\log p}$, which is equal to the value used in their simulation studies. Both benchmark methods also make use of a bias correction by an initial estimator, for which we again use the lasso estimator. Finally, we compare the performance against the recently developed Correlated Predictors Screening (CPS) method by Lan et al. (2016). In this method, for each regressor $x_j$ we find highly correlated regressors from the set of remaining columns in the regressor matrix. We then orthogonalize both $y$ and $x_j$ with respect to this set. Stopping rules for the size of the correlated set and estimation of the noise level can be found in Lan et al. (2016).

Both for our proposed methods and for JM and GBRD we estimate the noise level $\sigma^2$ using an estimator based on the lasso as defined in (5.19).

**Evaluation** The coverage rate is calculated as the percentage of cases in which the value of the coefficient in the data generating process falls inside the 95% confidence interval. The statistical power is calculated as the percentage of Monte Carlo replications in which zero is not included in the confidence interval of nonzero coefficients.

### 5.4.2 Simulation Results

#### Sparsity and signal strength

Table 5.1 shows the Monte Carlo simulation results for the set of experiments with an equicorrelated covariance matrix and Table 5.2 with a Toeplitz covariance matrix. The tables report the estimated coefficients, standard errors, coverage rates, and power of the Moore-Penrose pseudoinverse, random least squares, and ridge regression. Settings vary over the number ($s = 3, 15$) and signal strength ($b = 2, 5$; corresponding to coefficients of size 0.2 and 0.5) of nonzero coefficients.

The proposed methods obtain a coverage rate close to the nominal rate of 95%. The coverage rates are most precise in case of an equicorrelated covariance matrix in a sparse setting with a weak signal. We observe the largest deviations from the nominal rate for a Toeplitz covariance matrix in a non-sparse setting with a strong signal. In general, the quality of the results seem to be higher when an equicorrelated covariance matrix is used.

**Table 5.1:** Monte Carlo simulation: Equicorrelated Covariance Matrix

| | | $s = 3$ | | | | $s = 15$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| method | $b$ | coef. | SE | CR | power | coef. | SE | CR | power |
| MPI | 2 | 0.19 | 0.30 | 0.95 | 0.10 | 0.17 | 0.29 | 0.94 | 0.10 |
| | 0 | 0.00 | 0.30 | 0.95 | | 0.00 | 0.29 | 0.95 | |
| RLS | 2 | 0.19 | 0.28 | 0.95 | 0.10 | 0.17 | 0.27 | 0.94 | 0.11 |
| | 0 | 0.00 | 0.28 | 0.95 | | 0.00 | 0.27 | 0.95 | |
| RID | 2 | 0.19 | 0.29 | 0.95 | 0.10 | 0.17 | 0.28 | 0.94 | 0.11 |
| | 0 | 0.00 | 0.29 | 0.95 | | 0.00 | 0.28 | 0.95 | |
| GBRD | 2 | 0.17 | 0.20 | 0.94 | 0.13 | 0.16 | 0.20 | 0.93 | 0.14 |
| | 0 | 0.00 | 0.20 | 0.95 | | 0.01 | 0.20 | 0.96 | |
| JM | 2 | 0.06 | 0.05 | 0.15 | 0.14 | 0.09 | 0.05 | 0.21 | 0.27 |
| | 0 | 0.02 | 0.05 | 0.96 | | 0.03 | 0.05 | 0.91 | |
| CPS | 2 | 0.26 | 0.23 | 0.94 | 0.21 | 0.68 | 0.28 | 0.58 | 0.70 |
| | 0 | 0.10 | 0.23 | 0.92 | | 0.52 | 0.28 | 0.55 | |
| MPI | 5 | 0.47 | 0.30 | 0.94 | 0.35 | 0.44 | 0.34 | 0.94 | 0.27 |
| | 0 | 0.00 | 0.30 | 0.95 | | 0.01 | 0.34 | 0.96 | |
| RLS | 5 | 0.46 | 0.28 | 0.93 | 0.40 | 0.43 | 0.31 | 0.93 | 0.30 |
| | 0 | 0.00 | 0.28 | 0.95 | | 0.01 | 0.31 | 0.96 | |
| RID | 5 | 0.46 | 0.29 | 0.93 | 0.38 | 0.44 | 0.33 | 0.94 | 0.28 |
| | 0 | 0.00 | 0.29 | 0.95 | | 0.01 | 0.33 | 0.96 | |
| GBRD | 5 | 0.43 | 0.20 | 0.89 | 0.53 | 0.42 | 0.23 | 0.87 | 0.44 |
| | 0 | 0.01 | 0.20 | 0.96 | | 0.03 | 0.23 | 0.96 | |
| JM | 5 | 0.22 | 0.05 | 0.14 | 0.64 | 0.34 | 0.06 | 0.25 | 0.77 |
| | 0 | 0.02 | 0.05 | 0.95 | | 0.11 | 0.94 | 0.70 | |
| CPS | 5 | 0.67 | 0.24 | 0.89 | 0.79 | 1.70 | 0.47 | 0.27 | 0.94 |
| | 0 | 0.26 | 0.25 | 0.82 | | 1.30 | 0.50 | 0.26 | |

Note: this table reports the average over the estimated coefficients (coef.), standard errors (SE), coverage rates (CR) and statistical power of the Moore-Penrose pseudoinverse (MPI), random least squares (RLS), ridge regression (RID), and the methods of van de Geer et al. (2014) (GBRD), Javanmard and Montanari (2014) (JM) and Lan et al. (2016) (CPS). Results are based on 1000 replications of the linear model (5.36), with equicorrelated regressors as in (5.37). Results are provided separately for non-zero ($b \neq 0$) and zero ($b = 0$) coefficients. The number of observations is $n = 100$ and the number of regressors $p = 200$. The subspace dimension in RLS is $k = 0.9n$, we average over $N = 1000$ low-dimensional projections, and the penalty parameter for ridge regression is $\gamma = 1$. We vary the number ($s = 3, 15$) and signal strength ($b = 2, 5$) of nonzero coefficients.

Both the bias and the standard errors are smaller, and the coverage rate is very close to the nominal rate.

We find that ridge regularization results in an increase in power relative to the Moore-Penrose pseudoinverse estimator, but both estimators are outperformed by random least squares in all considered settings. Even though the number of variables is twice as large as the number of observations, the proposed methods achieve nontrivial power, varying from

**Table 5.2:** Monte Carlo simulation: Toeplitz Covariance Matrix

| | | $s = 3$ | | | | $s = 15$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| method | $b$ | coef. | SE | CR | power | coef. | SE | CR | power |
| MPI | 2 | 0.19 | 0.35 | 0.95 | 0.08 | 0.17 | 0.34 | 0.94 | 0.09 |
| | 0 | 0.00 | 0.35 | 0.95 | | 0.00 | 0.34 | 0.95 | |
| RLS | 2 | 0.19 | 0.30 | 0.95 | 0.09 | 0.17 | 0.29 | 0.94 | 0.10 |
| | 0 | 0.00 | 0.30 | 0.95 | | 0.01 | 0.29 | 0.95 | |
| RID | 2 | 0.19 | 0.32 | 0.95 | 0.09 | 0.17 | 0.31 | 0.94 | 0.10 |
| | 0 | 0.00 | 0.32 | 0.95 | | 0.01 | 0.31 | 0.95 | |
| GBRD | 2 | 0.18 | 0.21 | 0.94 | 0.15 | 0.15 | 0.20 | 0.94 | 0.13 |
| | 0 | 0.01 | 0.20 | 0.95 | | 0.02 | 0.20 | 0.96 | |
| JM | 2 | 0.10 | 0.05 | 0.41 | 0.31 | 0.10 | 0.05 | 0.28 | 0.32 |
| | 0 | 0.01 | 0.05 | 0.95 | | 0.03 | 0.95 | 0.92 | |
| CPS | 2 | 0.19 | 0.31 | 0.95 | 0.10 | 0.19 | 0.44 | 0.95 | 0.08 |
| | 0 | 0.00 | 0.32 | 0.95 | | 0.00 | 0.45 | 0.95 | |
| MPI | 5 | 0.46 | 0.35 | 0.94 | 0.28 | 0.42 | 0.34 | 0.91 | 0.26 |
| | 0 | 0.00 | 0.35 | 0.95 | | 0.01 | 0.66 | 0.95 | |
| RLS | 5 | 0.45 | 0.30 | 0.93 | 0.35 | 0.42 | 0.30 | 0.89 | 0.33 |
| | 0 | 0.00 | 0.30 | 0.95 | | 0.01 | 0.70 | 0.95 | |
| RID | 5 | 0.46 | 0.32 | 0.93 | 0.32 | 0.42 | 0.31 | 0.90 | 0.30 |
| | 0 | 0.00 | 0.32 | 0.95 | | 0.01 | 0.69 | 0.95 | |
| GBRD | 5 | 0.42 | 0.20 | 0.86 | 0.55 | 0.37 | 0.20 | 0.77 | 0.47 |
| | 0 | 0.01 | 0.20 | 0.96 | | 0.02 | 0.80 | 0.96 | |
| JM | 5 | 0.29 | 0.05 | 0.25 | 0.82 | 0.29 | 0.05 | 0.22 | 0.73 |
| | 0 | 0.01 | 0.05 | 0.95 | | 0.03 | 0.95 | 0.89 | |
| CPS | 5 | 0.50 | 0.37 | 0.95 | 0.28 | 0.48 | 0.84 | 0.95 | 0.09 |
| | 0 | 0.00 | 0.41 | 0.95 | | -0.01 | 0.88 | 0.95 | |

Note: this table reports the results for different Monte Carlo experiments where the regressors have a Toeplitz covariance as specified in (5.38). For additional information, see the note following Table 5.1.

0.10 to 0.40. The highest power is achieved in a sparse setting with a strong signal strength. In almost all cases, power is larger in settings with equicorrelated covariance matrix instead of Toeplitz.

We find some downward bias for the nonzero coefficients for the proposed methods in this chapter. The bias decreases in sparsity, which means that nonzero coefficients are more precisely estimated when there are relatively few of them. For all methods, the coefficients which are set to zero in the data generating process are estimated very close to zero.

Random least squares produces the most efficient estimates relative to ridge regression and Moore-Penrose pseudoinverse regression. Standard errors of the random least squares

estimates are lower than these estimators in all experiments. Ridge is again a more efficient estimator relative to the pseudo-inverse, in line with Theorem 5.3. Except for the non-sparse setting with a strong signal, standard errors are larger for a Toeplitz than an equicorrelated covariance matrix.

Compared to the benchmark models, the proposed models are less (downward) biased and obtain coverage rates substantially closer to the nominal rate. In all settings under consideration, the methods proposed in this chapter produce coverage rates that are closer to the nominal rates than the method of van de Geer et al. (2014). This can be explained by the large bias of the GBRD estimator in combination with small standard errors. The JM method produces coefficient estimates and standard errors that are both close to zero, which results in low coverage rates for the nonzero coefficients. Javanmard and Montanari (2014) present better results under the same choice for the tuning parameter. However, their simulation study considers a low-dimensional setting, where the number of variables does not exceed the number of observations. The method developed by Lan et al. (2016) performs well for Toeplitz designs. We see only a minor bias in the coefficient estimates, but substantially larger standard errors compared to the methods proposed in this chapter when the signal strength and/or the number of nonzero coefficients increase. For the equicorrelated design the coverage rates deteriorate and bias increases severely. Clearly this design does not satisfy the necessary conditions underlying the validity of CPS.

**Varying signal strength**

Since many economic processes can be characterized by a small number of large effects and a large number of small effects on the variable of interest, we now consider a setting in which the signal strength varies over the nonzero coefficients in the data generating process. Table 5.3 shows the Monte Carlo simulation results for this set of experiments for an equicorrelated and Toeplitz covariance matrix. The sparsity $s$ equals 15 and we randomly assign $b = 10$ to three nonzero coefficients and $b = 2$ to the 12 remaining nonzero coefficients.

In general, the findings for the proposed methods are similar to the settings discussed in the previous paragraph. The nonzero coefficients are estimated with some downward bias, which is larger in the Toeplitz setting relative to the equicorrelated covariance matrix. Estimates of coefficients that are zero in the data generating process are again estimated

**Table 5.3:** Monte Carlo simulation: Varying signal strength

| | | Equicorrelated | | | | Toeplitz | | | |
|---|---|---|---|---|---|---|---|---|---|
| method | $b$ | coef. | SE | CR | power | coef. | SE | CR | power |
| MPI | 10 | 0.94 | 0.31 | 0.93 | 0.84 | 0.92 | 0.34 | 0.91 | 0.75 |
| | 2 | 0.18 | 0.31 | 0.95 | 0.09 | 0.17 | 0.34 | 0.94 | 0.09 |
| | 0 | 0.00 | 0.31 | 0.95 | | 0.01 | 0.34 | 0.95 | |
| RLS | 10 | 0.93 | 0.28 | 0.93 | 0.89 | 0.91 | 0.29 | 0.89 | 0.85 |
| | 2 | 0.18 | 0.28 | 0.95 | 0.10 | 0.17 | 0.29 | 0.94 | 0.10 |
| | 0 | 0.00 | 0.28 | 0.96 | | 0.01 | 0.29 | 0.95 | |
| RID | 10 | 0.94 | 0.29 | 0.93 | 0.86 | 0.91 | 0.31 | 0.90 | 0.81 |
| | 2 | 0.18 | 0.29 | 0.95 | 0.09 | 0.17 | 0.31 | 0.94 | 0.09 |
| | 0 | 0.00 | 0.29 | 0.96 | | 0.01 | 0.31 | 0.95 | |
| GBRD | 10 | 0.90 | 0.21 | 0.85 | 0.97 | 0.87 | 0.20 | 0.81 | 0.95 |
| | 2 | 0.16 | 0.21 | 0.94 | 0.13 | 0.15 | 0.20 | 0.93 | 0.13 |
| | 0 | 0.02 | 0.21 | 0.96 | | 0.02 | 0.20 | 0.96 | |
| JM | 10 | 0.75 | 0.05 | 0.20 | 0.99 | 0.77 | 0.05 | 0.25 | 0.99 |
| | 2 | 0.11 | 0.05 | 0.24 | 0.34 | 0.10 | 0.05 | 0.26 | 0.31 |
| | 0 | 0.05 | 0.05 | 0.84 | | 0.03 | 0.05 | 0.92 | |
| CPS | 10 | 1.76 | 0.36 | 0.43 | 1.00 | 0.98 | 0.65 | 0.95 | 0.34 |
| | 2 | 1.09 | 0.40 | 0.39 | 0.78 | 0.19 | 0.74 | 0.95 | 0.06 |
| | 0 | 0.93 | 0.41 | 0.37 | | -0.01 | 0.75 | 0.95 | 0.00 |

Note: this table reports the results for Monte Carlo experiments with an equicorrelated and Toeplitz covariance matrix, where the nonzero coefficients of the regressors have different signal strengths. Three randomly chosen coefficients out of the 15 nonzero coefficients have signal strength $b = 10$ and the remaining 12 coefficients $b = 2$. For additional information, see the note following Table 5.1.

very close to zero. Although there is a large variation in signal strength, the standard errors are almost the same for coefficients of different strength and we find the same ranking in efficiency; random least squares produces the smallest standard errors, followed by the ridge regularized estimator.

The coverage rates for the zero coefficients are close to the nominal rate. The coverage rates for coefficients with a weak and moderately strong signal are slightly too low. The decrease in coverage rates holds especially for the Toeplitz setting, where standard errors are relatively larger, but also the bias increases relative to data generated from an equicorrelated covariance matrix.

We find that the power for coefficients with intermediate signal strength ($b = 2$) is comparable to settings with a constant signal strength in Table 5.1 and 5.2. As expected, the power for the strong signals is much larger, varying between 0.75 and 0.86. In general,

power increases for data generated from an equicorrelated covariance matrix relative to a Toeplitz.

Compared to the benchmark estimators, the proposed estimators show also superior performance in the settings with varying signal strength. The distance between the nominal coverage rate and the coverage rate attained by the methods GBRD and JM is in any case larger than for MPI, RLS, and RID. For the Toeplitz design, the coverage rate of CPS is excellent, but the standard errors are almost two times as large as for the competing methods.

**Estimation of the noise level** The validity of confidence intervals depends on a consistent estimator of the noise level $\sigma^2$. Appendix 5.C shows for each setting of the Monte Carlo experiments a box plot of the estimated $\sigma^2$ in each replication. We find that the noise level estimated by scaled lasso can be strongly biased, especially in settings where the data is generated from a Toeplitz covariance matrix, where the lasso estimator results in estimates that are always within one standard deviation from the true value. Therefore, the results in Table 5.1 and 5.2 are based on the estimator for the noise level $\sigma^2$ as defined in (5.19).

## 5.5 Empirical Application

This section applies the proposed estimators to the riboflavin data set that was considered in Bühlmann et al. (2014), and used by van de Geer et al. (2014) and Javanmard and Montanari (2014) to illustrate their methods.

The data, made available by DSM (Switzerland) consists of $n = 71$ observations on (the logarithm of) the riboflavin production rate. There are $p = 4088$ variables available that measure (the logarithm of) the gene expression level. We standardize each predictor by its standard deviation and consider the linear model. The coefficients $\beta$ are estimated in the regression equation

$$y = \alpha + X\beta + \varepsilon \tag{5.39}$$

where $y$ equals the logarithm of the riboflavin production rate, and $X$ the logarithm of the gene expression levels.

When estimating by random least squares, we choose the subspace dimension $k = 64$ and $N = 1000$ realizations of the regularized covariance matrix. The penalty parameter in

**Table 5.4:** Significant effects of genes on Riboflavin

| | MPI | | RLS | | RID | |
|---|---|---|---|---|---|---|
| variable | coef. | SE | coef. | SE | coef. | SE |
| ARGF_at | -0.303 | 0.069 | -0.288 | 0.058 | -0.303 | 0.068 |
| YOAB_at | -0.333 | 0.074 | -0.320 | 0.063 | -0.333 | 0.074 |
| YXLD_at | | | -0.260 | 0.058 | | |

Note: this table reports the estimated coefficients (coef.) and standard errors (SE) which are significantly different from zero on a five percent significance level, estimated by the Moore-Penrose pseudoinverse estimator (MPI), random least squares (RLS), and ridge regularization (RID).

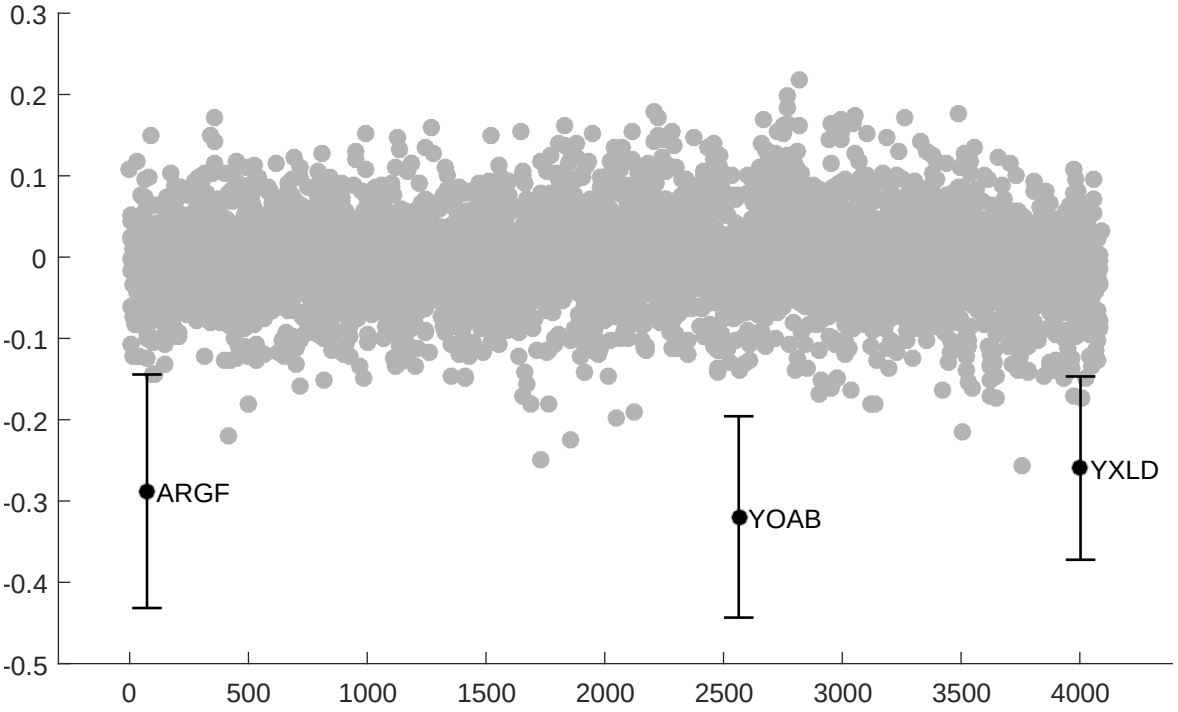
the lasso estimator for the lasso correction corresponds to the lowest mean squared error over a grid of one hundred values, and the penalty parameter in ridge regression is set to $\gamma = 1$ as in Bühlmann et al. (2013).

All previous approaches control the family-wise error rate at 5% by a Bonferonni correction. Bühlmann et al. (2014) find the gene YXLD_at to be significant using a sample splitting approach, while no significant genes are found using the projection estimator proposed in Bühlmann et al. (2013). The second finding is in agreement with the findings by van de Geer et al. (2014). However, Javanmard and Montanari (2014) find two significant variables: YXLD_at and YXLE_at.

Table 5.4 shows the estimated coefficients and standard errors which are significantly different from zero on a five percent significance level after Bonferonni correction. Using the approach based on the Moore-Penrose inverse, we find two significant genes that have not been identified by previous approaches: ARGF_at and YOAB_at. Using random least squares, we find in addition YXLD_at, which is in agreement with Javanmard and Montanari (2014). The ridge based approach yields the same findings as the Moore-Penrose inverse, showing that there is little power loss when one foregoes ridge regularization.

For ARGF_at, YOAB_at, and YXLD_at we find large positive coefficients of respectively 0.288, 0.320, and 0.260 with random least squares. Figure 5.1 shows that the remaining coefficients are close to zero.

**Figure 5.1:** Significant Coefficients Regression Riboflavin




Note: this figure shows the estimated coefficients in the regression of the genes on Riboflavin. Boldfaced coefficients are significantly different from zero on a five percent significance level after Bonferroni correction, and are accompanied by error bands constructed by two times the standard error.

## 5.6 Conclusion

This chapter proposes methods for constructing confidence intervals in high-dimensional linear regression models, where the number of unknown coefficients increases almost exponentially with the number of observations. We approximate the inverse of the singular empirical covariance matrix of the regressors by a diagonally scaled Moore-Penrose pseudoinverse. After a bias correction with the lasso this yields an asymptotically unbiased and normally distributed estimator. The covariance matrix of the estimates is available in closed form and free of tuning parameters. Confidence intervals can then be constructed using standard procedures.

We also consider two regularized estimators; random least squares, which relies on low-dimensional random projections of the data, and ridge regularization. These estimators are shown to have the same theoretical validity under suitable choices of the regularization parameters.

Monte Carlo experiments show that, even in small samples with a high dimensional regressor matrix, the proposed estimators provide valid confidence intervals with correct coverage rates. In an empirical application to riboflavin data, we show that the proposed methods have sufficient power to detect influential genes when only 71 observations are available for 4088 covariates.

# 5.A Preliminary lemmas

## 5.A.1 Concentration bounds

**Lemma 5.6** *Let $z_1^2, \ldots, z_p^2$ be independent subexponential variables with $E[z_i^2] = 1$. Define by $c_s > 0$ a constant such that $\sup_{l \geq 1} l^{-1/2} \left(E[|z_i|^l]\right)^{1/l} \leq c_s$. Then for every $\epsilon \geq 0$,*

$$P\left(\left|\frac{1}{p}\sum_{i=1}^{p} z_i^2 - 1\right| \geq \epsilon\right) \leq 2\exp\left[-cp\min\left(\frac{\epsilon^2}{4c_s^2}, \frac{\epsilon}{2c_s}\right)\right] \tag{5.40}$$

*with $c > 0$ an absolute constant.*

Proof: see Vershynin (2010), Proposition 5.16.

**Lemma 5.7 (Variant Johnson and Lindenstrauss (1984) lemma)** *Let $v$ be a fixed $p \times 1$ vector, and $U_n$ a $p \times n$ matrix that is distributed uniformly over the Stiefel manifold $V_{n,p}$. Then for $c_s$ as in Lemma 5.6 and $0 \leq \epsilon \leq 2c_s$,*

$$P\left(\frac{v'U_nU_n'v}{v'v} \geq (1+\epsilon)\frac{n}{p}, \quad \frac{v'U_nU_n'v}{v'v} \leq (1-\epsilon)\frac{n}{p}\right) \leq 2\exp\left(-\frac{c}{4c_s^2}\epsilon^2 n\right) \tag{5.41}$$

*for $c, c_s > 0$.*

Proof: Since $U_n \in V_{n,p}$, we have that $U_n'U_n = I_n$. Then $v'U_nU_n'v = ||P_{U_n}v||_2^2$, with the orthogonal projection matrix $P_{U_n} = U_n(U_n'U_n)^{-1}U_n'$. As $U_n$ is uniformly distributed on $V_{n,p}$, $P_{U_n}$ is uniformly distributed on the Grassmannian manifold $G_{n,p}$ (Chikuse (2012), theorem 2.2.2).

Instead of taking $P_{U_n}$ random and $v$ fixed, we can take the projection fixed and consider a random $v$. This holds as for any fixed $n \times n$ matrix $P \in G_{n,p}$ and $Q$ uniformly distributed in $\mathcal{O}(p)$, the product $QPQ'$ is uniformly distributed on the Grassmannian $G_{n,p}$ (Chikuse

(2012), theorem 2.2.2). Then, for uniformly random $P_{U_n}$, $v'P'_{U_n}v \overset{(d)}{=} v'QPQv$ where $P$ is fixed.

Since $Q$ is uniformly distributed on $\mathcal{O}(p)$, $Qv \overset{(d)}{=} z$ with $z$ uniformly on the unit sphere $S^{p-1}$. Without loss of generality, assume that the fixed projection matrix $P$ projects $z$ on its first $n$ coordinates. Then

$$\mathrm{E}\left[||Pz||_2^2\right] = \mathrm{E}\left[\sum_{i=1}^{n} z_i^2\right] = \sum_{i=1}^{n} \mathrm{E}\left[z_i^2\right] \tag{5.42}$$

Since $z$ is uniformly distributed $S^{p-1}$, $E[z'z] = \mathrm{E}\left[\sum_{i=1}^{p} z_i^2\right] = pE[z_1^2] = 1$. Then it follows from (5.42) that

$$\mathrm{E}[||Pz||_2^2] = \frac{n}{p} \tag{5.43}$$

To prove Lemma 5.7, we need a concentration result around this expectation. Since $z$ is uniformly distributed on the unit sphere, $z$ is subgaussian. The subvector consisting of the first $m$ coordinates is also subgaussian, as this is simply a linear transformation of $z$. The product of two subgaussian random variables is subexponential (Vershynin, 2010), and hence, we can invoke Lemma 5.6. We have $E[z_i^2] = \frac{1}{p}$, such that

$$P\left(\left|\frac{p}{n}\sum_{i=1}^{n} z_i^2 - 1\right| \geq \epsilon\right) \leq 2\exp\left(-\frac{c}{4c_s^2}\epsilon^2 n\right), \tag{5.44}$$

for $\epsilon, c, c_s > 0$. Note that we assume that $\epsilon^2/(4c_s^2) \leq \epsilon/(2c_s)$, which is satisfied for sufficiently small $\epsilon$. ■

## 5.A.2 Properties of elliptical distributions

Under Assumption A5.2, the concentration results from Appendix 5.A.1 bound the elements of the diagonally scaled Moore-Penrose pseudoinverse. To show this, we first introduce properties of matrices generated from elliptical and spherically symmetric distributions.

For $Z$ an $n \times p$ matrix with rows generated from a spherically symmetric distribution, $Z \overset{(d)}{=} ZT$ for $T \in \mathcal{O}(p)$. The matrix $Z$ can be decomposed by a singular value decomposition as

$$Z = VSU', \tag{5.45}$$

where $V \in \mathcal{O}(n)$, $S$ the $n \times p$ matrix of singular values, and $U \in \mathcal{O}(p)$. Since $Z$ is invariant under right multiplication with an orthogonal matrix, $U$ is uniformly distributed on $\mathcal{O}(p)$. When $n < p$,

$$Z = V S_n U_n', \tag{5.46}$$

where $S_n$ is an $n \times n$ matrix with the non-zero singular values on its diagonal, and $U_n$ is a $p \times n$ matrix that satisfies $U_n' = [I_n, O_{n,p-n}]U'$. Since $U$ is uniformly distributed over $\mathcal{O}(p)$, $U_n$ is uniformly distributed over the Stiefel manifold $V_{n,p}$ defined as $V_{n,p} = \{A \in R^{p \times n} : A'A = I_n\}$.

**Definition 1 (Matrix Angular Central Gaussian distribution, Chikuse (1990))** *Suppose the entries of a $p \times n$ matrix $W$ are independent standard normally distributed, and $\Sigma$ an invertible $p \times p$ matrix. Define $H = \Sigma^{1/2} W (W'\Sigma W)^{-1/2}$. Then $H$ has the density function*

$$f_H = |\Sigma|^{-n/2} |H'\Sigma^{-1}H|^{-p/2}, \tag{5.47}$$

*and is generated from the Matrix Angular Central Gaussian distribution with parameter $\Sigma$, denoted as MACG$(\Sigma)$, and defined on the Stiefel manifold $V_{n,p}$. For $n = 1$, this reduces to the Angular Central Gaussian distribution ACG$(\Sigma)$ on the unit sphere $S^{p-1}$.*

**Lemma 5.8 (Chikuse (2012))** *Define $W$ as a $p \times n$ matrix with independent standard normal entries. For any matrix $U_n$ that is distributed uniformly over $V_{n,p}$, we have that*

$$U_n = W(W'W)^{-1/2}. \tag{5.48}$$

**Lemma 5.9 (Chikuse (2012))** *Let $H$ be a $p \times n$ random matrix on the Stiefel manifold $V_{n,p}$, which is decomposed as*

$$H = [h_1, H_2], \tag{5.49}$$

*where $h_1$ is a $p \times 1$ vector and $H_2$ is a $p \times n - 1$ matrix. Then we can write*

$$h_1 = G(H_2)T, \tag{5.50}$$

*where $G(H_2)$ is any $p \times p-n+1$ matrix chosen so that $[H_2, G(H_2)] \in \mathcal{O}(p)$, and $T$ a $(p-n+1)\times 1$ vector. As $H_2$ takes values in $V_{n-1,p}$, $T$ takes values in $V_{1,p-n+1}$ and the relationship is one-to-one.*

**Lemma 5.10 (Wang and Leng (2015))** *Let $H$ be a $p\times n$ random matrix on the Stiefel manifold $V_{n,p}$. Suppose $H \sim$ MACG$(\Sigma)$. Decompose the Stiefel manifold $H = (G(H_2)T, H_2)$ as in Lemma 5.9, with $T$ a $(p-n+1)\times 1$ and $H_2$ a $p\times(n-1)$ matrix. Then,*

$$T|H_2 \sim ACG(G(H_2)'\Sigma G(H_2)). \tag{5.51}$$

*Since $h_1 = G(H_2)T$, which is a linear transformation of $T$,*

$$h_1|H_2 \sim ACG(\tilde{\Sigma}). \tag{5.52}$$

*where $\tilde{\Sigma} = G(H_2)G(H_2)'\Sigma G(H_2)G(H_2)'$.*

**Lemma 5.11 (Fan and Lv (2008); Wang and Leng (2015))** *Denote the first row of $H$ by $h_1' = [h_{11}, h']$. We then have*

$$e_1 HH'e_2 \overset{(d)}{=} h_{11}h_{21}\Big|\left\{e_1 HHe_1 = h_{11}^2\right\}. \tag{5.53}$$

Proof: For $Q \in \mathcal{O}(n)$

$$e_1'HH'e_2 = e_1'HQQ'H'e_2. \tag{5.54}$$

Now define $\tilde{Q} \in \mathcal{O}(n-1)$ and $Q = \begin{pmatrix} 1 & 0_{1\times n-1} \\ 0_{n-1\times 1} & \tilde{Q} \end{pmatrix}$. Choose $Q$ such that it rotates $H$ into a frame where $e_1'\tilde{H} = \left[\tilde{h}_{11}, 0_{1\times n-1}\right]$. In terms of the rotated frame, we have

$$e_1'HH'e_2 = e_1'\tilde{H}\tilde{H}e_2 = \tilde{h}_{11}\tilde{h}_{21}, \tag{5.55}$$

implying that

$$e_1'HH'e_2 \overset{(d)}{=} h_{11}h_{21}\Big|\left\{e_1'H = h_{11}\right\}. \tag{5.56}$$

Denote the first row of $H$ by $h_1' = [h_{11}, h']$. Then $e_1'HH'e_1 = h_{11}^2 + h'h$ and thus $e_1'H = [h_{11}, 0_{1\times n-1}]$ if and only if $e_1'HH'e_1 = h_{11}^2$. Substituting this into (5.56) completes the proof. ■

# 5.B Proofs

## 5.B.1 Proof of Lemma 5.1

Under Assumption A5.2 and the decomposition (5.46) in Appendix 5.A.2,

$$X'(XX')^{-1}X = \Sigma_2^{1/2}U_n(U_n'\Sigma_2 U_n)^{-1}U_n'\Sigma_2^{1/2}. \tag{5.57}$$

By Lemma 5.8 in Appendix 5.A, we can write $U_n = W(W'W)^{-1/2}$ with the elements of $W$ standard normal and independently distributed. Substituting into (5.57) gives

$$X'(XX')^{-1}X = \Sigma_2^{1/2}W(W'\Sigma_2 W)^{-1}W'\Sigma_2^{1/2} = HH', \tag{5.58}$$

where $H = \Sigma_2^{1/2}W(W'\Sigma_2 W)^{-1/2}$.

We separately bound the diagonal and off-diagonal elements of $HH'$. The proof extends the approach by Wang and Leng (2015).

**Diagonal terms of $HH'$** The diagonal elements of $HH'$ are themselves not of particular interest, as we choose the diagonal matrix $D$ such that the diagonal elements of $MX$ are all equal to one. However, to bound the off-diagonal elements, we require a bound on the diagonal elements of $HH'$. We first construct bounds under the assumption that $\Sigma = I_p$, and then connect these to the case where $\Sigma_2 \neq I_p$.

When $\Sigma_2 = I_p$, we can invoke Lemma 5.7 in Appendix 5.A to show that

$$P\left(e_1'U_nU_n'e_1 > c_\epsilon\frac{n}{p}, \quad e_1'U_nU_n'e_1 < \frac{1}{c_\epsilon}\frac{n}{p}\right) \leq 2\exp\left(-\frac{c}{4c_s^2}\epsilon^2 n\right), \tag{5.59}$$

with $c, c_s > 0$, and $c_\epsilon = \frac{1+\epsilon}{1-\epsilon} > 1$ is introduced to reduce notation.

We will now use these results to establish a bound when $\Sigma_2 \neq I$. The diagonal terms can be bounded by noting that for any vector $v$,

$$\begin{aligned} v'HH'v &= v'\Sigma_2^{\frac{1}{2}}U_n(U_n'\Sigma_2 U_n)^{-1}U_n'\Sigma_2^{\frac{1}{2}}v \\ &\leq \kappa v'U_nU_n'v, \end{aligned} \tag{5.60}$$

where the condition number $\kappa = \frac{\lambda_{\max}(\Sigma_2)}{\lambda_{\min}(\Sigma_2)} < \infty$ by Assumption A5.3. Similarly

$$v'HH'v \geq \frac{1}{\kappa} v'U_n U_n' v. \tag{5.61}$$

Since $U_n \overset{(d)}{=} QU_n$ with $Q \in \mathcal{O}(p)$, upon choosing $Q$ such that $Qv = e_1$, we obtain

$$P\left(e_1'HH'e_1 > c_\epsilon \kappa \frac{n}{p}, \quad e_1'HH'e_1 < \frac{1}{c_\epsilon \kappa}\frac{n}{p}\right) \leq 2\exp\left(-\frac{c}{4c_s^2}\epsilon^2 n\right). \tag{5.62}$$

**Off-diagonal elements** The proof for the off-diagonal elements is more involved. For $i = 1$ and $j = 2$, we bound with high probability the ratio $\frac{|e_i'HH'e_j|}{e_i'HH'e_i}$. A union bound is used to extend the results to arbitrary $i$ and $j$.

We separate three cases: (a) $e_1'HHe_1 \geq c_\epsilon \kappa \frac{n}{p}$, (b) $c_\epsilon \kappa \frac{n}{p} > e_1'HH'e_1 > \frac{1}{c_\epsilon \kappa}\frac{n}{p}$, and (c) $e_1'HHe_1 \leq c_\epsilon \kappa \frac{n}{p}$. Conditioning on these three cases and using the trivial fact that for any probability $P(\cdot) \leq 1$, it follows that

$$\begin{aligned}
P\left(\frac{|e_1'HH'e_2|}{e_1'HH'e_1} \geq t\right) &\leq P\left(e_1'HH'e_1 \geq c_\epsilon \kappa \frac{n}{p}\right) + P\left(e_1'HH'e_1 \leq \frac{1}{c_\epsilon \kappa}\frac{n}{p}\right) \\
&\quad + \int_{\frac{1}{c_\epsilon \kappa}\frac{n}{p}}^{c_\epsilon \kappa \frac{n}{p}} P\left(\frac{|e_1'HH'e_2|}{e_1'HH'e_1} \geq t \,\middle|\, e_1'HH'e_1 = t_1^2\right) P\left(e_1'HH'e_1 = t_1^2\right) dt_1^2 \\
&\leq P\left(e_1'HH'e_1 \geq c_\epsilon \kappa \frac{n}{p}\right) + P\left(e_1'HH'e_1 \leq \frac{1}{c_\epsilon \kappa}\frac{n}{p}\right) \\
&\quad + P\left(\frac{|e_1'HH'e_2|}{e_1'HH'e_1} \geq t \,\middle|\, e_1'HH'e_1 = t_*^2\right).
\end{aligned} \tag{5.63}$$

where $t_*$ is the value of $t_1$ that maximizes $P\left(\frac{|e_1'HH'e_2|}{e_1'HH'e_1} \geq t \,\middle|\, e_1'HH'e_1 = t_1^2\right)$.

The first two terms of (5.63) are bounded by (5.62), so we focus on the final term of (5.63). Denote the $i,j$-th element of $H$ by $h_{ij}$. Lemma 5.11 in Appendix 5.A states that

$$e_1'HH'e_2 \overset{(d)}{=} h_{11}h_{21} \mid \left\{h_{11}^2 = e_1'HH'e_1\right\}, \tag{5.64}$$

from which it follows that

$$e_1'HH'e_2 \mid \left\{e_1'HH'e_1 = t_1^2\right\} \overset{(d)}{=} h_{11}h_{21} \mid \left\{h_{11}^2 = t_1^2\right\}. \tag{5.65}$$

We decompose $H = [h_1, H_2]$, with $h_1$ a $p \times 1$ vector, and $H_2$ a $p \times n-1$ matrix. As in Lemma 5.10, $h_1 = G(H_2)T$ with $G(H_2)$ such that $[H_2, G(H_2)] \in \mathcal{O}(p)$. Then by Lemma 5.10 in Appendix 5.A, $h_1|H_2 \overset{(d)}{=} \frac{y}{\sqrt{y_1^2+\ldots+y_p^2}}$, where $y = (y_1, \ldots, y_p) \sim N(0, \tilde{\Sigma})$ with $\tilde{\Sigma} = G(H_2)G(H_2)'\Sigma G(H_2)G(H_2)'$.

Using the above results, $h_{11}h_{21}|\{h_{11}^2 = t_1^2\} \overset{(d)}{=} \frac{y_1 y_2}{y_1^2+\ldots+y_p^2}$. Since $\frac{y_1^2}{y_1^2+\ldots+y_p^2} = t_1^2$, we have $y_1^2 = \frac{t_1^2}{1-t_1^2}\left(y_2^2 + \ldots + y_p^2\right)$. Then

$$\frac{|y_1 y_2|}{y_1^2 + \ldots + y_p^2} = \frac{(1-t_1^2)|y_1 y_2|}{y_2^2 + \ldots + y_p^2} \leq \frac{\sqrt{1-t_1^2}|t_1||y_2|}{\sqrt{y_2^2 + \ldots + y_p^2}}. \tag{5.66}$$

Now we establish the following upper bound

$$\begin{aligned} P\left(\frac{|e_1'HH'e_2|}{e_1'HH'e_1} \geq t \middle| h_{11}^2 = t_1^2\right) &= P\left(\frac{|h_{11}h_{21}|}{h_{11}^2} \geq t \middle| h_{11}^2 = t_1^2\right) \\ &\leq P\left(\frac{\sqrt{1-t_1^2}|y_2|}{\sqrt{y_2^2 + \ldots + y_p^2}} \geq |t_1|t\right) \\ &= P\left(\frac{|y_2|}{\sqrt{y_2^2 + \ldots + y_p^2}} \geq t\sqrt{\frac{t_1^2}{1-t_1^2}}\right) \\ &\leq P\left(\frac{|y_2|}{\sqrt{y_2^2 + \ldots + y_p^2}} \geq t\sqrt{\frac{1}{c_\epsilon \kappa}\frac{n}{p}}\right). \end{aligned} \tag{5.67}$$

where we use that $t_1^2/(1-t_1^2)$ is a monotonically increasing function in $t_1^2$, and the minimum value of $t_1^2$ that we need to consider equals $\frac{1}{c_\epsilon \kappa}\frac{n}{p}$. This is then our choice for $t_*$ in (5.63).

Since by definition, $G(H_2)'G(H_2) = I_{p-n+1}$, $\lambda_{\max}(\tilde{\Sigma}) \leq \lambda_{\max}(\Sigma)$. Similarly, we have $\lambda_{\min}(\tilde{\Sigma}) \geq \lambda_{\min}(\Sigma)$. Then by Lemma 5.6 in Appendix 5.A,

$$\begin{aligned} P\left(|y_2| \geq \sqrt{\lambda_{\max}(\Sigma)}\sqrt{1+\epsilon_1}\right) &\leq 2e^{-\frac{c}{2c_s}\epsilon_1} \\ P\left(\sqrt{y_2^2 + \ldots + y_p^2} \leq \sqrt{\lambda_{\min}(\Sigma)(p-n)(1+\epsilon_2)}\right) &\leq 2e^{-\frac{c}{4c_s^2}\epsilon_2^2(p-n)}, \end{aligned} \tag{5.68}$$

where we assumed that $\epsilon_1$ is such that $\epsilon_1/(2c_s) < \epsilon_1^2/(4c_s^2)$, which will be justified below, and $\epsilon_2$ such that $\epsilon_2^2/(4c_s^2) \leq \epsilon_2/(2c_s)$.

Using Bonferonni's inequality, (5.68) implies

$$P\left(\frac{|y_2|}{\sqrt{y_2^2+\ldots+y_p^2}} \geq \sqrt{\kappa\frac{1+\epsilon_1}{1+\epsilon_2}\frac{1}{p-n}}\right) \leq 2e^{-\frac{c}{2c_s}\epsilon_1} + 2e^{-\frac{c}{4c_s^2}\epsilon_2^2(p-n)}. \tag{5.69}$$

Take $c_p$ a constant such that $p/n \geq c_p > 1$, then also

$$P\left(\frac{|y_2|}{\sqrt{y_2^2+\ldots+y_p^2}} \geq \sqrt{\frac{\kappa}{(1-c_p^{-1})}\frac{1+\epsilon_1}{1+\epsilon_2}\frac{1}{p}}\right) \leq 2e^{-\frac{c}{2c_s}\epsilon_1} + 2e^{-\frac{c}{4c_s^2}\epsilon_2^2(p-n)}. \tag{5.70}$$

We are interested in the case where $\sqrt{\frac{\kappa}{(1-c_p^{-1})}\frac{1+\epsilon_1}{1+\epsilon_2}\frac{1}{p}} = t\sqrt{\frac{1}{c_\epsilon\kappa}\frac{n}{p}}$, which holds for

$$t = \kappa\sqrt{\frac{c_p c_\epsilon}{c_p-1}\frac{1+\epsilon_1}{1+\epsilon_2}\frac{1}{n}}. \tag{5.71}$$

Since $\kappa^2 c_\epsilon c_p/(c_p-1) > 1$, we can take $\epsilon_2 = \kappa^2 c_\epsilon c_p/(c_p-1) - 1$. Then choosing $\epsilon_1 = a^2\log p - 1$, we have

$$P\left(\frac{|e_1' HH' e_2|}{e_1' HH' e_1} > a\sqrt{\frac{\log p}{n}}\right) \leq 2e^{-\frac{c}{2c_s}a^2\log p} + 2e^{-\frac{c}{4c_s^2}[\kappa^2 c_\epsilon c_p/(c_p-1)-1]^2(p-n)}. \tag{5.72}$$

Note that for this choice of $\epsilon_1$, for $p$ sufficiently large $\epsilon_1/(2c_s) < \epsilon_1^2/(4c_s^2)$, which was used in (5.68).

Finally, taking the union bound over all pairs $e_i, e_j$ we have that

$$P\left(\frac{|e_i' HH' e_j|}{e_i' HH' e_i} > a\sqrt{\frac{\log p}{n}}\right) = O\left(p^{-\tilde{c}}\right) \qquad \forall i,j \in \{1,\ldots,p\} \tag{5.73}$$

with $\tilde{c} = \frac{c}{2c_s}a^2 - 2$. ■

## 5.B.2 Proof of Lemma 5.2

The bound in Lemma 5.2 is shown by Bühlmann and Van De Geer (2011) to hold under the following compatibility condition

**Definition 2 (Compatibility condition)** *Denote by $S_0$ the true set of $s_0 = ||S_0||_0$ non-zero coefficients, then the compatibility condition is satisfied for this set if*

$$||\beta_{S_0}||_1 \leq \frac{\sqrt{s_0}||X\beta||_2}{\sqrt{n}\phi_0}, \tag{5.74}$$

*for all $\beta$ for which $||\beta_{S_0^c}||_1 \leq 3||\beta_{S_0}||_1$ and $\phi_0 > 0$.*

This condition is satisfied under Assumption A5.2. Note that $||\beta_{S_0}||_1 \leq \sqrt{s_0}||\beta_{S_0}||_2$, so it is sufficient if

$$||\beta||_2^2 \leq \frac{\beta' \frac{1}{n} X'X\beta}{\phi_0}. \tag{5.75}$$

Using Assumption A5.2, we have

$$\begin{aligned} \beta' \frac{1}{n} X'X\beta &= \beta' \Sigma^{1/2} \frac{1}{n} US'SU' \Sigma^{1/2} \beta \\ &\geq \frac{1}{c_Z} \frac{p}{n} v' U_n U_n' v, \end{aligned} \tag{5.76}$$

where $v = \Sigma^{1/2}\beta$, and the last line holds since the non-zero eigenvalues $S'S$ are the same as the eigenvalues of $ZZ'$ which are bounded by Assumption A5.2. Since our results should hold for all $s_0$-sparse vectors $\beta$, we apply a union bound in combination with Lemma 5.7. This shows that with probability at least $1 - 2\exp(-\frac{c}{4c_s^2}\epsilon^2 n + s_0 \log p)$, we have that

$$\begin{aligned} \beta' \frac{1}{n} X'X\beta &\geq \frac{1}{c_Z c_\epsilon} \beta' \Sigma \beta \\ &\geq \frac{1}{c_Z c_\epsilon} \lambda_{\min}(\Sigma) ||\beta||_2^2. \end{aligned} \tag{5.77}$$

Choosing $\phi_0 \leq \frac{1}{c_Z c_\epsilon} \lambda_{\min}(\Sigma)$, and if $s_0 \log p = o(n)$, we have the desired result. ■

## 5.B.3 Proof of Lemma 5.3

Building on Wang and Leng (2015), we rewrite the noise term of $\hat{\beta}_i^c$ as

$$Z_i = \sqrt{n} d_i x_i' (XX')^{-1} \varepsilon \overset{(d)}{=} \sqrt{n} d_i ||x_i'(XX')^{-1}||_2 \frac{\sigma x_i'(XX')^{-1} u}{||x_i'(XX')^{-1}||_2}, \tag{5.78}$$

where $u \sim N(0, I_n)$.

We first bound the norm term

$$\sqrt{n}d_i||x_i'(XX')^{-1}||_2 = \sqrt{n}\frac{||x_i'(XX')^{-1}||_2}{x_i'(XX')^{-1}x_i}. \tag{5.79}$$

Using standard norm inequalities, we have

$$\frac{1}{\lambda_{\max}(XX')}x_i'(XX')^{-1}x_i \leq ||x_i'(XX')^{-1}||_2^2 \leq \frac{1}{\lambda_{\min}(XX')}x_i'(XX')^{-1}x_i. \tag{5.80}$$

The eigenvalues of $XX' = \Sigma_1^{1/2}Z\Sigma_2 Z'\Sigma_1^{1/2}$ satisfy

$$\begin{aligned}\lambda_{\max}(\Sigma_1^{1/2}Z\Sigma_2 Z'\Sigma_1^{1/2}) &\leq \lambda_{\max}(\Sigma_1)\lambda_{\max}(\Sigma_2)\lambda_{\max}(ZZ'),\\ \lambda_{\min}(\Sigma_1^{1/2}Z\Sigma_2 Z'\Sigma_1^{1/2}) &\geq \lambda_{\min}(\Sigma_1)\lambda_{\min}(\Sigma_2)\lambda_{\min}(ZZ').\end{aligned} \tag{5.81}$$

The eigenvalues of $ZZ'$ are bounded by Assumption A5.2, and using (5.62), it follows that with probability exceeding $1 - 2e^{-c\epsilon^2 n} - 2e^{-C_Z n}$ we have that

$$\begin{aligned}\left(\frac{1}{\lambda_{\max}(\Sigma_1)\lambda_{\max}(\Sigma_2)}\frac{n}{p}\frac{1}{c_\epsilon\kappa\frac{n}{p}}\right)^{1/2} &\leq \sqrt{n}d_i||x_i'(XX')^{-1}||_2\\ &\leq \left(\frac{1}{\lambda_{\min}(\Sigma_1)\lambda_{\min}(\Sigma_2)}\frac{n}{p}\frac{1}{\frac{1}{c_\epsilon\kappa}\frac{n}{p}}\right)^{1/2},\end{aligned} \tag{5.82}$$

By Assumption A5.3, the eigenvalues of $\Sigma_1$ and $\Sigma_2$ are finite. Then

$$\sqrt{n}d_i||x_i'(XX')^{-1}||_2 = O_p(1). \tag{5.83}$$

We now turn to the second term of (5.78)

$$\frac{\sigma x_i'(XX')^{-1}u}{||x_i'(XX')^{-1}||_2} = \frac{\sigma\frac{1}{\sqrt{n}}x_i'\left(\frac{1}{p}XX'\right)^{-1}u}{\left|\left|\frac{1}{\sqrt{n}}x_i'\left(\frac{1}{p}XX'\right)^{-1}\right|\right|_2}. \tag{5.84}$$

When $u \sim N(0, I_n)$, it is clear that

$$\frac{1}{\sqrt{n}}X'\left(\frac{1}{p}XX'\right)^{-1}u \sim N\left[0, \frac{1}{n}X'\left(\frac{1}{p}XX'\right)^{-2}X\right]. \tag{5.85}$$

and hence $Z_i \sim N(0, \sigma^2\Omega_{ii})$ with $\Omega_{ii} = n\frac{x_i'(XX')^{-2}x_i}{(x_i'(XX')^{-1}x_i)^2} = O_p(1)$. ■

## 5.B.4 Proof of Lemma 5.3 for non-gaussian errors

**Lemma 5.12** *Suppose assumptions A5.2 and A5.3 hold. The errors $\varepsilon_i$ are independent and identically distributed with variance $\sigma^2$, and satisfy*

$$E\left[|\varepsilon_i|^{2+\delta}\right] \leq c < \infty \tag{5.86}$$

*for $i = 1, \ldots, n$. Then as $n \to \infty$,*

$$\frac{1}{\sqrt{n}} e_i' M\varepsilon \xrightarrow{(d)} N(0, \sigma^2 e_i' MM' e_i/n). \tag{5.87}$$

Proof: When $u_i \sim i.i.d(0,1)$, we will show that Lyupanov's condition is satisfied, and therefore a central limit theorem applies ensuring that, as $n \to \infty$,

$$\frac{\sigma x_i' \left(XX'\right)^{-1} u}{\left|\left|x_i' \left(XX'\right)^{-1}\right|\right|_2} \xrightarrow{(d)} N(0, \sigma^2). \tag{5.88}$$

Define

$$r_{ik} = \frac{\left[\left(XX'\right)^{-1} x_i\right]_k}{||\left(XX'\right)^{-1} x_i||_2} \tag{5.89}$$

where the numerator denotes the $k$-th component of the $n$-dimensional vector $\left(XX'\right)^{-1} x_i$. Furthermore, we have $\mathrm{E}[r_{ik}u_k] = 0$, $\mathrm{Var}[r_{ik}u_k] = \left(||\left(XX'\right)^{-1} x_i||_2^{-1} \left[\left(XX'\right)^{-1} x_i\right]_k\right)^2$, $s_n^2 = \sum_{k=1}^n \mathrm{Var}[r_{ik}u_k] = 1$. To prove that a central limit theorem applies to $\sum_{k=1}^n r_{ik}u_k$ we prove that Lyapunov's condition,

$$LC = \lim_{n\to\infty} \sum_{k=1}^{n} |r_{ik}u_k|^{2+\delta} = 0, \tag{5.90}$$

holds. By assumption we have

$$LC \leq c \lim_{n\to\infty} \sum_{k=1}^{n} |r_{ik}|^{2+\delta}. \tag{5.91}$$

By Assumption 5.2 the summand satisfies with probability exceeding $1 - \exp(-C_Z n)$

$$|r_{ik}| \leq c_Z^2 \frac{||x_i||_\infty}{||x_i||_2}. \tag{5.92}$$

By the results in Appendix 5.B.2, we have that, again with high probability, $||x_i||_2 \geq \frac{\lambda_{\min}(\Sigma_1)\lambda_{\min}(\Sigma_2)}{c_Z} c_\epsilon^{-1} n$. We can then continue our string of inequalities as

$$|r_{ik}| \leq c_Z^3 c_{\kappa,1} c_{\kappa_2} c_\epsilon \frac{||z_i||_\infty}{n}, \tag{5.93}$$

where $z_i$ denotes the $i$-th row of the matrix $Z$ defined in Assumption 5.2.

Since by assumption each element of $Z$ is independent and identically distributed with variance 1, following Chebyshev's inequality

$$P(|z_{ik}| \geq a) \leq a^{-2}. \tag{5.94}$$

Then applying a union bound over $k \in \{1, \ldots, n\}$ gives

$$P(||z_i||_\infty \geq a) \leq n a^{-2}. \tag{5.95}$$

Choosing $a = c_a n^{1/2(1+\alpha)}$, the right-hand side tends to zero, and uniformly over $k$,

$$|z_{ik}| \leq c_Z^3 c_{\kappa,1} c_{\kappa_2} c_\epsilon n^{-1/2(1-\alpha)}. \tag{5.96}$$

In this case

$$LC \leq c_Z^3 c_{\kappa,1} c_{\kappa_2} c_\epsilon n^{\alpha - \delta/2 + \alpha\delta/2}, \tag{5.97}$$

which tends to zero as $n \to \infty$ if

$$\alpha - \delta/2 + \alpha\delta/2 < 0 \Rightarrow \alpha \leq \frac{\delta}{2+\delta}. \tag{5.98}$$

This shows that for an individual parameter $\beta_i$,

$$\sum_{k=1}^{n} r_{ik} u_k \xrightarrow{d} N(0,1) \tag{5.99}$$

which completes the proof. The extension to a fixed subset of $\beta$ follows from a union bound of the size of the subset. ■

We can extend the results of Lemma 5.12 to hold uniformly over $i \in \{1, \ldots, p\}$, by making the additional assumption that the rows of $Z$ are subgaussian and the number of variables does not increase too fast with the number of observations.

**Lemma 5.13** *Suppose assumptions A5.2 and A5.3 hold, but strenghten Assumption A5.2 such that the rows of $Z$ are also subgaussian. As in Lemma 5.12, suppose that the errors $\varepsilon_i$ are independent and identically distributed with variance $\sigma^2$, and satisfy $E\left[|\varepsilon_i|^{2+\delta}\right] \leq c < \infty$ for $i = 1, \ldots, n$. In addition, the number of regressors grows at a rate*

$$\log p = o\left(n^{1-\frac{1}{2+\delta}}\right). \tag{5.100}$$

*Then, as $n \to \infty$*

$$\frac{1}{\sqrt{n}} e_i' M \varepsilon \xrightarrow{(d)} N(0, \sigma^2 e_i' M M' e_i / n). \tag{5.101}$$

*and this result holds uniformly over $i \in \{1, \ldots, p\}$.*

Proof: In this case, instead of Chebyshev's inequality (5.94) we use

$$P(|z_{ik}| \geq a) \leq 2 \exp\left(-a^2/2\right). \tag{5.102}$$

Applying again a union bound over all $k \in \{1, \ldots, n\}$ and $i \in \{1, \ldots, p\}$ gives uniformly over $i$

$$P(||Z||_{\max} \geq a) \leq 2 \exp\left(-a^2/2 + \log p + \log n\right) \tag{5.103}$$

The right-hand side now goes to zero if $a > \sqrt{2(\log p + \log n)}$. In this case, we have

$$LC \leq c \lim_{n \to \infty} \left( \frac{\log p + \log n}{n^{1-\frac{1}{2+\delta}}} \right)^{2+\delta} \tag{5.104}$$

Ignoring the lower order term $\log n$, we see that $LC \to 0$ uniformly over $i \in \{1, \ldots, p\}$, when $n \to \infty$ and $\log p = o\left(n^{1-\frac{1}{2+\delta}}\right)$. This completes the proof. ■

### 5.B.5 Proof of Lemma 5.4: random least squares

**Size of the bias** Consider the eigenvalue decomposition

$$\frac{1}{n}X'X = \hat{U}_n\hat{\Lambda}\hat{U}_n'. \tag{5.105}$$

where $\hat{U}_n$ is a $p \times n$ matrix, and $\hat{\Lambda}$ an $n \times n$ diagonal matrix of eigenvalues. We list three properties of the expectation $\mathrm{E}[R(R'X'XR)^{-1}R']X'X$ established in Marzetta et al. (2011).

First, using the eigenvalue decomposition (5.105) and the fact that only $n$ eigenvalues are non-zero,

$$\mathrm{E}[R(R'X'XR)^{-1}R']X'X \overset{(d)}{=} \hat{U}_n\mathrm{E}[\Phi(\Phi'\hat{\Lambda}\Phi)^{-1}\Phi']\hat{\Lambda}\hat{U}_n', \tag{5.106}$$

with $\Phi$ an $n \times k$ matrix of independent standard normal random variables. The proof relies on the fact that for any orthogonal matrix $\hat{U}$ independent of $R$, we have that $\hat{U}'R \overset{(d)}{=} \Phi$.

Second, $\mathrm{E}[\Phi(\Phi'\hat{\Lambda}\Phi)^{-1}\Phi']\hat{\Lambda}$ is a diagonal matrix. This follows since a matrix $A$ is diagonal if and only if for all diagonal unitary matrices $\Omega$, we have that $\Omega A\Omega^* = A$ with $\Omega^*$ the complex conjugate of $\Omega$. Indeed,

$$\begin{aligned}\Omega\mathrm{E}[\Phi(\Phi'\hat{\Lambda}\Phi)^{-1}\Phi']\hat{\Lambda}\Omega^* &= \Omega\mathrm{E}[\Phi(\Phi'\hat{\Lambda}\Phi)^{-1}\Phi']\Omega^*\hat{\Lambda} \\ &= \Omega\mathrm{E}[\Phi(\Phi'\Omega^*\Omega\hat{\Lambda}\Omega^*\Omega\Phi)^{-1}\Phi']\Omega^*\hat{\Lambda} \\ &\overset{(d)}{=} \mathrm{E}[\Psi(\Psi'\hat{\Lambda}\Psi)^{-1}\Psi']\hat{\Lambda},\end{aligned} \tag{5.107}$$

where $\Psi$ is again an $n \times k$ matrix of standard normals, and using as above that $\Omega\Phi \overset{(d)}{=} \Psi$ for any unitary matrix $\Omega$.

The final property is that we can rewrite

$$\mathrm{E}[\Psi(\Psi'\hat{\Lambda}\Psi)^{-1}\Psi']\hat{\Lambda} = I - V, \tag{5.108}$$

where

$$V = \mathrm{E}[\Xi(\Xi'\hat{\Lambda}^{-1}\Xi)^{-1}\Xi']\hat{\Lambda}^{-1} \tag{5.109}$$

is an $n \times n$ diagonal matrix with $\Xi$ is a $n \times (n-k)$ matrix with independent standard normal entries.

Using (5.108), it follows that

$$\mathrm{E}_R[R(R'X'XR)^{-1}R']X'X = \hat{U}(I-V)\hat{U}'. \tag{5.110}$$

Now, $\hat{U}\hat{U}'$ is the Moore-Penrose pseudoinverse post-multiplied by $X$, which is identical to (5.58) in Appendix 5.B.1, so that we have

$$\hat{U}\hat{U}' = X'(XX')^{-1}X = HH'. \tag{5.111}$$

Therefore, one expects that if the entries of $\hat{U}V\hat{U}'$ are sufficiently small compared to $\hat{U}\hat{U}'$, then the results obtained under the Moore-Penrose inverse will continue to hold.

Denote by $\hat{u}_i = \hat{U}'e_i$. We can use the following string of inequalities

$$\begin{aligned} P\left(\frac{|\hat{u}_i'(I-V)\hat{u}_j|}{\hat{u}_i'(I-V)\hat{u}_i} \geq t\right) &\leq P\left(\frac{|\hat{u}_i'(I-V)\hat{u}_j|}{\hat{u}_i'\hat{u}_i(1-||V||_2)} \geq t\right) \\ &\leq P\left(\frac{|\hat{u}_i'\hat{u}_j|}{\hat{u}_i'\hat{u}_i} + \frac{|\hat{u}_i'V\hat{u}_j|}{\hat{u}_i'\hat{u}_i} \geq t(1-||V||_2)\right) \\ &\leq P\left(\frac{|\hat{u}_i'\hat{u}_j|}{\hat{u}_i'\hat{u}_i} + ||V||_2\sqrt{\frac{\hat{u}_j'\hat{u}_j}{\hat{u}_i'\hat{u}_i}} \geq t(1-||V||_2)\right). \end{aligned} \tag{5.112}$$

For $\hat{u}_i'\hat{u}_i = e_i'HH'e_i$ and $\hat{u}_j'\hat{u}_j = e_j'HH'e_j$, we can apply the bounds established in (5.62) in Appendix 5.B.1. Denote by $\mathcal{E}$ the event that $e_j'HH'e_j \leq c_\epsilon\kappa\frac{n}{p}$, $e_j'HH'e_j \geq (c_\epsilon\kappa)^{-1}\frac{n}{p}$, then the string of inequalities (5.112) proceeds as

$$\begin{aligned} &\leq P\left(\frac{|e_i'HH'e_j|}{e_i'HH'e_i} + ||V||_2 c_\epsilon\kappa \geq t(1-||V||_2)\middle|\,\mathcal{E}\right)\left(1-2e^{-\frac{c}{4c_s^2}\epsilon^2 n}\right) + 2e^{-\frac{c}{4c_s^2}\epsilon^2 n} \\ &= P\left(\frac{|e_i'HH'e_j|}{e_i'HH'e_i} \geq t - ||V||_2\left(t + c_\epsilon\kappa\right)\right)\left(1-2e^{-\frac{c}{4c_s^2}\epsilon^2 n}\right) + 2e^{-\frac{c}{4c_s^2}\epsilon^2 n}. \end{aligned} \tag{5.113}$$

We now need to find a choice of the projection dimension $k$ such that $t(1-||V||_2) - ||V||_2 c_\epsilon\kappa = \tilde{a}\sqrt{\log p/n}$. This will then allow us to apply the previously derived bounds on $|e_i'HH'e_j|/e_i'HH'e_i$.

We first analyze the $l_2$ norm $||V||_2$ is more detail. Denote by $\hat{\lambda}_i$ the $i$-th diagonal element of the diagonal matrix of empirical eigenvalues $\hat{\Lambda}$, $\xi_i$ the $i$-th row of $\Xi$ defined in (5.109),

and $A_{-i} \equiv \sum_{j \neq i} \hat{\lambda}_j^{-1} \xi_j \xi_j'$. It holds that

$$\begin{aligned} [V]_{ii} &= \hat{\lambda}_i^{-1} \xi_i' (\Xi' \hat{\Lambda}^{-1} \Xi)^{-1} \xi_i \\ &= \hat{\lambda}_i^{-1} \xi_i' \left( A_{-i} + \hat{\lambda}_i^{-1} \xi_i \xi_i' \right)^{-1} \xi_i \\ &= \frac{\hat{\lambda}_i^{-1} \nu_i}{1 + \hat{\lambda}_i^{-1} \nu_i}, \end{aligned} \tag{5.114}$$

where

$$\nu_i = \xi_i' A_{-i}^{-1} \xi_i = \xi_i' \left( \Xi_{-i} \hat{\Lambda}_{-i}^{-1} \Xi_{-i} \right)^{-1} \xi_i > 0, \tag{5.115}$$

and the Sherman-Morrison formula is used to obtain the last line of (5.114). This shows that random least squares performs a type of generalized ridge regression, where the penalty is different for each eigenvalue. By Jensen's inequality and the fact that $x/(1+x)$ with $x > 0$ is a concave function,

$$[V]_{ii} \leq \frac{\hat{\lambda}_i^{-1} \mathrm{E}[\nu_i]}{1 + \hat{\lambda}_i^{-1} \mathrm{E}[\nu_i]} \leq \frac{\hat{\kappa} \frac{n-k-1}{k}}{1 + \hat{\kappa} \frac{n-k-1}{k}}, \tag{5.116}$$

where $\hat{\kappa} = \max_i \hat{\lambda}_i / \min_i \hat{\lambda}_i \leq c_\kappa$ by (5.81). We can now solve for which $k$ we have that $t(1 - ||V||_2) - ||V||_2 c_\epsilon \kappa = \tilde{a}\sqrt{\log p/n}$. In order for the bias of the estimator to vanish compared to the noise, we require $t = a\sqrt{\log p/n}$. After some rewriting, we then find that $k$ should satisfy

$$k = \left( 1 + \frac{a - \tilde{a}}{\tilde{a} c_\kappa} \frac{\tilde{a}(c_\epsilon \kappa)^{-1} \sqrt{\log p/n}}{1 + \tilde{a}(c_\epsilon \kappa)^{-1} \sqrt{\log p/n}} \right)^{-1} (n-1). \tag{5.117}$$

Assuming $\tilde{a}(c_\epsilon \kappa)^{-1} \sqrt{\log p/n}$ to be sufficiently small, we have

$$k = \left( 1 - c_k \sqrt{\frac{\log p}{n}} \right) (n-1), \tag{5.118}$$

with $c_k = (a - \tilde{a})/(\kappa c_\epsilon c_\kappa)$ a positive constant. Under this choice of $k$, the approximate inverse obtained by random least squares satisfies

$$\begin{aligned} P\left( \frac{|\hat{u}_i'(I-V)\hat{u}_j|}{\hat{u}_i'(I-V)\hat{u}_i} \geq a\sqrt{\frac{\log p}{n}} \right) &\leq P\left( \frac{|e_i' H H' e_j|}{e_i' H H' e_i} \geq \tilde{a}\sqrt{\frac{\log p}{n}} \right) \\ &= O\left( p^{-\tilde{c}} \right) \end{aligned} \tag{5.119}$$

with $\tilde{c}$ as in Lemma 5.1 with $a$ replaced by $\tilde{a} < a$.

**Order of the variance term** What remains to be shown is that the variance of the noise term satisfies

$$||\sqrt{n}d_i^{\text{RLS}}e_i\text{E}\left[R(R'X'XR)^{-1}R'\right]X'||_2 = O_p(1). \tag{5.120}$$

We rewrite this as

$$\left(\sqrt{n}d_i^{\text{RLS}}||e_i\text{E}\left[R(R'X'XR)^{-1}R'\right]X'||_2\right)^2 = n\frac{e_i'\hat{U}_n(I-V)\hat{\Lambda}^{-1}(I-V)\hat{U}_n'e_i}{(e_i'\hat{U}_n(I-V)\hat{U}_n'e_i)^2}, \tag{5.121}$$

which can be lower and upper bounded as

$$\begin{aligned}\frac{1}{\lambda_{\max}(\hat{\Lambda})}n\frac{e_i'\hat{U}_n(I-V)^2\hat{U}_n'e_i}{(e_i'\hat{U}_n(I-V)\hat{U}_n'e_i)^2} &\leq n\frac{e_i'\hat{U}_n(I-V)\hat{\Lambda}^{-1}(I-V)\hat{U}_n'e_i}{(e_i'\hat{U}_n(I-V)\hat{U}_n'e_i)^2} \\ &\leq \frac{1}{\lambda_{\min}(\hat{\Lambda})}n\frac{e_i'\hat{U}_n(I-V)^2\hat{U}_n'e_i}{(e_i'\hat{U}_n(I-V)\hat{U}_n'e_i)^2}.\end{aligned} \tag{5.122}$$

Under stated assumptions, the eigenvalues satisfy $c_1 p \leq \lambda_{\min}(\hat{\Lambda}) \leq \lambda_{\max}(\hat{\Lambda}) \leq c_2 p$ for $0 \leq c_1 \leq c_2$. Also, from the previous paragraph we know that the elements of $V$ satisfy $0 \leq [V]_{ii} \leq c_V\sqrt{\frac{\log p}{n}} \leq 1$ for some $c_V > 0$. Then,

$$\begin{aligned}\frac{1}{c_2}\left[\frac{n}{p}\left(1-c_V\sqrt{\frac{\log p}{n}}\right)^2\frac{1}{e_i'\hat{U}_n\hat{U}_n'e_i}\right] &\leq n\frac{e_i'\hat{U}_n(I-V)\hat{\Lambda}^{-1}(I-V)\hat{U}_n'e_i}{(e_i'\hat{U}_n(I-V)\hat{U}_n'e_i)^2} \\ &\leq \frac{1}{c_1}\left[\frac{n}{p}\left(1-c_V\sqrt{\frac{\log p}{n}}\right)^{-2}\frac{1}{e_i'\hat{U}_n\hat{U}_n'e_i}\right].\end{aligned} \tag{5.123}$$

Finally, (5.59) in Appendix 5.A and the fact that $e_i'\hat{U}_n\hat{U}_n'e_i = e_iHH'e_i$ shows that

$$n\frac{e_i'\hat{U}_n(I-V)\hat{\Lambda}^{-1}(I-V)\hat{U}_n'e_i}{(e_i'\hat{U}_n(I-V)\hat{U}_n'e_i)^2} = O_p(1). \tag{5.124}$$

This completes the proof. ■

### 5.B.6 Proof of Lemma 5.5: ridge regression

**Order of bias term** The proof largely follows the strategy under random least squares. We first show that $(X'X + \gamma I_p)^{-1}X'X$ also satisfies the right-hand side of (5.110) in Appendix 5.B.5.

By substituting $X = \hat{V}\hat{S}\hat{U}$ and defining $\hat{\Lambda} = \hat{S}'\hat{S}$, we have

$$\begin{aligned}(X'X + \gamma I_p)^{-1}X'X &= (\hat{U}\hat{\Lambda}\hat{U}' + \gamma I_p)^{-1}\hat{U}\hat{\Lambda}\hat{U}' \\ &= \hat{U}_n(I_n - V)\hat{U}_n',\end{aligned} \tag{5.125}$$

where $\hat{\Lambda}_n$ is a diagonal matrix with on the diagonal the nonzero eigenvalues of $X'X$, $U_n$ consists of the first $n$ rows of $\hat{U}$ and $V = (\hat{\Lambda}_n + \gamma I_n)^{-1}\gamma I_n$.

Now following (5.113), $V$ should be such that $t(1 - ||V||_2) - ||V||_2 c_\epsilon \kappa = \tilde{a}\sqrt{\log p/n}$ for $t = a\sqrt{\log p/n}$. This implies $||V||_2 = \frac{(a-\tilde{a})\sqrt{\log p/n}}{a\sqrt{\log p/n} + c_\epsilon \kappa}$. Since $V$ is diagonal, and the non-zero eigenvalues satisfy $c_1 p \leq \lambda_{\min}(\hat{\Lambda}) \leq \lambda_{\max}(\hat{\Lambda}) \leq c_2 p$ for $0 \leq c_1 \leq c_2$,

$$||V||_2 = \max_{i=1,\ldots,n} \frac{\gamma}{\hat{\lambda}_i + \gamma} \leq \frac{\gamma}{c_1 p + \gamma}. \tag{5.126}$$

It follows that we need to set

$$\gamma \leq c_1 p \frac{||V||_2}{1 - ||V||_2}, \tag{5.127}$$

Using the expression for $||V||_2$ and assuming $\tilde{a}\sqrt{\log p/n}/(c_\epsilon \kappa)$ sufficiently small, we have

$$\gamma = c_\gamma \sqrt{\frac{\log p}{n}} p, \tag{5.128}$$

with $c_\gamma = c_1(a - \tilde{a})/(c_\epsilon \kappa)$. ■

**Order of the variance** What remains to be shown is

$$||\sqrt{n} d_i^{\text{RI}} e_i'(X'X + \gamma I_p)^{-1}X'||_2 = O_p(1). \tag{5.129}$$

This follows from the same argument as made for random least squares.

### 5.B.7 Proof of Theorem 5.3

Define the diagonal matrix $A = \mathrm{E}[R(R'\hat{\Lambda}R)^{-1}R']\hat{\Lambda}$, then

$$\begin{aligned} ||e_i\hat{U}\mathrm{E}[R(R'\hat{\Lambda}R)^{-1}R'X'||_2^2 &= e_i\hat{U}A\hat{\Lambda}^{-1}A\hat{U}'e_i \\ &= e_i\hat{U}\hat{\Lambda}^{-1/2}A^2_{\text{RLS}}\hat{\Lambda}^{-1/2}\hat{U}'e_i, \end{aligned} \tag{5.130}$$

where $A^2_{\text{RLS}}$ is a diagonal matrix with diagonal elements $0 \leq A^2_{ii} \leq 1$.

Similarly, for the ridge regularized inverse, we have

$$\begin{aligned} ||e_i(X'X + \gamma I_p)^{-1}X'||_2^2 &= e_i(X'X + \gamma I_p)^{-1}X'X(X'X + \gamma I_p)^{-1}e_i \\ &= e_i\hat{U}_n(\hat{\Lambda} + \gamma I_p)^{-2}\hat{\Lambda}\hat{U}_n'e_i \\ &= e_i\hat{U}_n\hat{\Lambda}^{-1/2}A^2_{\text{RID}}\hat{\Lambda}^{-1/2}\hat{U}_n'e_i, \end{aligned} \tag{5.131}$$

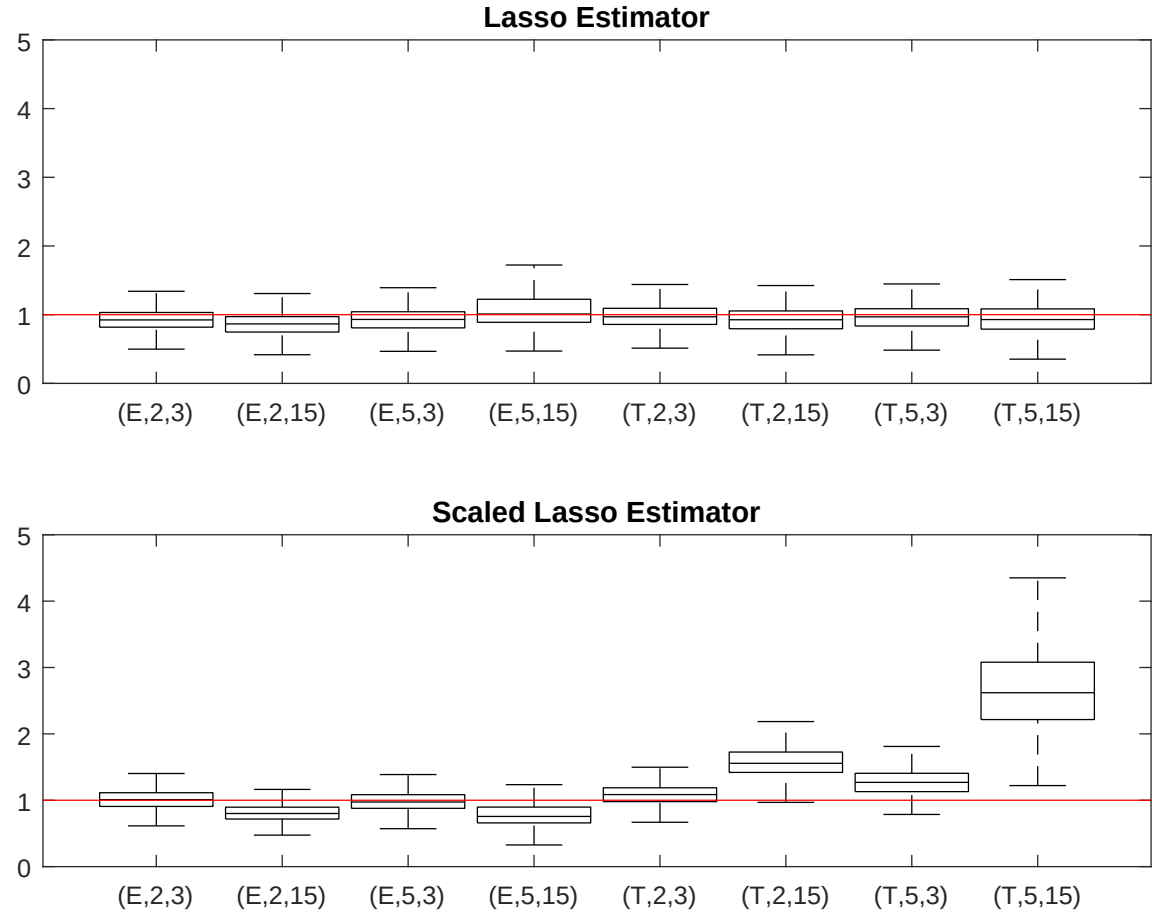
with $A^2_{\text{RID}}$ is a diagonal matrix with the diagonal elements satisfying $0 \leq A^2_{ii} \leq 1$. For the Moore-Penrose pseudoinverse we have

$$\begin{aligned} ||e_iX'(XX')^{-1}||_2^2 &= e_iX'(XX')^{-2}Xe_i \\ &= e_i\hat{U}\hat{\Lambda}^{-1}\hat{U}'e_i. \end{aligned} \tag{5.132}$$

Since for both RLS and RID $A^2$ is a diagonal matrix with diagonal elements satisfying $0 \leq A^2_{ii} \leq 1$, the claim in Theorem 5.3 follows. ■

# 5.C Estimation of the noise level

**Figure 5.2:** Estimates noise level Monte Carlo experiments




Note: this figure shows for each Monte Carlo experiment a box plot for the estimates of the noise level $\sigma^2$ in each replication. The first panel shows these plots for the estimator based on lasso, as defined in (5.19), and the second panel for the estimator based on scaled lasso as in Sun and Zhang (2012). The red horizontal line indicates the value of $\sigma^2 = 1$ in the data generating process. Settings are indicated by (covmat,$b$,$s$), where the covariance matrix covmat varies between equicorrelated (E) and Toeplitz (T), the signal strength ($b = 2, 5$) and sparsity ($s = 3, 15$). For additional information, see the note following Table 5.1.

# Chapter 6

# A High-Dimensional Multinomial Choice Model

## 6.1 Introduction

Many multinomial choice problems involve large choice sets relative to the number of observations. For instance, video streaming providers have choice data on millions of movies, the assortment size in grocery stores exceeds thousands of products, and we have the choice between almost every place in the world as holiday destination.

To identify factors that explain observed choices, choice behavior is related to characteristics of decision makers. Streaming providers take past ratings for different kind of genres of their subscribers into account before recommending movies. With the advent of loyalty cards, grocery stores know exactly in which neighborhoods their costumers live. Online travel agencies ask for your household composition before offering travel deals. These individual characteristics divide the sample of decision makers into a large number of different categories.

Multinomial choice models help to understand the relation between discrete choices and the characteristics of the decision makers. Since the parameters in these discrete choice models are alternative-specific, the number of parameters increases linearly with the size of the choice set. Furthermore, when the explanatory variables describe categorical characteristics, these variables enter the model as sets of dummies, with for each category a dummy variable. With several categorical variables and large numbers of categories, the number of parame-

ters needed to specify the effect on one choice alternative is already large. When both the number of choice alternatives and the number of explanatory categories is large, the number of parameters easily approaches the number of observations.

This chapter proposes a Bayesian method to manage the number of parameters in high-dimensional multinomial choice models in a data-driven way. A two-way Dirichlet process prior on the model parameters of a multinomial probit model encourages the alternative-specific parameters to cluster over both outcome and explanatory categories. With positive probability, the two-way mixture choice model reduces the high-dimensional parameter space in both directions. The result is a decrease in parameter uncertainty and an enhancement of the parameter interpretability, without imposing any model restrictions.

Although pooling of categories is ubiquitous in practice, we are, to the best of our knowledge, the first to estimate from the data which categories can be pooled together, for both dependent and independent categorical variables. We set up a Gibbs sampler that draws model parameters clustered over outcome and explanatory categories, and draws two-way cluster assignments over both dimensions. Since we can formulate the multinomial probit model in terms of latent normally distributed utilities, the cluster assignments can be sampled according to the sample steps developed for mixtures of normals of Ishwaran and James (2002). By jointly estimating the number of clusters, cluster assignments, and model parameters, the posterior parameter distributions incorporate the parameter uncertainty together with the uncertainty in the number of clusters, which is ignored by fixing a priori the number of clusters. The estimated model parameters retain their interpretation as in a standard multinomial choice model, and prior distributions can be parametrized according to prior beliefs about the number of distinct effects over outcome and explanatory categories.

The practical implications of this two-way mixture choice model are illustrated in a simulation study. We show that when parameters are clustered over outcome categories or explanatory categories in the data generating process, estimating the parameters in a standard multinomial choice model on a moderate sample size can lead to biased and noisy estimates. The two-way mixture choice model accurately identifies the clusters of unique parameter values, both over the choice alternatives and the explanatory dummy categories. The posterior parameter distributions improve in accuracy and precision upon the standard multinomial choice model on a range of posterior parameter diagnostics, and show the best in-sample and out-of-sample predictive performance.

In an empirical application we estimate the effect of household composition on holiday destinations. We apply the two-way mixture choice model to survey data of a market research company on holiday behavior of Dutch households. Using 4000 observed holiday choices, we estimate the effect of 21 explanatory variables describing household characteristics, of which 10 control variables and 11 household composition dummy variables, on the choice out of 49 holiday destinations. The Dirichlet process prior reduces the number of estimated parameters from 1029 to 144 in the two-way mixture choice model. The estimated holiday destination clustering is very different from an ad hoc grouping based on, for instance, geographical location. The cluster size of holiday destinations varies from 1 to 28 and only in Europe we already find nine different clusters. The mixture over explanatory dummy paramaters distinguishes different holiday preferences for single households from households of two or more persons. Based on their estimated base preferences, singles are less inclined to visit conventional holiday destinations close to home, but more adventurous to explore countries further away.

When confronted by a large number of alternatives, researchers commonly focus on a subset of alternatives, or alternatives are a priori aggregated to a higher level (Zanutto and Bradlow, 2006; Carson and Louviere, 2014). This is not a solution when all available categories are of interest. Cramer and Ridder (1991) propose a statistical test for pooling outcome categories. However, testing for all different combinations of subsets is computationally expensive and the order of tests can change the final clustering. At the cost of departing from the standard discrete choice model parameter interpretation, Ho and Chong (2003) and Jacobs et al. (2016) circumvent the pooling problem by introducing an additional set of latent variables. Instead of estimating separate parameters for each choice alternative, the explanatory variables influence the choice probabilities via a relatively small set of latent variables.

Large sets of explanatory categories are, similar to choice alternatives, often clustered on expert opinion to ease the curse of dimensionality. Evidently, this leads to suboptimal results when the expert is wrong. More recently, regularization techniques for high-dimensional regressor matrices, such as the lasso introduced by Tibshirani (1996), are also applied to categorical data. In addition to shrinkage or selection, it is for a categorical explanatory variable also of interest which categories should be distinguished when modelling the effect on the outcome variable (Tutz and Gertheiss, 2016). Bondell and Reich (2009) and Gertheiss et al. (2010) show that by choosing a specific functional form for the penalty in the lasso,

categories are clustered to a smaller set of dummies. Although these methods are tailored to the categorical nature of the data, the relation between the lasso penalty parameter and the number of distinguished categories is opaque.

Since choice probabilities and expected utilities of choice alternatives within the same parameter cluster will be identical, we interpret the clusters over outcome categories as preference sets. Preference sets only reduce the parameter space when individuals are assumed to assign in expectation the same utility to different choice alternatives. An alternative assumption is that individuals only have a utility function for a subset of the alternatives in the total choice set, and the alternatives outside this consideration set get a zero probability of being chosen, see for example Liu and Arora (2011) and Manzini and Mariotti (2014). Andrews and Srinivasan (1995), Chiang et al. (1998), and Mehta et al. (2003) model the probabilities of all potential consideration sets. As the number of possible consideration sets is exponential in the number of choice alternatives, estimation becomes computationally infeasible in large choice sets. Bronnenberg and Vanhonacker (1996) and Van Nierop et al. (2010) model the consideration set inclusion probabilities for all alternatives, which does not alleviate the curse of dimensionality in a standard multinomial choice model. Gilbride and Allenby (2004) and Terui et al. (2011) impose a hard constraint on choice alternatives to be included in the consideration set. Just as for the preference sets, the number of parameters increase quadratical in the number of choice alternatives.

The potential of the Dirichlet process prior has gained an increasing amount of attention in different fields in econometrics. The most popular application of the prior is the modelling of unknown error distributions, without resorting to strong parametric assumptions. Hirano (2002) puts a Dirichlet process prior on the error distribution in dynamic panel data models, Van Hasselt (2011) in sample selection models, and Conley et al. (2008) and Wiesenfarth et al. (2014) in instrumental variable models. On the other hand, Dirichlet process priors are used to model parameter heterogeneity. Hu et al. (2015) specify the prior on the model parameters in an instrumental variable model to allow for heterogeneity in treatment effects. Bauwens et al. (2017) use the prior to model time-variation in the parameters of autoregressive moving average models. Burda et al. (2008) models unobserved heterogeneity across individuals by a Dirichlet process prior on individual-specific parameters in a choice model. We employ the properties of a Dirichlet process prior to embattle choice models for high-

dimensional choice sets. Instead of mixing over individuals or over time, the prior clusters parameters over choice alternatives and explanatory categories.

The outline of the remainder of this chapter is as follows. Section 6.2 discusses the general specification of the multinomial probit model and introduces the mixture models. Section 6.3 explains the Bayesian inference methods. Section 6.4 explains the properties of the mixture methods using simulated data and compares the performance to a standard multinomial choice model. Section 6.5 applies the two-way mixture model to survey data on holiday destinations. We conclude with a discussion in Section 6.6.

## 6.2 Model specification

This section discusses the specification of a high-dimensional multinomial choice model. Section 6.2.1 introduces the baseline specification of a multinomial probit model. Section 6.2.2 shows how we cluster parameters over over the categories in the categorical dependent and independent variables in this model. Section 6.2.3 introduces the technique that drives the clustering, the Dirichlet process prior.

### 6.2.1 Multinomial probit model

Let $y_i$ be an observable unordered random categorical variable, such that $y_i \in \{1, 2, \ldots, J\}$, with $J$ the number of choice alternatives, and $i = 1, \ldots, N$, with $N$ the number of individuals. Let $x_i$ be a $K$-dimensional vector with explanatory variables, potentially with dummy coded categorical variables. As is common for multinomial choice models, we introduce latent utilities driving the decisions. Let $z_i = (z_{i1}, \ldots, z_{iJ})'$ be a $J \times 1$ vector of continuous latent random variables, such that

$$y_i(z_i) = j \text{ if } z_{ij} = \max(z_i), \tag{6.1}$$

where $\max(z_i)$ is the largest element of the vector $z_i$. The latent utilities are modeled as

$$z_i = \beta x_i + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \Sigma), \tag{6.2}$$

where $\beta = (\beta_1, \ldots, \beta_J)'$ is a $J \times K$ matrix of coefficients, and $\varepsilon_i$ an independent normally distributed disturbance vector with covariance matrix $\Sigma$. We now have defined the conditional density $f(y_i|x_i, \beta, \Sigma)$, where the covariates in $x_i$ are constant across different outcome categories, but the $K$-dimensional model parameter vectors $\beta_j$, $j = 1, \ldots, J$, vary over the outcome categories.

The parameters $\beta_j$ and $\Sigma$ in the multinomial probit model specified in (6.1) and (6.2) are not identified (Bunch, 1991). There are two parameter identification problems. First, $y_i(z_i + c) = y_i(z_i)$ for each scalar $c$. To overcome this additive redundancy we set $\beta_1 = 0$. Second, we still have $y_i(cz_i) = y_i(z_i)$ for each positive scalar $c$, even if the aforementioned restriction is imposed. We follow Gilbride and Allenby (2004) and Terui et al. (2011) and set the covariance matrix $\Sigma$ in (6.2) to be the identity matrix. This restriction identifies the model parameters and avoids covariance parameter estimation problems when $J$ is large.

A conventional multiplicative identifying assumption is to only restrict the first element of the covariance matrix to be equal to one (McCulloch et al., 2000). Burgette and Nordheim (2012) restrict the trace of the covariance matrix to sample identified parameters. Instead of restricting the covariance matrix, McCulloch and Rossi (1994) report the posterior of the model parameters up to a scaling factor. Imai and Van Dyk (2005) introduce a new parameter to link identified to unidentified parameters.

Although these approaches lead to models with formally identified parameters, Keane (1992) shows that, in the absence of alternative-specific explanatory variables, parameter identification in multinomial probit models is extremely fragile because "it is difficult to disentangle covariance parameters from regressor coefficients". Many economic applications suffer from this problem, which is the reason that the multinomial probit model with a diagonal covariance matrix is most commonly used in applied research (Rossi et al., 2005). However, in practice, even in a diagonal covariance matrix different values for the variances can hardly be identified, especially in the high-dimensional settings we study in this chapter.

### 6.2.2 Parameter clustering over categories

When the choice set is large, the number of parameters in the $J \times K$ matrix $\beta$ easily approaches the number of observations. Large numbers of parameters amplify overfitting concerns, increase parameter uncertainty, and make it a difficult exercise to extract useful

insights. For the data to be informative on the parameters without additional restrictions, the number of outcome categories and the number of explanatory variables need to be relatively small.

Two features of many large scale empirical applications of choice models exacerbate the curse of dimensionality. First, the observed choices $y_1, \ldots, y_N$ often are not evenly distributed over the choice set. This results in a small number of observed choices to estimate the parameters $\beta_j$ for the least chosen alternatives $j$, even for large $N$ relative to $J$. Second, the individual choice behavior is usually explained by, among other variables, categorical variables indicating characteristics of individuals. These categorical variables are implemented by means of dummies, resulting in sets of binary variables for each explanatory category. Therefore, the number of explanatory variables $K$ can become large in models with categorical variables consisting of many explanatory categories.

When subsets of categories can be treated as a single category, this parsimonious model is preferred. Section 6.2.2 discusses parameter clustering over outcome categories, Section 6.2.2 over explanatory categories, and Section 6.2.2 over both dimensions.

**Parameter clustering over outcome categories**

The latent utility model in (6.2) can be written as

$$z_{ij} = \beta_j' x_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \mathcal{N}(0,1), \quad j = 1, \ldots, J, \tag{6.3}$$

where the vector $\beta_j$ contains alternative-specific parameters. Equivalently, we can say that the parameters in $\beta_j$ vary over an infinite number of clusters, where the number of clusters equals the total number of choice alternatives $J$ when each choice alternative has a different parameter vector. Within the clusters the parameters are assumed to be identical, but across clusters the parameters are allowed to be different.

The cluster representation of the latent utility model in (6.3) is

$$z_{ij} = \tilde{\beta}_{C_j}' x_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \mathcal{N}(0,1), \quad j = 1, \ldots, J, \tag{6.4}$$

where $\beta_j = \tilde{\beta}_{C_j}$ can vary over $L_J \to \infty$ clusters. The classification variables $C_j \in \{1, \ldots, L_J\}$ take integer values indicating the cluster for choice category $j$, and identify the corresponding cluster parameter vector $\tilde{\beta}_{C_j}$.

The model in (6.4) imposes the parameter clustering over choice alternatives to be the same for each individual. Although this seems restrictive at first sight, the clustering does not imposes the same expected utility ordering for each individual. Since the utilities are conditional on individual-specific characteristics in $x_i$, the expected preference ordering over outcome categories is also individual-specific. However, when repeated observations per individual are available, we can easily extend (6.4) to a more flexible model with individual-specific clustering.

**Parameter clustering over explanatory categories**

To cluster over categories within a categorical explanatory variable, we make an explicit distinction in the regressor vector $x_i = (w_i', d_i')'$. The $K_d$ dummies in $d_i$ correspond to the categories in the categorical explanatory variable, and $w_i$ contains the $K_w$ remaining explanatory variables without an intercept. We rewrite the model in (6.3) to

$$z_{ij} = \beta_j' x_i + \varepsilon_{ij} = \gamma_j' w_i + \kappa_j' d_i + \varepsilon_{ij}, \tag{6.5}$$

where $\beta_j = (\gamma_j', \kappa_j')'$ and where the parameter values in $\kappa_j = (\kappa_{j,1}, \ldots, \kappa_{j,K_d})$ correspond to the dummy categories in $d_i = (d_{i,1}, \ldots, d_{i,K_d})$. We cluster the dummy parameters over the categories of only one categorical explanatory variable in (6.5). However, the methods can easily be extended to account for parameter clustering over multiple explanatory categorical variables.

We let the explanatory dummy parameters vary over an infinite number of clusters. The formulation of the latent utility model in (6.5) conditional on the classification variables is

$$z_{ij} = \gamma_j' w_i + \sum_{k=1}^{K_d} \tilde{\kappa}_{j,D_k} d_{ik} + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \mathcal{N}(0,1), \quad j = 1, \ldots, J, \tag{6.6}$$

where $\kappa_{jk} = \tilde{\kappa}_{j,D_k}$ can vary over $L_D \to \infty$ clusters. The classification variables $D_k \in \{1, \ldots, L_D\}$ take integer values indicating the cluster for explanatory category $k$. Within a cluster $l$, dummies have identical parameter values $\tilde{\kappa}_{jl}$ and are equivalently aggregated to

a new dummy variable. As a result, the explanatory categories within one cluster have the same effect on the dependent variable and we have a smaller set of dummies.

The dummy parameter clustering in (6.6) perfectly fits categorical variables without a natural ordering, such as profession. However, the modelling framework does not take ordering in the explanatory categories into account. Ordered explanatory categories, for instance income categories, fit in as well but could be handled more efficiently when the ranking in the categories can be taken into account.

### Two-way parameter clustering

Combining parameter clustering over outcome categories in (6.4) with parameter clustering over explanatory categories in (6.6) results in two-way parameter clustering,

$$z_{ij} = \tilde{\gamma}'_{C_j} w_i + \sum_{k=1}^{K_d} \tilde{\kappa}_{C_j,D_k} d_{ik} + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \mathcal{N}(0,1), \quad j = 1, \ldots, J, \tag{6.7}$$

where $\gamma_j = \tilde{\gamma}_{C_j}$ can vary over $L_J$ clusters and $\kappa_{jk} = \tilde{\kappa}_{C_j,D_k}$ over $L_J \times L_D$ clusters.

The parameter clustering in (6.7) over the outcome category dimension is unconditional on the clustering over the explanatory category dimension. This means that the clustering over dummy parameters is the same for each outcome category cluster. Allowing for conditional clustering, where each cluster of outcome categories may have another division of explanatory categories, results in an overly flexible model specification which causes difficulties in parameter estimation and parameter interpretation.

Researchers often pool categories a priori when they consider the number of parameters to be estimated large relative to the number of observations. The frequentist framework provides a variety of statistical tests for testing whether categories share the same parameter value. For instance, Cramer and Ridder (1991) propose a likelihood based test for pooling outcome categories. They test for the equality of two alternative-specific parameter vectors, apart from the intercepts. However, to test for all different combinations of subsets is computationally expensive and the order of tests can change the final clustering. Therefore, researchers arbitrarily aggregate outcome and explanatory categories into subsets in practice (Zanutto and Bradlow, 2006; Carson and Louviere, 2014).

### 6.2.3 Dirichlet process mixture model

The key to our parameter clustering approach is the specification of a cluster assignment probability distribution for each category. The probability distribution is modelled by a Dirichlet process mixture model that implicitly integrates out the cluster probabilities, while allowing for as many clusters as categories. Since there is a positive probability that two categories share a cluster, the Dirichlet process mixture encourages a parsimonious model without imposing any model restrictions.

#### Dirichlet process prior

A data-driven parameter clustering approach is obtained by specifying a Dirichlet process prior for the parameter vector $\beta_j$ in (6.3),

$$\begin{aligned} \beta_j|P &\sim P, \\ P|\alpha_J, H &\sim DP(\alpha_J, H_\beta), \end{aligned} \tag{6.8}$$

where the prior of $\beta_j$ is a random distribution $P$ generated by a Dirichlet process. Conditionally on $P$, the parameter vectors $\beta_j$, $j = 1, \ldots, J$, are independently and identically distributed. The Dirichlet process $DP(\alpha_J, H)$, has a positive scalar concentration parameter $\alpha_J$ and continuous base distribution $H_\beta$.

The expectation over the Dirichlet process equals the base distribution, and the concentration parameter governs the dispersion around the base distribution. When $\alpha_J$ is large, the distributions $P$ and $H$ are more similar. Since $P$ is a discrete random distribution, there is a positive probability that different $\beta_j'$s take the exact same value. A cluster of outcome categories is defined as the choice alternatives with identical parameter vectors $\beta_j$. Therefore, the model in (6.8) is known as a Dirichlet process mixture model, which in this case clusters over choice alternatives.

A standard multinomial choice model puts a prior on the parameters that assumes that each $\beta_j$ is independent and identically distributed; $\beta_j \sim iid\, H_\beta$ for $j = 1, \ldots, J$. In this case, the parameters values $\beta_j$ are unique. A Dirichlet process prior also allows the parameters to vary over $j$. However, the prior clusters similar categories into groups with unique values of $\beta_j$ with a positive probability.

### Stick-breaking representation

Sethuraman (1994) shows that a Dirichlet process prior is equivalently formulated by the stick-breaking representation,

$$P = \sum_{l=1}^{L_J} p_l \delta(\tilde{\beta}_l), \qquad \tilde{\beta}_l \sim H_\beta, \tag{6.9}$$

where $L_J \to \infty$ and $\delta(\tilde{\beta}_l)$ denotes a unit-mass measure concentrated at $\tilde{\beta}_l$. The Dirichlet process is a distribution over independent and identically distributed draws from the base distribution, with random weights

$$p_1 = V_1, \quad p_l = (1 - V_1)(1 - V_2) \dots (1 - V_{l-1})V_l, \quad l = 2, \dots, L_J, \tag{6.10}$$

where $V_l \sim \text{Beta}(1, \alpha_J)$. This process is also written as $p = \{p_l\}_{l=1}^{L_J} \sim \text{stick}(\alpha_J)$. Since $\sum_{l=1}^{L_J} p_l = 1$, it follows that $p$ can be interpreted as probabilities, and $P$ is a distribution over discrete probability measures.

The construction of the weights $p_l$ in (6.10) is named after the process of iteratively breaking up a stick into pieces. Starting with a unit-length stick, in each step we break off a random proportion of the remaining stick. When we write (6.10) as

$$p_l = V_l \prod_{k=1}^{l-1} (1 - V_k), \tag{6.11}$$

we can interpret $V_l$ as the proportion of the remaining stick which has length $p_l$. After breaking off the first $l - 1$ pieces, the length of the remainder of the stick is $\prod_{k=1}^{l-1}(1 - V_k)$. Since $\text{E}[V_l] = \frac{1}{1+\alpha_J}$, a small $\alpha_J$ results on average in a few large sticks, and the lengths of the remaining sticks are close to zero. For a large value for $\alpha_J$, the weights in $p$ are more evenly distributed.

### Mixture model over outcome categories

The Dirichlet process mixture model in (6.8) can be equivalently formulated by means of the classification variables $C = (C_1, \dots, C_J)$ in (6.4). Using the stick-breaking representation

of the Dirichlet process prior in (6.9), we have

$$\begin{aligned} z_{ij} &= \tilde{\beta}'_{C_j} x_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \mathcal{N}(0,1), \\ C_j|p &\sim \sum_{l=1}^{L_J} p_l \delta(l), \quad p \sim \text{stick}(\alpha_J), \quad \tilde{\beta}_l \sim H_\beta. \end{aligned} \tag{6.12}$$

The probability that an outcome category is assigned to cluster $l$ is denoted by $p_l$. The conditional distribution of $z_{ij}$ now takes the form of a mixture distribution over the outcome categories with random weights $p = (p_1, \ldots, p_{L_J})$,

$$f(z_{ij}|x_i, \tilde{\beta}_1, \ldots, \tilde{\beta}_{L_J}, p) = \sum_{l=1}^{L_J} p_l f_N(z_{ij}|\tilde{\beta}'_l x_i, 1), \tag{6.13}$$

where $f_N(x|\mu, \sigma^2)$ is a normal density with expectation $\mu$ and variance $\sigma^2$ evaluated at $x$. The conditional distribution in (6.13) is an infinite mixture of normal distributions.

The Dirichlet process mixture model over choice alternatives infers whether the parameters of a subset of categories can be treated as a single parameter, or whether the alternative-specific parameters are significantly different to distribute them over different clusters. Even when there are differences between categories, but there is not enough power to distinguish all differences between the values in the category-specific parameter vectors, it may be that the efficiency gain of clustering still outweighs the loss in accuracy. Therefore, the Dirichlet process mixture model only introduces a new parameter vector for a outcome category when this category is significantly different from the other ones.

**Mixture model over explanatory categories**

Along the same lines as for outcome categories, we specify a Dirichlet process mixture model over explanatory categories. We specify a Dirichlet process prior for the explanatory dummy parameters in (6.5),

$$\begin{aligned} z_i &= \gamma w_i + \sum_{k=1}^{K_d} \kappa_{1:J,k} d_{ik} + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, I_J), \\ \gamma &\sim H_\gamma, \quad \kappa_{1:J,k}|Q \sim Q, \quad Q|\alpha_D, H_\kappa \sim DP(\alpha_D, H_\kappa), \end{aligned} \tag{6.14}$$

where $\gamma = (\gamma_1, \ldots, \gamma_J)'$ and $\kappa_{1:J,k} = (\kappa_{1k}, \ldots, \kappa_{Jk})'$, $H_\kappa$ is the base distribution of the parameters $\kappa_{1:J,k}$, and $H_\gamma$ the prior distribution for $\gamma$. In the same way as for the outcome category cluster probabilities, we let the explanatory category cluster probabilities $q = \{q_l\}_{l=1}^{L_D} \sim \text{stick}(\alpha_D)$. Now the stick breaking representation conditional on the classification vector $D = (D_1, \ldots, D_{K_d})$ is

$$\begin{aligned} z_i &= \gamma w_i + \sum_{k=1}^{K_d} \tilde{\kappa}_{1:J,D_k} d_{ik} + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, I_J), \\ D_k|q &\sim \sum_{l=1}^{L_D} q_l \delta(l), \quad q \sim \text{stick}(\alpha_D), \quad \gamma \sim H_\gamma, \quad \tilde{\kappa}_{1:J,l} \sim H_\kappa, \end{aligned} \tag{6.15}$$

where $\tilde{\kappa}_{1:J,l} = (\tilde{\kappa}_{1l}, \ldots, \tilde{\kappa}_{Jl})'$, and $q_l$ is the probability that an explanatory category is assigned to cluster $l$. The elements of the explanatory cluster assignment probability vector $q = (q_1, \ldots, q_{L_D})$ add up to one, $\sum_{l=1}^{L_D} q_l = 1$.

From (6.15) follows the specification of the Dirichlet process mixture model that mixes over the parameter values corresponding to the dummies variables $d_i$,

$$f(z_i|x_i, \gamma, \tilde{\kappa}_{1:J,1}, \ldots, \tilde{\kappa}_{1:J,L_D}, q) = \sum_{l=1}^{L_D} q_l f_N(z_{ij}|\gamma w_i + \sum_{k=1}^{K_d} \tilde{\kappa}_{1:J,l} d_{ik}, 1). \tag{6.16}$$

The Dirichlet process mixture model over explanatory categories in (6.16) clusters dummy parameters together. A subset of explanatory dummy categories with identical parameters is equivalent to aggregating the corresponding explanatory dummy variables.

### Two-way mixture model

We specify a mixture model for the two-way parameter clustering in Section 6.2.2 by combining the mixture model over outcome categories in (6.12) with a mixture model over explanatory categories in (6.16). The result is a two-way Dirichlet process mixture model

$$\begin{aligned} z_{ij} &= \tilde{\gamma}'_{C_j} w_i + \sum_{k=1}^{K_d} \tilde{\kappa}_{C_j,D_k} d_{ik} + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \mathcal{N}(0,1), \quad j = 1, \ldots, J, \\ C_j|p &\sim \sum_{l=1}^{L_J} p_l \delta(l), \quad p \sim \text{stick}(\alpha_J), \quad D_k|q \sim \sum_{k=1}^{L_D} q_k \delta(k), \quad q \sim \text{stick}(\alpha_D), \\ \tilde{\gamma}_l &\sim H_\gamma, \quad \tilde{\kappa}_{lk} \sim H_{\kappa_j}, \quad l = 1, \ldots, L_J, \quad k = 1, \ldots, L_D. \end{aligned} \tag{6.17}$$

The Dirichlet process mixture models in (6.12), (6.16), and (6.17), are tailored for a high-dimensional multinomial probit model. However, they can easily be extended to a mix of a multinomial and conditional choice model by adding a conditional part to (6.2), in which the covariates vary across different outcome categories, but the model parameters are constant over outcome categories. The methods are of less interest to the conditional choice model itself, since the parameters do not grow in the number of choice alternatives. The same holds for ordered choice models, in which only the intercepts are alternative-specific. Another interesting application of the cluster methods is the rank ordered model, in which the model parameters are also alternative-specific.

### Model interpretation

The parameter clusters over outcome categories can be interpreted as preference sets. The latent utility model in (6.3) shows that the deterministic part of the individual utilities for distinct outcome categories $j$ and $k$ is identical when $\beta_j = \beta_k$. From this observation follows directly that the expected utility of choice alternatives within the same parameter cluster is exactly equal. This motivates the preference set interpretation of the parameter clusters over outcome categories. There is an expected preference relation across the sets, while the expected preferences are the same within a preference set. Due to the idiosyncratic part of the utilities, the utilities can still differ across choice alternatives within a preference set.

Since the expected utilities for choice alternatives are identical within a preference set, the probability that an individual prefers one alternative over the other equals 0.5. The probability that the utility for choice alternative $j$ is larger than the utility for choice alternative $k$ equals

$$P(z_{ij} > z_{ik}|x_i) = P(\beta_j' x_i + \varepsilon_{ij} > \beta_k' x_i + \varepsilon_{ik}) = P\left(\eta_{ijk} < \theta_{ijk}\right), \tag{6.18}$$

where $\theta_{ijk} = \frac{(\beta_j - \beta_k)' x_i}{\sqrt{2}}$, and $\eta_{ijk} = \frac{\varepsilon_{ik} - \varepsilon_{ij}}{\sqrt{2}}$ is a standard normally distributed variable. When $\beta_j = \beta_k$,

$$P(z_{ij} > z_{ik}|x_i) = P\left(\eta_{ijk} < 0\right) = \int_{-\infty}^{0} \frac{1}{\sqrt{2\pi}} e^{\frac{-\eta_{ijk}^2}{2}} d\eta_{ijk} = \frac{1}{2}. \tag{6.19}$$

This result shows that the model implicitly assumes that individuals are indifferent between choice alternatives within a preference set.

We can also show that the choice probabilities of the alternatives within a preference set are identical. The odds ratio between two alternatives $j$ and $k$ within the same preference set equals

$$\frac{P(y_i = j|x_i)}{P(y_i = k|x_i)} = \frac{P(z_{ij} > z_{il}\,\forall l|x_i)}{P(z_{ik} > z_{il}\,\forall l|x_i)} = \frac{P(\varepsilon_{il} - \varepsilon_{ij} < (\beta_j - \beta_l)'x_i\,\forall l)}{P(\varepsilon_{il} - \varepsilon_{ik} < (\beta_k - \beta_l)'x_i\,\forall l)} = 1. \tag{6.20}$$

The last equality in (6.20) follows from the fact that $\beta_j = \beta_k$ and all $\varepsilon_{ij}$ are identically distributed.

To illustrate the concept of preference sets, we consider some examples. A supermarket costumer may be more inclined to buy a diet coke, but does not care about the brand. In this case diet cokes are preferred over regular cokes, but all different diet cokes get the same probability of being chosen. A lot of holidaymakers prefer a holiday by car over a holiday involving air travel, but do not have strong preferences for different neighbouring countries. In other words, they cluster choice alternatives in preference sets. There is a preference ranking over categories across these sets, but individuals are more or less indifferent between categories within a set.

Preference sets only reduce the parameter space when individuals are assumed to assign in expectation the same utility to different choice alternatives. An alternative assumption is that individuals only have a utility function for a subset of the alternatives in the total choice set. The alternatives in this consideration set are evaluated along the lines of random utility maximization, and the remaining alternatives get a zero probability of being chosen.

The concept of preference sets is closely related to the framework of consideration sets. When the utilities in one preference set go to minus infinity, and the remaining subsets only contain one category, we have a consideration set. However, by clustering over alternative-specific parameters, and implicitly grouping outcome categories in preference sets, the final choice is made from the total choice set while potentially greatly reducing the number of parameters to be estimated.

Estimating the mixture model on data generated from this two-stage consideration set process, requires $(CS + 1) \times K + 1$ parameter estimates, where $CS$ denotes the number of choice alternatives in the consideration set. We have $CS \times K$ alternative specific parameters

in the consideration set, $K$ parameters equal to minus infinity corresponding to the alternatives outside the consideration set, and the concentration parameter controlling the clustering over the unconsidered choice alternatives. The number of required parameters for a preference set data generating process is $PS \times K + 1$, where $PS$ is the number of preference sets, and $J \times K + 1$ parameters are sufficient for a standard data generating process with unique parameter values.

The consideration set literature can be divided into two approaches to consideration set modelling. First, we have models that impose a hard constraint on choice alternatives to be included in the consideration set. For instance, Gilbride and Allenby (2004) use the deterministic part of the utility function to screen alternatives, or set threshold values on each explanatory variable. The first approach does not reduce the number of (alternative-specific) parameters in a multinomial setting, but the second only requires $(CS + 1) \times K$ parameters. However, when there are $PS$ preference sets with positive probability present in the data, the consideration set model needs $J \times K$ parameter estimates.

Second, we have models that separately model the consideration set probability from the choice probability. Andrews and Srinivasan (1995), Chiang et al. (1998), and Mehta et al. (2003) model the $2^J - 1$ probabilities of all potential consideration sets. As the number of possible consideration sets is exponential in the number of choice alternatives, estimation becomes computationally infeasible in large choice sets. Bronnenberg and Vanhonacker (1996) and Van Nierop et al. (2010) model the consideration set inclusion probabilities for all alternatives. This approach results in $J$ additional probabilities to the set of choice probabilities for the considered alternatives, and therefore does not alleviate the curse of dimensionality in a standard multinomial choice model.

## 6.3 Bayesian Inference

To estimate the posterior distributions of the parameters in $\beta$, we approximate the two-way Dirichlet process mixture model by truncating the Dirichlet processes at the $L$th term by setting $V_L = 1$ for a finite number $L$. For inference in the truncated Dirichlet process we build upon the Gibbs sampler described by Ishwaran and James (2002) that is simpler than corresponding samplers for the full Dirichlet process, while displaying favorable mixing

properties. This sampler is extended to suit a multinomial choice model and we introduce new sampling steps for clustering over outcome and explanatory categories.

### 6.3.1 Truncation level

When we set the truncation level $L$ equal to the number of available categories, the truncated model is in practice equal to the full Dirichlet process mixture model. However, smaller values for $L$ lead to smaller computational times. When $L$ is still larger than the expected number of clusters in the data, posterior results are indistinguishable from results based on the full Dirichlet process. The value for the truncation level can be different for the clustering over outcome categories and the clustering over explanatory categories.

The stick-breaking representation of the Dirichlet process prior, as in Section 6.2.3, provides a guideline for selecting the truncation level $L$. When the higher order probabilities $p = \{p_l\}_{l=L}^{\infty}$ in (6.9), or $q$ for the explanatory categories, are small enough, the approximation error is negligible. Ishwaran and Zarepour (2000) derive the moments of the tail probability $\sum_{l=L}^{\infty} p_l$,

$$\mathrm{E}\left[\sum_{l=L}^{\infty} p_l\right] = \left(\frac{\alpha}{\alpha+1}\right)^{L-1}, \quad \mathrm{var}\left[\sum_{l=L}^{\infty} p_l\right] = \left(\frac{\alpha}{\alpha+2}\right)^{L-1} - \left(\frac{\alpha}{\alpha+1}\right)^{2L-2}, \tag{6.21}$$

which are the mean and the variance of the tail probability, respectively. Using the mean and variance of the tail probability, we can test for a particular concentration parameter $\alpha$ whether the truncation level results in a small enough approximation error. The fact that the mean tail probability increases in the concentration parameter $\alpha$, confirms that the number of clusters is proportional to $\alpha$.

### 6.3.2 Concentration parameter

The concentration parameter $\alpha$ controls the number of clusters. Hence, the value for $\alpha$ implies a prior distribution for the number of unique parameter values $L^*$,

$$Pr[L^* = j|\alpha] = c(j, J)J!\alpha^j \frac{\Gamma(\alpha)}{\Gamma(\alpha + J)}, \tag{6.22}$$

where we cluster over $j = 1, \ldots, J$ parameters, and $c(j, J) = Pr[L^* = j|\alpha = 1]$. This implied prior distribution over the number of clusters is derived by Antoniak (1974), and Escobar and West (1995) discuss how the factors $c(j, J)$ are calculated. The distribution runs from $L^* = 1$, which means no parameter variation at all, to $L^* = J$ with unique parameter values for each $j$.

Suppose we have a prior belief about the number of clusters $L^*$. Van den Hauwe (2015) proposes to choose a value for the concentration parameter $\alpha$ that sets the prior mode of $L^*$ equal to that belief. The concentration parameter that matches the belief mode$[L^*] = m^*$ is

$$\alpha_{m^*} = \frac{1}{2}\left(\exp(-\delta c(m^* + 1)) + \exp(-\delta c(m^*))\right), \tag{6.23}$$

with $\delta c(1) = \log(c(1, J))$ and $\delta c(m^*) = \log(c(m^*, J)) - \log(c(m^* - 1, J))$ for numerical stability.

By choosing $\alpha$ as in (6.23) we control the prior mode of the distribution of the number of clusters. Conley et al. (2008) shows that for a range of fixed values for the concentration parameters, the prior distributions on the number of clusters are very informative. By putting a prior on the concentration parameter, we can also govern the variance of the distribution of the number of clusters.

We specify a prior distribution on $\alpha$ with prior mean equal to the value in (6.23). To check whether the prior induces enough dispersion around the prior mode of $L^*$, we evaluate the marginal prior probability density function

$$f(L^*) = \int f(L^*|\alpha) f(\alpha) d\alpha, \tag{6.24}$$

where $f(L^*|\alpha)$ is the probability function in (6.22) and $f(\alpha)$ is the prior probability density function of $\alpha$. The integral in (6.24) is evaluated using Monte Carlo integration.

### 6.3.3 Prior distributions

The Dirichlet process mixture model is defined by a Dirichlet process prior on the parameters $\beta$. To complete the prior specification for $\beta$, we specify the base distribution $H_\beta$,

$$\beta_{1k} \sim \mathcal{N}(0, 0) \text{ and } \beta_{jk}|\sigma_\beta^2 \sim \mathcal{N}(0, \sigma_\beta^2), \tag{6.25}$$

where $j = 2, \ldots, J$, $k = 1, \ldots, K$, and $\sigma_\beta \in \mathcal{R}^+$. Note that the normal distribution turns into the Dirac delta function $\delta(0)$ when the variance is zero. We let the data determine the number of clusters by treating the concentration parameter $\alpha$ as unknown with a prior distribution,

$$\alpha|\eta_1, \eta_2 \sim \text{Gamma}(\eta_1, \eta_2), \tag{6.26}$$

where Gamma$(\eta_1, \eta_2)$ denotes a gamma distribution with mean $\eta_1/\eta_2$. The values $(\eta_1, \eta_2) \in \mathcal{R}^+$ directly effect the number of estimated clusters through the concentration parameter, where larger values for $\alpha$ encourage more distinct values for the coefficients. We set $\eta_1/\eta_2$ equal to $\alpha_{m^*}$ in (6.23) and use $\eta_2$ to govern the dispersion around the mean.

### 6.3.4 Posterior distribution

To estimate the posterior distributions of the parameters, we rely on a Markov Chain Monte Carlo sampler with data augmentation. The representations of the mixture models in (6.12) and (6.15) condition on the choice alternative classification variable $C$ and the explanatory dummy category classification variable $D$, respectively. Using these representations of the model allows for clustering over both the outcome category and the explanatory category dimension, by simulating the latent classification variables alongside the model parameters in $\beta$ and the cluster probabilities $p$ and $q$.

Ishwaran and Zarepour (2000) and Ishwaran and James (2002) derive a Gibbs sampler for finite normal mixture models using a truncation approximation of the Dirichlet process. The sampler is developed for normal mixtures over the observations $i = 1, \ldots, N$. We extend the sample algorithm to suit a multinomial choice model and we introduce new sampling steps for clustering over outcome and explanatory categories.

In each iteration of the Gibbs sampler, we sample the parameters $\beta$, $p$, and $q$ together with the latent classification variables $C$ and $D$ from their full conditional distributions, given the data $y = (y_1, \ldots, y_N)'$ and $x = (x_1, \ldots, x_N)'$. The Markov Chain Monte Carlo simulation scheme is as follows:

1. Sample $z|\tilde{\beta}, C, D, y, x$
2. Sample $\tilde{\beta}|C, D, \sigma_\beta, z, x$

3. Sample $C|p, \tilde{\beta}, D, z, x$

4. Sample $D|q, \tilde{\beta}, C, z, x$

5. Sample $p|C, \alpha_J$ and $q|D, \alpha_D$

6. Sample $\alpha_J|p, \eta_1, \eta_2$ and $\alpha_D|q, \zeta_1, \zeta_2$

The first sampling step distinguishes the sampling algorithm for the multinomial probit model from a normal mixture model. Since the multinomial probit model can be represented by a set of Gaussian latent variables, as we show in (6.2), sampling the latent variables $z = (z_1, \ldots, z_N)'$ conditional on the observed choices in $y$ is sufficient (Allenby and Rossi, 1998). Since $z$ contains continuous normally distributed variables, it can serve as dependent variable in the sampling steps of Ishwaran and James (2002).

The model parameters in the $L_j$ by $K_w + L_D$ parameter matrix $\tilde{\beta}$, with rows $\tilde{\beta}_l = (\tilde{\gamma}_l', \tilde{\kappa}_{l,1}, \ldots, \tilde{\kappa}_{l,L_D})$, are sampled in the second step conditional on $z$. This step extends the sampler of Ishwaran and James (2002) in two directions. First, their sample algorithm is developed for normal mixtures over the observations $i = 1, \ldots, N$, which is relatively straightforward since clusters of observations are independent of each other. We sample parameter values for clusters over the dimensions $j = 1, \ldots, J$ and $k = 1, \ldots, K_d$. Second, we extend sampling model parameters over one-way clusters to two-way clustering, by sampling the model parameters simultaneously over the outcome and explanatory category clusters.

The third and fourth sampling steps draw the classification vectors. For identification purposes, the first outcome category is in every iteration of the sampler assigned to the first cluster, in which the parameter values $\tilde{\beta}_1$ are equal to zero. Since the categorical explanatory regressors $d_i$ may be correlated with explanatory variables in $w_i$, there is potential dependence between parameters corresponding to different category dummies, which should be taken into account when sampling the classification variables $D$.

The probabilities of each cluster of outcome categories and the probabilities of each cluster of explanatory categories are sampled in the fifth step, and finally we resample the concentration parameters.

### 6.3.5 Posterior simulation

Let $C^* = \{C_1^*, \ldots, C_{m_j}^*\}$ denote the current $m_j$ unique values of $C$ excluding $C = 1$, and $r_l$ the number of values in $C$ which equal $l$. Let $D^* = \{D_1^*, \ldots, D_{m_d}^*\}$ denote the current $m_d$ unique values of $D$. The sampling steps in each iteration of the sampler are:

**Step 0.** Initialize the sampler by a draw from the prior distributions. Sample the initial draw for the model parameters as $\tilde{\beta}_l|\sigma_\beta \sim \mathcal{N}(0, \sigma^2)$, where $\sigma^2 = \sigma_\beta^2$ when $l \neq 1$ and $\sigma^2 = 0$ when $l = 1$. The initial draw for the concentration parameters is $\alpha_J|\eta_1, \eta_2 \sim$ Gamma$(\eta_1, \eta_2)$ and $\alpha_D|\zeta_1, \zeta_2 \sim$ Gamma$(\zeta_1, \zeta_2)$ and for the latent variables $p|\alpha_J \sim$ stick$(\alpha_J)$, $C_j|p \sim \sum_{l=1}^{L_J} p_l \delta(l)$, $q|\alpha_D \sim$ stick$(\alpha_D)$, and $D_k|q \sim \sum_{l=1}^{L_D} q_l \delta(l)$.

Initialize the latent variables $z_i$ by a draw from a standard normal distribution and center the vector at zero. Permute the elements so that the maximum of each $z_i$ coincides with $y_i$.

**Step 1.** Given $\tilde{\beta}$, $C$, $D$, $y$, and $x$, sample the latent variables $z_{ij}$ for $i = 1, \ldots, N$ and for $j = 1, \ldots, J$. Following from (6.1), $z_{ij} \geq \max(z_i^{(j)})$ if $y_i = j$ and $z_{ij} \leq \max(z_i^{(j)})$ if $y_i \neq j$, where $z_i^{(j)} = (z_{i1}, \ldots, z_{i,j-1}, z_{i,j+1}, \ldots, z_{iJ})$. Hence, sample $z_{ij}$ according to

$$z_{ij}|z_i^{(j)}, \tilde{\beta}_{C_j}, D, y_i, x_i \sim \mathcal{N}_{+\max(z_i^{(j)})}\left(\tilde{\gamma}_{C_j}' w_i + \sum_{k=1}^{K_d} \tilde{\kappa}_{C_j, D_k} d_{ik}, 1\right) \text{ for } j = y_i,$$

$$z_{ij}|z_i^{(j)}, \tilde{\beta}_{C_j}, D, y_i, x_i \sim \mathcal{N}_{-\max(z_i^{(j)})}\left(\tilde{\gamma}_{C_j}' w_i + \sum_{k=1}^{K_d} \tilde{\kappa}_{C_j, D_k} d_{ik}, 1\right) \text{ for } j \neq y_i,$$

where $\mathcal{N}_{+a}(\mu, \sigma^2)$ and $\mathcal{N}_{-a}(\mu, \sigma^2)$ represent a normal distribution with expectation $\mu$ and variance $\sigma^2$ truncated from below or above by $a$, respectively.

**Step 2.** Given $C$, $D$, $\sigma_\beta$, $z$, and $x$, sample the model parameters $\tilde{\beta}_l$ for $l = 1, \ldots, L_J$. Distinguish three different cases to sample all parameters in $\tilde{\beta}_l$:

1. For $l \in \{C_1^*, \ldots, C_{m_j}^*\}$ and $\tilde{\kappa}_{l,k}$ with $k \in \{D_1^*, \ldots, D_{m_d}^*\}$, sample $\beta_l^* = (\tilde{\gamma}_l', \tilde{\kappa}_{l,D_1^*}, \ldots, \tilde{\kappa}_{l,D_{m_d}^*})$ in

$$Z_l = \beta_l^* X_l' + \eta, \tag{6.27}$$

where $\eta$ is a $1 \times (r_l \times N)$ matrix with independent and identically standard normal distributed elements. The dependent variable $Z_l = (z_1^l, \ldots, z_N^l)$ is defined as a $1 \times (r_l \times N)$ matrix, in which $z_i^l$ are row vectors stacking all $z_{ij}$ for which $C_j = l$. Aggregate the dummies within each cluster, $x_i^* = (w_i', \sum_{k:D_k=D_1^*} d_{ik}, \ldots, \sum_{k:D_k=D_{m_d}^*} d_{ik})'$, set $x^* = (x_1^*, \ldots, x_N^*)'$, and stack $r_l$ times the matrix $x^*$ in the $(r_l \times N) \times (K_w + m_d)$ matrix $X_l$. Sample $\beta_l^*$ according to

$$\beta_l^* | C, D, \sigma_\beta, z, x \sim \mathcal{N}(b, B^{-1}),\ b = Z_l X_l B^{-1},\ B = X_l' X_l + \frac{1}{\sigma_\beta^2} I_{K_w + m_d}.$$

2. For $l \in C - \{C_1^*, \ldots, C_{m_j}^*\}$ and $\tilde{\kappa}_{l,k}$ with $k \in \{D_1^*, \ldots, D_{m_d}^*\}$, sample $\beta_l^*$ from the base distribution as $\beta_l^* | C, D, \sigma_\beta, z, x \sim \mathcal{N}(0, \sigma^2 I_{K_w + m_d})$, where $\sigma^2 = \sigma_\beta^2$ when $l \neq 1$ and $\sigma^2 = 0$ when $l = 1$.

3. For $l \in C$ and $\tilde{\kappa}_{l,k}$ with $k \in D - \{1, D_1^*, \ldots, D_{m_d}^*\}$, sample $\tilde{\kappa}_{lk}$ from the base distribution as $\tilde{\kappa}_{lk} | C, D, \sigma_\beta, z, x \sim \mathcal{N}(0, \sigma^2)$, where $\sigma^2 = \sigma_\beta^2$ when $l \neq 1$ and $\sigma^2 = 0$ when $l = 1$.

**Step 3.** Given $p$, $\tilde{\beta}$, $D$, $z$, and $x$, sample the classification vector of the outcome categories $C = (1, C_2, \ldots, C_J)$ according to

$$C_j | p, \tilde{\beta}, D, z, x \sim \sum_{l=1}^{L_J} \pi_{lj} \delta_l, \tag{6.28}$$

for $j = 2, \ldots, J$. The conditional cluster probability $\pi_{lj}$ is a function of the unconditional cluster probability $p_l$ and the likelihood contributions of the latent utilities of each outcome category $z_j$ and the observed explanatory variables $x$, for the parameter value $\tilde{\beta}_l$. Since $z_{i1}, \ldots, z_{iJ}$ are conditionally independent,

$$\begin{aligned}(\pi_{1j}, \ldots, \pi_{L_J,j}) \propto \Bigg( & p_1 \exp\left( -\frac{1}{2} \sum_{i=1}^{N} (z_{ij} - \tilde{\gamma}_1' w_i + \sum_{k=1}^{K_d} \tilde{\kappa}_{1,D_k} d_{ik}) \right), \ldots, \\ & p_{L_j} \exp\left( -\frac{1}{2} \sum_{i=1}^{N} (z_{ij} - \tilde{\gamma}_{L_j}' w_i + \sum_{k=1}^{K_d} \tilde{\kappa}_{L_J,D_k} d_{ik}) \right) \Bigg).\end{aligned} \tag{6.29}$$

**Step 4.** Given $q$, $\tilde{\beta}$, $C$, $z$, and $x$, sample the classification vector of the explanatory categories $D = (D_1, \ldots, D_{K_d})$ according to

$$D_k | q, \tilde{\beta}, C, z, x \sim \sum_{l=1}^{L_D} \psi_{lk} \delta_l, \tag{6.30}$$

for $k = 1, \ldots, K_d$. Since clusters of different explanatory dummies are not necessarily independent, we cannot distinguish between likelihood contributions of each explanatory category, as we do for individuals or outcome categories. To measure likelihood contribution of each cluster value for the different category dummy coefficients, we introduce

$$\ddot{\kappa}_{C_j,kl} = (\tilde{\kappa}_{C_j,D_1}, \ldots, \tilde{\kappa}_{C_j,D_{k-1}}, \tilde{\kappa}_l, \tilde{\kappa}_{C_j,D_{k+1}}, \ldots, \tilde{\kappa}_{C_j,D_{K_d}}),$$

which is the coefficient vector $\tilde{\kappa}_{C_j}$ based on the classification vector $D$ of the previous iteration of the sampler, where the coefficient corresponding to the $k$th dummy is replaced by the coefficient value of cluster $l$. Now the conditional cluster probabilities $\psi_{lk}$ are a function of the unconditional cluster probabilities $q_l$ and the data $(z, x)$,

$$(\psi_{1k}, \ldots, \psi_{L_D,k}) \propto \left( q_1 \exp\left( -\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{J} (z_{ij} - \tilde{\gamma}'_{C_j} w_i - \ddot{\kappa}_{C_j,k1} d_i) \right), \ldots, \right.$$
$$\left. q_{L_d} \exp\left( -\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{J} (z_{ij} - \tilde{\gamma}'_{C_j} w_i - \ddot{\kappa}_{C_j,k,L_D} d_i) \right) \right).$$

**Step 5.** Given $C$ and $\alpha_J$, sample the unconditional cluster probabilities for the outcome categories from $p|C, \alpha_J$ according to

$$p_1 = V_1^*, \quad p_l = (1 - V_1^*)(1 - V_2^*) \ldots (1 - V_{l-1}^*) V_l^*, \text{ for } l = 2, \ldots, L_J - 1,$$

where

$$V_l^* \sim \text{Beta}\left( 1 + r_l, \alpha_J + \sum_{k=l+1}^{L_J} r_k \right), \quad l = 1, \ldots, L_J - 1.$$

Given $D$ and $\alpha_D$, sample the unconditional cluster probabilities for the explanatory categories $q$ in the same way as for $p$.

**Step 6.** Given $p$, $\eta_1$, and $\eta_2$, sample the concentration parameter for the outcome categories $\alpha_J$ according to

$$\alpha_J|p,\eta_1,\eta_2 \sim \text{Gamma}\left(L_J+\eta_1-1, \eta_2-\sum_{l=1}^{L_J-1}\log(1-V_l^*)\right). \tag{6.31}$$

Given $q$, $\zeta_1$, and $\zeta_2$, sample the concentration parameter for the explanatory categories $\alpha_D$ in the same way as for $\alpha_J$.

**Step 7.** Go to Step 1.

Note that this sample algorithm clusters parameters over both outcome and explanatory categories. In case we only want to cluster over outcome categories, we simply put all explanatory variables in $w_i$. The vector $d_i$ remains empty, which means that we do not have to restructure the dummy variables and sample their parameters $\tilde{\kappa}$ in Step 2, and ignore Step 4 of the sample algorithm. On the other hand, when we only cluster parameters over explanatory variables, we set $L_J = J$, $C = (1, 2, \ldots, J)$, and skip Step 3.

### 6.3.6 Predictive Distributions

To construct a predictive density, we make use of the in-sample posterior conditional cluster probabilities and the truncation level of the Dirichlet process. We draw the cluster assignment of an out-of-sample observation from the posterior mixture distribution of that observation. Moreover, since the truncation level is assumed to be much larger than the number of in-sample clusters, the model allows for new clusters of model parameters out-of-sample.

The predictive densities of $y_i$ for different individuals $i = N+1, \ldots, N+h$ are simulated by means of (6.1) and (6.2) in each iteration of the sampler, together with the parameter draws obtained in that sample iteration. In iteration $s$ of the sampler, we have

$$y_i^{(s)}(z_{ij}^{(s)}) = j \text{ if } z_{ij}^{(s)} = \max(z_i^{(s)}), \tag{6.32}$$

where $z_i^{(s)} = (z_{i1}^{(s)}, \ldots, z_{iJ}^{(s)})$ is a vector with draws of the latent utilities in iteration $s$ of the sampler. We obtain a draw from the predictive density of $z_{ij}$ as follows

$$z_{ij}^{(s)} = \tilde{\gamma}_{C_j^{(s)}}^{(s)\prime} w_i + \sum_{k=1}^{K_d} \tilde{\kappa}_{C_j^{(s)}, D_k^{(s)}}^{(s)} d_{ik} + \varepsilon_{ij}^{(s)}, \quad \varepsilon_{ij}^{(s)} \sim \mathcal{N}(0, I_J), \tag{6.33}$$

for $j = 1, \ldots, J$, where $\tilde{\gamma}^{(s)}$ and $\tilde{\kappa}^{(s)}$ are the parameter draws for $\tilde{\gamma}$ and $\tilde{\kappa}$ in iteration $s$ of the sampler. We sample $C_j^{(s)}|\pi^{(s)} \sim \sum_{l=1}^{L_J} \pi_{lj}^{(s)} \delta(l)$ and $D_k^{(s)}|\psi^{(s)} \sim \sum_{l=1}^{L_D} \psi_{lk}^{(s)} \delta(l)$, where $\pi^{(s)}$ and $\psi^{(s)}$ are the matrices with conditional cluster probabilities in iteration $s$ of the sampler.

The draw from the predictive density is conditional on the draws for the model parameters $\tilde{\beta}^{(s)}$ and the conditional cluster probabilities $\pi^{(s)}$ and $\psi^{(s)}$ in iteration $s$ of the sampler. Since the truncation levels $L_J$ and $L_D$ are assumed to be much larger than the expected number of clusters in the outcome and explanatory categories, respectively, future model parameter values can be drawn from new clusters which are not present in-sample.

### 6.3.7 Label-Switching

The posterior distribution is invariant to the labels of the clusters. Between iterations of the sampler, the classification values in $C$ and $D$ can switch between different clusters. If a label switch occurs in the classification vectors during the posterior simulation, statistics such as the cluster specific posterior means and standard deviations of model parameters become uninformative (Frühwirth-Schnatter, 2001; Geweke, 2007b; Bauwens et al., 2017).

Moreover, the number of unique parameter values can vary over the iterations of the sampler. Since the model parameters have a positive cluster probability for each cluster in each iteration of the sampler, model parameters can be distributed over more clusters in one iteration than the other. This complicates the calculation of cluster specific functions even further.

When the statistics of interest are label invariant, we can ignore the label-switching. The posterior distribution of the number of different clusters, the posterior distributions of the model parameters per category, instead of per cluster, and the predictive distributions are examples of invariant functions of the posterior draws.

### 6.3.8 Convergence diagnostics

To infer whether we use enough draws from our posterior simulator, we analyze inefficiency factors $1 + 2\sum_{f=1}^{\infty} \rho_f$, where $\rho_f$ is the $f$th order autocorrelation of the chain of draws for a specific parameter. We use the Bartlett kernel as in Newey and West (1987) with a bandwidth of four percent of the sample draws. The inefficiency factors equal the variance of the mean of the posterior draws from the sampler, divided by the variance of the mean assuming independent draws. When we require the variance of the mean of the posterior draws to be limited to at most one percent of the variation due to the data, the inefficiency factor provides an indication of the minimum number of draws to achieve this, see Kim et al. (1998).

We also test for convergence of the sampler by the Geweke (1992) t-test for the null hypothesis of equality of the means computed from the first 20 percent and the last 40 percent of the sample draws. We compute the variances of the means using the Newey and West (1987) heteroskedasticity and autocorrelation robust variance estimator with a bandwidth of four percent of the sample sizes.

## 6.4 Simulation Study

This section examines the practical implications of the developed parameter clustering methods on simulated data. We estimate the two-way Dirichlet process mixture model in Section 6.2.3, and show the clustering over outcome and explanatory categories separately at work. Each study shows the properties of the mixture methods and compares the performance to a standard multinomial choice model.

### 6.4.1 General set-up

The choice data are generated from a multinomial choice model with control variables and a categorical explanatory variable. The outcome categories and the explanatory categories vary both over two parameter clusters. The data generating process takes the form

$$\begin{aligned} y_i(z_i) &= j \text{ if } z_{ij} = \max(z_i), \\ z_{ij} &= \gamma_j' w_i + \kappa_j' d_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \mathcal{N}(0,1), \end{aligned} \tag{6.34}$$

with $j = 1, \ldots, J$ and $i = 1, \ldots, N$. The vector $w_i = (w_{i1}, w_{i2})$ includes two standard normally distributed variables $w_i \sim \mathcal{N}(0, I_2)$. The categorical dummies in $d_i$ are drawn from a multinomial distribution

$$d_i = (d_{i1}, \ldots, d_{i,K_d}) \sim \text{Multinomial}\left(1 - p_{d_i}, \frac{p_{d_i}}{K_d - 1}, \ldots, \frac{p_{d_i}}{K_d - 1}\right), \tag{6.35}$$

where $p_{d_i} = \frac{\exp(w_{i2})}{1+\exp(w_{i2})}$ so that the explanatory variables are correlated with each other.

We mimick the dimensions of the empirical application in Section 6.5 and apply the Gibbs sampler to $N = 4000$ observations simulated from the data generating process, and use another $1000$ observations for out-of-sample analysis. We set the number of outcome categories to $J = 50$ and the number of explanatory categories to $K_d = 10$. The outcome and explanatory categories are clustered into two groups, with model parameter values $\tilde{\beta}_l = (\tilde{\gamma}_l', \tilde{\kappa}_{l,1}, \ldots, \tilde{\kappa}_{l,L_D})$ equal to

$$\begin{aligned} \tilde{\beta}_1 &= (\tilde{\gamma}_1', \tilde{\kappa}_{1,1}, \tilde{\kappa}_{1,2}) = \begin{pmatrix} 0 & 0 & 0 & 0 \end{pmatrix}, \\ \tilde{\beta}_2 &= (\tilde{\gamma}_2', \tilde{\kappa}_{2,1}, \tilde{\kappa}_{2,2}) = \begin{pmatrix} -1 & 1 & -1 & 2 \end{pmatrix}, \end{aligned} \tag{6.36}$$

where $\beta_j = (\tilde{\gamma}_{C_j}', \tilde{\kappa}_{C_j,D_1}, \ldots, \tilde{\kappa}_{C_j,D_{10}})$ with $C_j = 1$ for $j = 1, \ldots, 25$ and $C_j = 2$ for $j = 26, \ldots, 50$, and $D_k = 1$ for $k = 1, \ldots, 5$ and $D_k = 2$ for $k = 6, \ldots, 10$.

We set $L_J = L_D = 10$. Since $L_D = K_d$, we estimate a full Dirichlet process for the explanatory categories. By truncating the number of possible outcome category clusters to $L_J$, we obtain a potential approximation error. The expectation and the variance of the aggregated higher order probabilities equal $0.001$ and $2.979 \times 10^{-5}$ for the sampled $\alpha_J$ in the last iteration of the Gibbs sampler for the two-way mixture model. These numbers confirm that the approximation error is negligible in the posterior simulation. The priors for the concentration parameters are parametrized according to the procedure in Section 6.3.2. The means of the prior distributions equal the concentration parameters that match the prior belief that the mode of unique parameter values equals five. That results in the distributions $\alpha_J \sim \text{Gamma}(1.30 \times 10, 10)$ with $\text{var}(L_J^*) = 4.27$ and $\alpha_D \sim \text{Gamma}(3.47 \times 1, 1)$ with $\text{var}(L_D^*) = 2.99$. The prior variance of the model parameters is set to $\sigma_\beta^2 = 1$, which allows for a wide range of plausible values within a multinomial choice model.

Posterior results are based on 100,000 iterations of the Gibbs sampler, from which the first 50,000 are discarded and we use a thinning value of 10.

### 6.4.2 Evaluation criteria

We examine the estimation performance of the Bayesian method by a set of diagnostics of the posterior parameter densities; the posterior mean, the mean squared error (MSE) and the mean absolute error (MAE) of the posterior draws, and the interquartile range (IQR) of the posterior parameter distributions.

The in-sample and out-of-sample fit is evaluated by the hit rate and the root mean squared probability error (RMSPE). The hit rate is defined as the fraction of correct predictions in the predictive distribution

$$H_{\text{cat}} = \frac{1}{NS} \sum_{s=1}^{S} \sum_{i=1}^{N} I(y_i = y_i^{(s)}), \tag{6.37}$$

where $S$ denotes the number of samples from the predictive density, $I(A)$ is an indicator function that equals one if event $A$ occurs and zero otherwise, and $y_i^{(s)}$ is defined in (6.32) as a draw from the predictive density of $y_i$ conditional on the draws for the model parameters in iteration $s$ of the sampler.

Since the data generating process divides the choice set into two sets within individuals have identical expected preferences for alternatives, it is also of interest whether the model can correctly predict which preference set is chosen,

$$H_{\text{set}} = \frac{1}{NS} \sum_{s=1}^{S} \sum_{i=1}^{N} I((y_i \leq 25 \text{ and } y_i^{(s)} \leq 25) \text{ or } (y_i > 25 \text{ and } y_i^{(s)} > 25)). \tag{6.38}$$

The hit rate only weighs correct predictions, but does not reward predictions which are close to the realized value. Therefore, we also evaluate model performance based on the implied category probabilities. The root mean squared probability error measure benefits posterior conditional category probabilities which are close to these probabilities implied by

the data generating process,

$$\text{RMSPE}_{\text{cat}} = \sqrt{\frac{1}{NJ}\sum_{i=1}^{N}\sum_{j=1}^{J}\Big(\widehat{P(y_i = j|x_i)} - P(y_i = j|x_i)\Big)}, \tag{6.39}$$

where $\widehat{P(y_i = j|x_i)} = \frac{1}{S}\sum_{s=1}^{S} I(y_i^{(s)} = j)$, and $P(y_i = j|x_i)$ is simulated from the data generating process. The RMSPE per preference set is calculated as

$$\text{RMSPE}_{\text{set}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\Big(\widehat{P(y_i > 25|x_i)} - P(y_i > 25|x_i)\Big)}. \tag{6.40}$$

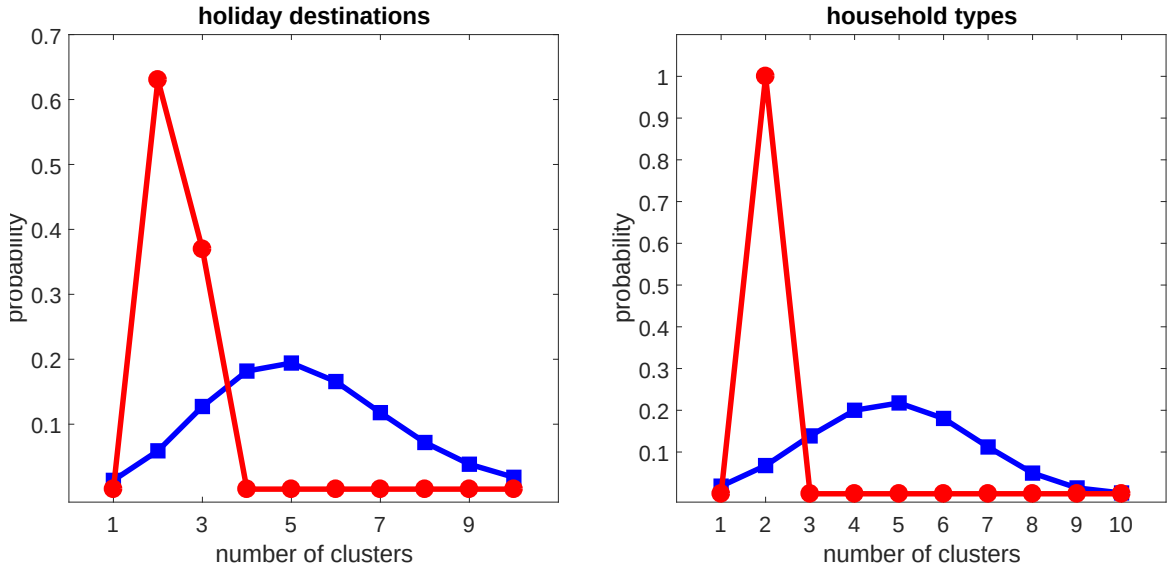
### 6.4.3 Two-way parameter clustering

We estimate the parameters in the two-way mixture model in (6.17), as outlined in Section 6.3, on a draw from the data generating process in (6.34). We obtain posterior distributions for each parameter in $\beta$ and the classification parameters $C$ en $D$. Based on the posterior results we find that the mixture model accurately estimates the number of different parameter clusters and correctly assigns categories to clusters. Relative to a standard multinomial choice model, clustering parameters over categories concentrates the modes of the posterior parameter distributions around the parameter values in the data generating process. Moreover, the uncertainty around these values strongly decreases. The gains in accuracy and precision are confirmed by a range of performance measures.

We find a posterior mode for the number of distinct parameter values of two for both the outcome categories and the explanatory categories, which is equal to the number of clusters in the data generating process. Figure 6.1 shows the posterior distributions together with the prior distributions. The posterior probability of two parameter clusters over outcome categories is 63 percent, and 37 percent for three clusters. The posterior for the explanatory categories shifts all probability mass under the relatively uninformative prior to two clusters.

Because of label-switching, we cannot compute cluster-specific statistics. Moreover, as we show in Figure 6.1, the number of parameter clusters is not a fixed value but can differ over sample iterations. However, after the sampler converged, we can check which categories tend to cluster together by calculating the percentage of sample iterations in which

**Figure 6.1:** Two-way: Distribution number of unique parameter values



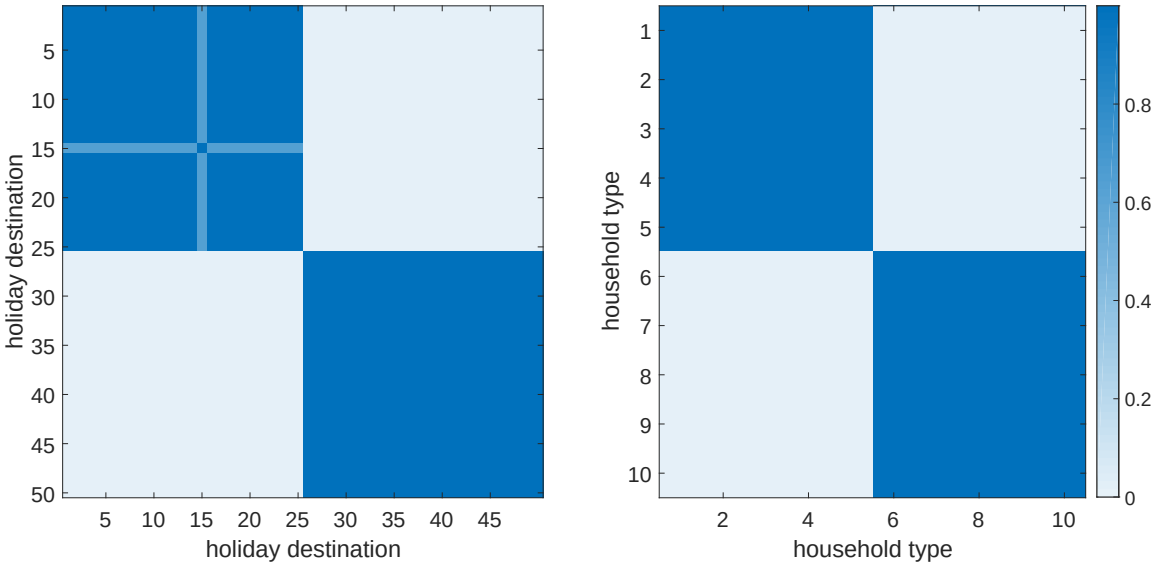

This figure shows the prior (□ in blue) and posterior (◯ in red) distribution over the number of unique parameter values $L^*$ over the outcome categories (left panel) and the explanatory categories (right panel) in the two-way mixture model.

two particular categories are in the same parameter cluster. Figure 6.2 shows the posterior probabilities that two categories share the same parameter values.

Figure 6.2 displays that the cluster assignment of the categories is very close to what is expected based on the data generating process. With posterior probability equal to zero, two categories share a parameter cluster when that is not the case in the data generating process. Almost all first 25 outcome categories are assigned to a cluster with the base category, in which all parameter values are exactly zero. Since the 15th outcome category is assigned its own cluster in some sample iterations, this category has a slightly lower posterior probability of sharing a cluster with the base category. All final 25 categories are in one and the same cluster with probability one. Since the posterior distribution of the number of explanatory parameter clusters in Figure 6.1 is concentrated at two, there is even less uncertainty around the cluster memberships of the explanatory variables.

Mixing a large number of categories into a relatively small number of parameter clusters improves in parameter estimation efficiency. Figure 6.3 shows the posterior parameter distributions of the two-way mixture model and a standard multinomial probit model. The upper two panels show the distributions of the parameters of the variables in $w_i$ over all outcome categories, and the lower panel the parameters of the dummies in $d_i$ over all outcome and explanatory categories. The mixture model estimates considerable thinner posterior distri-

**Figure 6.2:** Two-way: Posterior probabilities cluster memberships



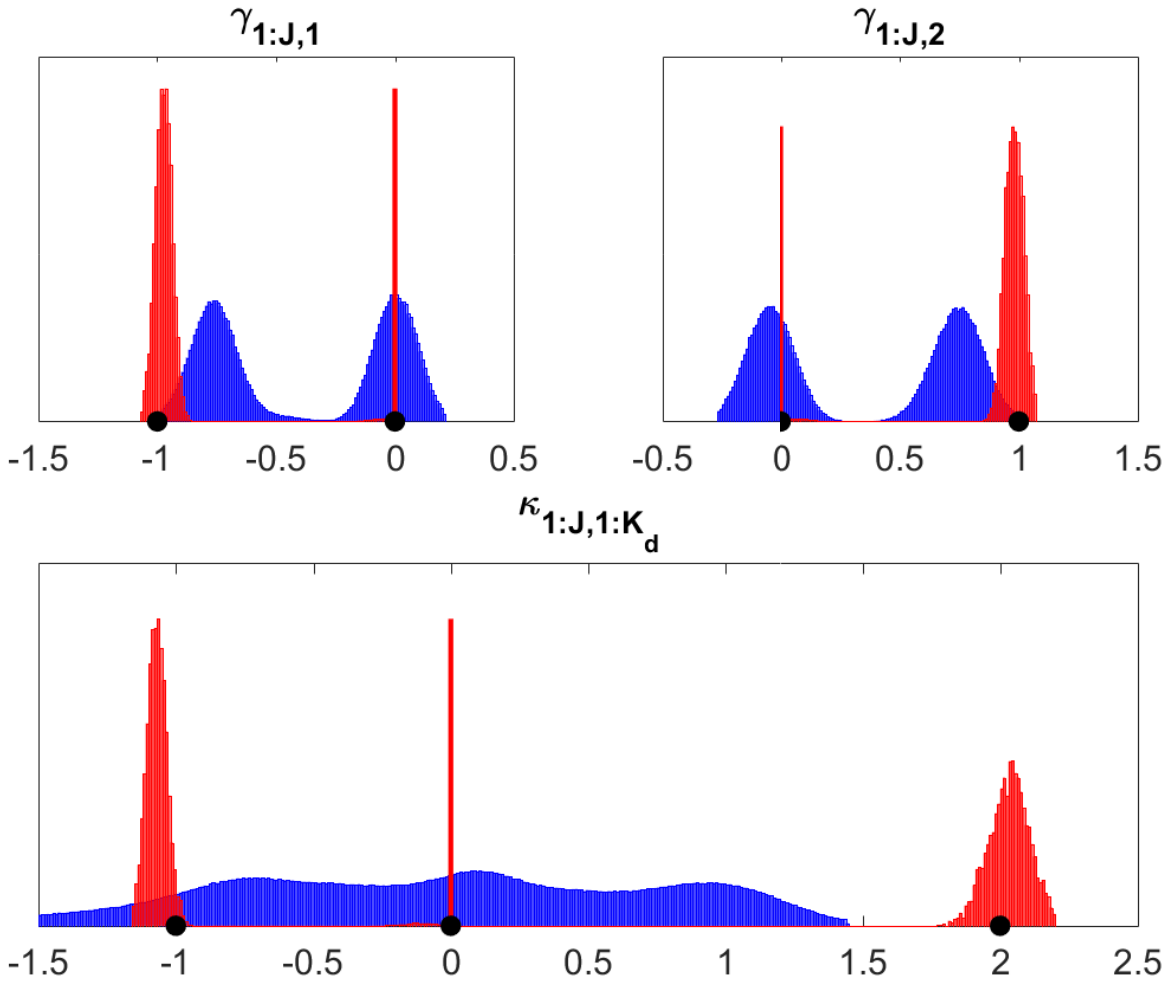

This figure shows the posterior probabilities that the outcome category at a specific row is in the same cluster as the outcome category at a specific column (left panel) and the posterior probabilities that explanatory categories at the rows and columns are in the same cluster (right panel) in the two-way clustering model. The posterior probabilities range from zero (light blue) to one (dark blue).

butions compared to the standard multinomial choice model. While the standard model separately estimates parameters for each outcome and explanatory category, the mixture model decreases parameter uncertainty by only estimating distinct parameter values per cluster.

Moreover, the posterior parameter distributions of the mixture model are centered around the parameter values in the data generating process, where the posterior modes of the standard model deviate from these values. The black dots in Figure 6.3 represent the parameter values in the data generating process. The standard multinomial choice model seems to be biased towards zero. The nonzero parameters in $\gamma$ are overestimated for negative values and underestimated for positive values. The standard model slightly overestimates the effect of the first five explanatory categories, at the expense of the estimated effects of the other explanatory categories in $\kappa_{jk}$. Note that outcome categories clustered with the first outcome category have parameter values exactly equal to zero, resulting to an accumulation of probability mass at zero in the posterior distributions of the mixture model. The standard multinomial choice model only sets the parameters of the first category exactly to zero.

The posterior parameter distribution diagnostics in Table 6.1 formalize the gains in estimation performance due to parameter clustering. The first panel shows performance measures for the mixture model and the fourth panel for the standard multinomial choice model,

**Figure 6.3:** Posterior parameter distributions two-way clustering




This figure shows the posterior parameter distributions of a standard multinomial probit model (fat, in blue) and the two-way clustering model (thin, in red). The upper left panel shows the parameter distributions of $\gamma_{j1}$ and the upper right panel of $\gamma_{j2}$, for $j = 1, \ldots, J$. The lower panel shows the parameter distributions of $\kappa_{jk}$ for $j = 1, \ldots, J$ and $k = 1, \ldots, K_d$. The black dots represent the parameter values in the data generating process in (6.36).

averaged over the outcome categories and explanatory categories per parameter cluster in the data generating process. The mixture model outperforms the benchmark for all diagnostics on each estimated parameter. The posterior parameter means of the mixture model are much closer to the parameter values in the data generating process, which is confirmed by the values of the mean squared error and the mean absolute error of the parameter draws. The relative interquartile range values show that the Dirichlet process prior is more efficient in exploiting sample information than the standard multinomial choice model.

Table 6.2 shows that two-way parameter clustering not only improves in accuracy and precision of the posterior parameter densities, but also yields higher hit rates and smaller root

**Table 6.1:** Diagnostics posterior parameter distributions

| | | $j = 1, \ldots, 25$ | | | | $j = 26, \ldots, 50$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | $\gamma_{j1}$ | $\gamma_{j2}$ | $\bar{\kappa}_{j,1:5}$ | $\bar{\kappa}_{j,6:10}$ | $\gamma_{j1}$ | $\gamma_{j2}$ | $\bar{\kappa}_{j,1:5}$ | $\bar{\kappa}_{j,6:10}$ |
| | DGP | 0.000 | 0.000 | 0.000 | 0.000 | -1.000 | 1.000 | -1.000 | 2.000 |
| two-way | Mean | -0.001 | 0.001 | -0.002 | -0.017 | -0.975 | 0.980 | -1.072 | 2.034 |
| | MSE | 0.000 | 0.000 | 0.000 | 0.024 | 0.002 | 0.002 | 0.007 | 0.007 |
| | MAE | 0.001 | 0.001 | 0.002 | 0.017 | 0.035 | 0.035 | 0.073 | 0.067 |
| | IQR | 0.001 | 0.001 | 0.003 | 0.031 | 0.047 | 0.054 | 0.049 | 0.100 |
| choice | Mean | -0.007 | -0.051 | -0.013 | -0.380 | -0.969 | 0.929 | -1.093 | 1.723 |
| | MSE | 0.001 | 0.004 | 0.016 | 0.351 | 0.003 | 0.007 | 0.033 | 0.155 |
| | MAE | 0.026 | 0.055 | 0.097 | 0.440 | 0.046 | 0.074 | 0.144 | 0.335 |
| | IQR | 0.042 | 0.045 | 0.126 | 0.392 | 0.056 | 0.063 | 0.162 | 0.357 |
| dummy | Mean | 0.038 | -0.073 | 0.102 | -0.818 | -0.757 | 0.750 | -0.820 | 0.958 |
| | MSE | 0.009 | 0.014 | 0.062 | 1.078 | 0.072 | 0.074 | 0.141 | 1.169 |
| | MAE | 0.077 | 0.097 | 0.175 | 0.847 | 0.244 | 0.250 | 0.276 | 1.042 |
| | IQR | 0.096 | 0.105 | 0.190 | 0.697 | 0.114 | 0.122 | 0.273 | 0.295 |
| standard | Mean | 0.003 | -0.048 | 0.211 | -0.790 | -0.757 | 0.745 | -0.703 | 0.903 |
| | MSE | 0.008 | 0.012 | 0.125 | 1.062 | 0.072 | 0.076 | 0.290 | 1.293 |
| | MAE | 0.072 | 0.088 | 0.281 | 0.830 | 0.243 | 0.256 | 0.440 | 1.097 |
| | IQR | 0.096 | 0.107 | 0.272 | 0.737 | 0.112 | 0.120 | 0.384 | 0.308 |

This table shows the performance measures for the parameters averaged over the outcome categories in the first cluster in the first four columns, and for the second cluster in the last four columns. The parameter draws for the dummies in the same cluster are averaged in $\bar{\kappa}_{j,1:5} = \frac{1}{5}\sum_{k=1}^{k=5}\kappa_{jk}$ and $\bar{\kappa}_{j,6:K_d} = \frac{1}{5}\sum_{k=6}^{k=K_d}\kappa_{jk}$. The first row shows the parameter values in the data generating process, the next panel the posterior mean, the mean squared error, the mean absolute error, and the interquartile range for the two-way clustering model. The next panels report these diagnostics for clustering over choice categories, clustering over dummy categories, and the standard multinomial probit model.

mean squared prediction errors of the choice category predictions. The first row shows hit rates and root mean squared prediction errors for the two-way mixture model. We compare these numbers to the performance of the standard multinomial choice model in the fourth row and a naive method in the fifth row. This naive method only uses the information in the dependent variable by calculating the category probabilities as the percentages observed in the data, and the category with the largest probability is always chosen. Out-of-sample the standard multinomial probit model and the naive data method are outperformed. In-sample, the standard model achieves a better hit rate, probably due to the large number of parameters. However, this gain in accuracy comes at the cost of efficiency. Therefore, the out-of-sample hit rate is slightly better for the mixture model. Based on the RMSPE there is no doubt

**Table 6.2:** Diagnostics in-sample and out-of-sample model fit

| | hit-rate | | | | RMSPE | | | |
|---|---|---|---|---|---|---|---|---|
| | category | | set | | category | | set | |
| | in | out | in | out | in | out | in | out |
| two-way | 0.035 | 0.035 | 0.881 | 0.871 | 0.003 | 0.003 | 0.022 | 0.024 |
| choice | 0.035 | 0.035 | 0.879 | 0.870 | 0.004 | 0.004 | 0.026 | 0.028 |
| dummy | 0.037 | 0.035 | 0.862 | 0.855 | 0.008 | 0.008 | 0.046 | 0.044 |
| standard | 0.038 | 0.035 | 0.857 | 0.851 | 0.009 | 0.009 | 0.049 | 0.047 |
| naive | 0.029 | 0.023 | 0.586 | 0.591 | 0.017 | 0.017 | 0.427 | 0.422 |

This table shows in-sample and out-of sample performance measured by hit rates and root mean squared prediction errors. The performance measures are reported for predicting actual category choices and predicting the right choice category cluster. The performance of the parameter clustering methods is compared to a standard multinomial probit model and a naive method in which the category probabilities are calculated as percentage observed in the data, and the category with the largest probability is always chosen.

which method performs best in estimating the category probabilities. The two-way mixture model outperforms the standard multinomial probit model and the naive data method by a large margin.

Since individuals have the same expected utility between alternatives within a choice cluster, it may be useful to target their preference set instead of only their actual choices. Although we cannot observe the actual outcome category clusters or preference sets in real data, simulated data allows us to study the power to distinguish these category parameter clusters from the data. When the model predicts a choice category which is in the same cluster as the actual choice in the data generating process, we count it as an correct prediction. Table 6.2 shows the hit rates and RMSPE for these set predictions. Just as for the category predictions, we find that model predictions improve on a naive method, and that parameter clustering outperforms a standard multinomial choice model. Two-way clustering correctly predicts out-of-sample choice sets in more than 87 percent of the cases.

The data generating process in (6.34) assumes a specific signal-to-noise ratio and distributes the observed choices relatively evenly over the choice alternatives. Moreover, by clustering over both the outcome and explanatory categories, the data generating process suits the two-way Dirichlet process mixture particularly well. We consider estimation results on simulated data from three different data generating processes, in which we only modify one aspect from the set-up in Section 6.4.1. To analyze the effect of prior beliefs

that are not in line with the data, we also keep the same prior settings. Detailed results are available upon request.

When all parameter values are unique while prior beliefs are in favour of parameter clustering, the two-way mixture model deteriorates in performance. Apart from the parameter vector corresponding to the first outcome category, which is set to zero for identification, all parameter values are independently and identically normally distributed with mean zero and variance 0.5. The signal-to-noise ratio decreases and the observed choices are irregularly distributed over the choice alternatives. The data generating process implies as many clusters as there are categories, both for the outcome and explanatory categories. The posterior mode for the number of distinct parameter values over the outcome categories is eight, which is in the tail of the prior distribution in Figure 6.1.

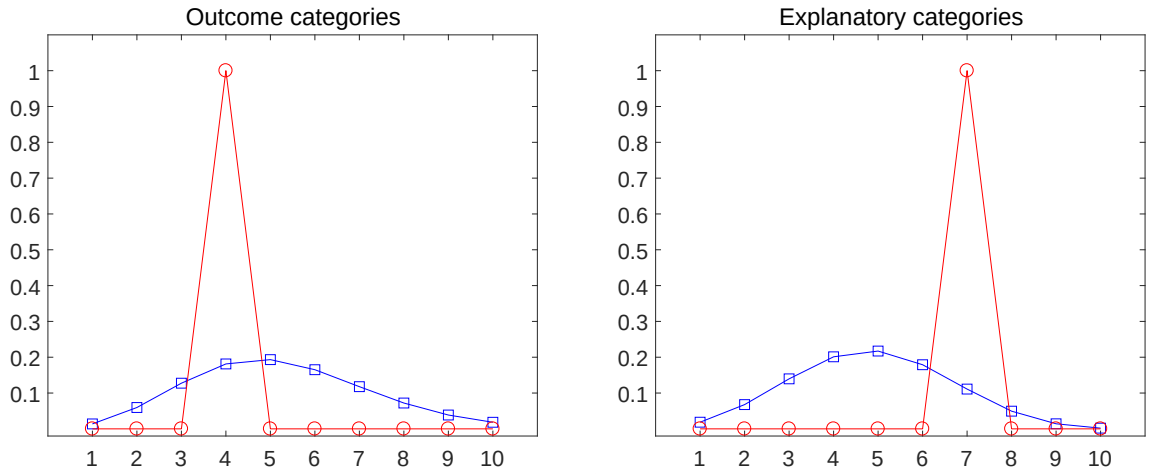
Parameter clustering over only one dimension is less harmful, but still affected by prior beliefs that encourage clustering over both dimensions. Again we draw independently and identically normally distributed parameters with mean zero and variance 0.5. However, we only sample one vector of parameter values per category cluster. The posterior mode for the number of distinct parameter values over the clustered dimension in the data generating process is equal to two. The posterior probabilities of the cluster memberships show an almost perfect cluster assignment. The posterior mode for the number of distinct parameter values over the non clustered dimension is located in the right tail of the prior distribution.

### 6.4.4 Parameter clustering over outcome categories

We have seen in Section 6.4.3 that the two-way mixture model improves upon a standard multinomial choice model for high-dimensional choice data for a wide range of different performance measures. To distinguish the gains from parameter clustering over outcome categories from parameter clustering over explanatory categories, this section estimates a model that only clusters over choice alternatives. The next section estimates a model that clusters explanatory dummy parameters. The models are estimated on the exact same simulated data as the two-way mixture model in Section 6.4.3.

The distribution over the number of parameter clusters for the outcome categories shifts to the right. The left panel in Figure 6.4 shows that the posterior distribution of the one-way mixture model is concentrated at four, while the two-way mixture model only puts positive

**Figure 6.4:** One-way: Distribution number of unique parameter values

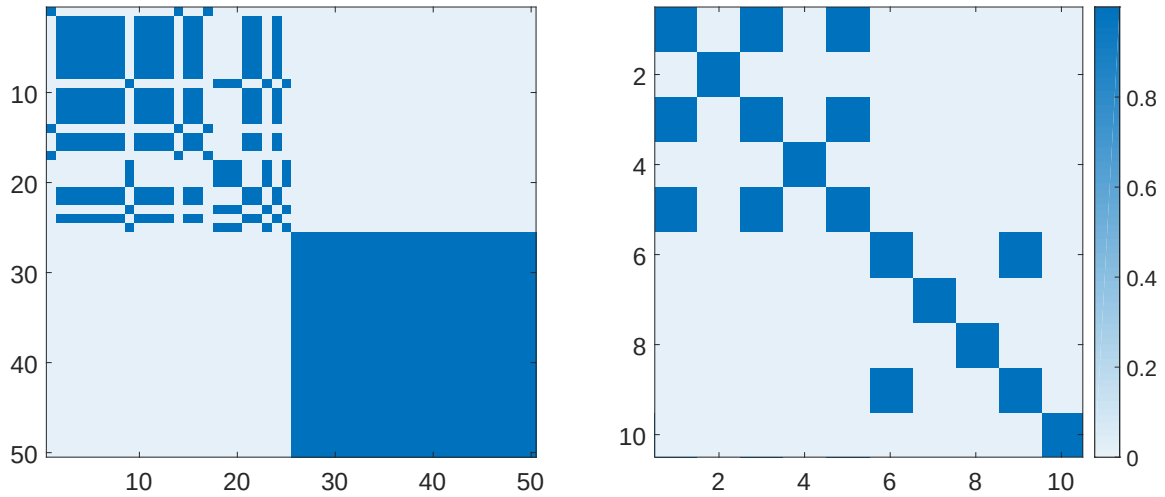



This figure shows the prior (□ in blue) and posterior (◯ in red) distribution over the number of unique parameter values $L^*$ over the outcome categories in the choice mixture model (left panel) and the explanatory categories in the dummy mixture model (right panel).

probability mass at two or three distinct unique parameter values over choice categories. The left panel in Figure 6.5 shows that the increase in clusters is utilized for estimating the parameter values for the first half of choice categories. Although the first 25 categories are never clustered together with a category with different parameter values in the data generating process, the clustering is not efficient in the sense that the model estimates three distinct unique parameter values for these categories while the actual parameter values are identical. The last 25 categories are again correctly mixed into one and the same cluster.

Although still slim compared to the standard multinomial probit model, Figure 6.6 exposes that the posterior parameter distributions of the outcome category mixture model gain up relative to the two-way mixture model. The dispersion of probability mass can be explained by the increase in parameter values to be estimated. The absence of a mixture over the explanatory categories increases the number of dummy parameters from two to ten. Since the outcome category parameters are distributed over more clusters in the one-way mixture model, the parameter uncertainty increases even further. The posterior probabilities of the cluster memberships in the left panel of Figure 6.5 show that the base category shares its parameter cluster with only two other choice categories. Therefore, probability mass is less concentrated at zero. We find that almost all posterior modes are still centered around the parameter values in the data generating process. However, the parameter draws for the last five explanatory categories seem to be biased towards zero.

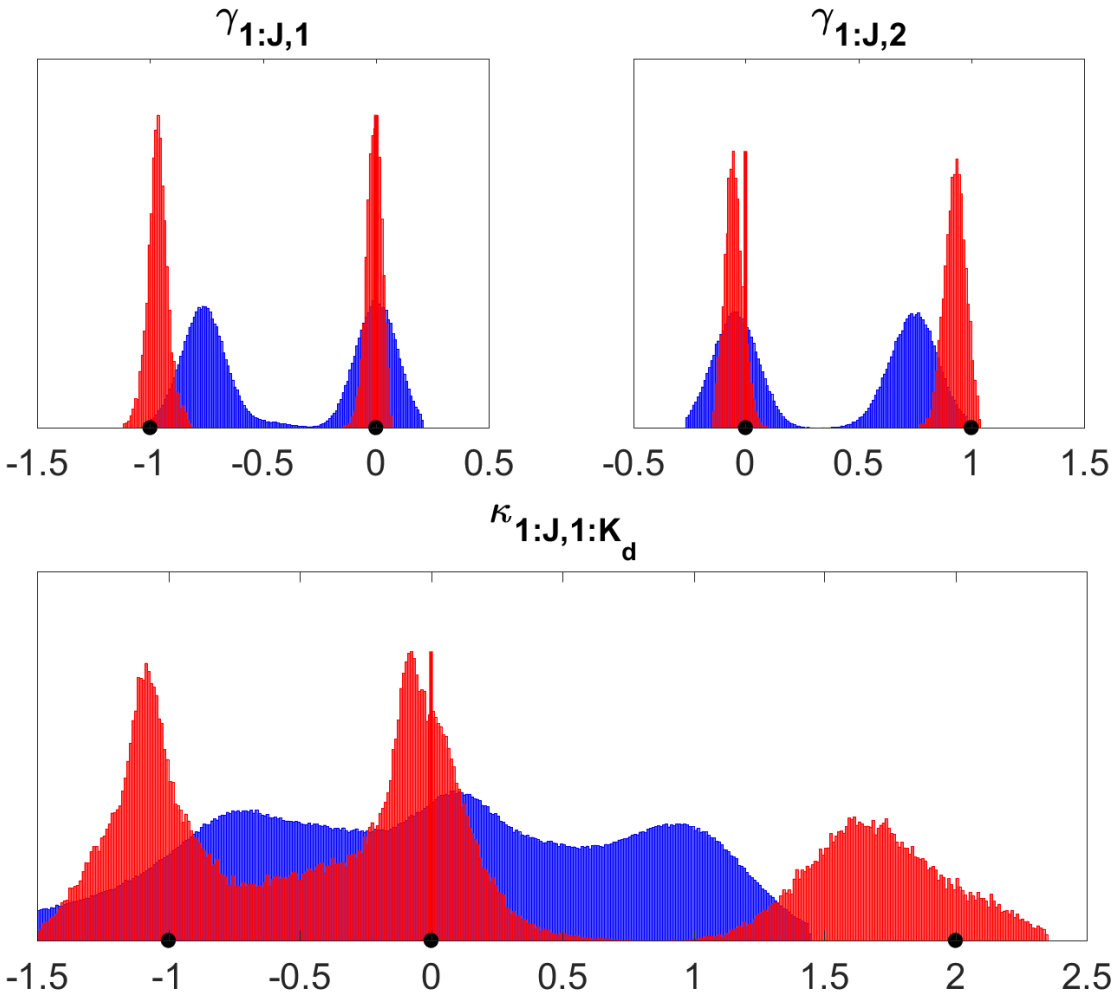
**Figure 6.5:** One-way: Posterior probabilities cluster memberships




This figure shows the posterior probabilities that the outcome category at a specific row is in the same cluster as the outcome category at a specific column in the outcome category parameter clustering model (left panel) and the posterior probabilities that explanatory categories at the rows and columns are in the same cluster in the explanatory category parameter clustering model (right panel). The posterior probabilities range from zero (light blue) to one (dark blue).

Diagnostics on the posterior parameter distributions and the overall model fit show that parameter clustering only over choice categories improves upon a standard multinomial choice model, but is not competitive to two-way parameter clustering. On all performance measures reported in Table 6.1 the choice category mixture model, in the second panel of this table, is outperformed by the two-way mixture model for each parameter. Although the posterior mean of the standard multinomial choice model is sometimes closer to the actual parameter value than the choice mixture model, the mean squared error and mean absolute error are always smaller for the latter. Apart from the last five explanatory categories, we can conclude the same for the interquartile range. The second row of Table 6.2 shows the hit rates and root mean squared prediction errors for the choice category mixture model. The hit rates of the choice category predictions of the two-way mixture model and the choice category mixture model are almost the same. A slight decrease in performance of one-way clustering is found on the set predictions, while still outperforming the standard multinomial choice model and the naive prediction method. This also holds for the root mean squared prediction errors for category and set predictions, for which only two-way parameter clustering decreases these measures relative to clustering parameters over choice categories.

**Figure 6.6:** Posterior parameter distributions outcome category clustering




This figure shows the posterior parameter distributions of a standard multinomial probit model (fat, in blue) and the choice category mixture model (thin, in red). For additional information, see the note following Figure 6.3.

### 6.4.5 Parameter clustering over explanatory categories

The performance of a one-way mixture model over explanatory categories deteriorates further relative to two-way clustering than the mixture over outcome categories. Estimating unique parameter values for each of the fifty outcome categories, instead of two, is a greater burden in terms of parameter uncertainty than estimating distinct parameter values for ten explanatory categories.

The one-way mixture over explanatory categories tends to perceive the increase in noise as parameter heterogeneity. The right panel of Figure 6.4 shows that the number of distinct parameter values for the dummy categories equals seven with posterior probability one,

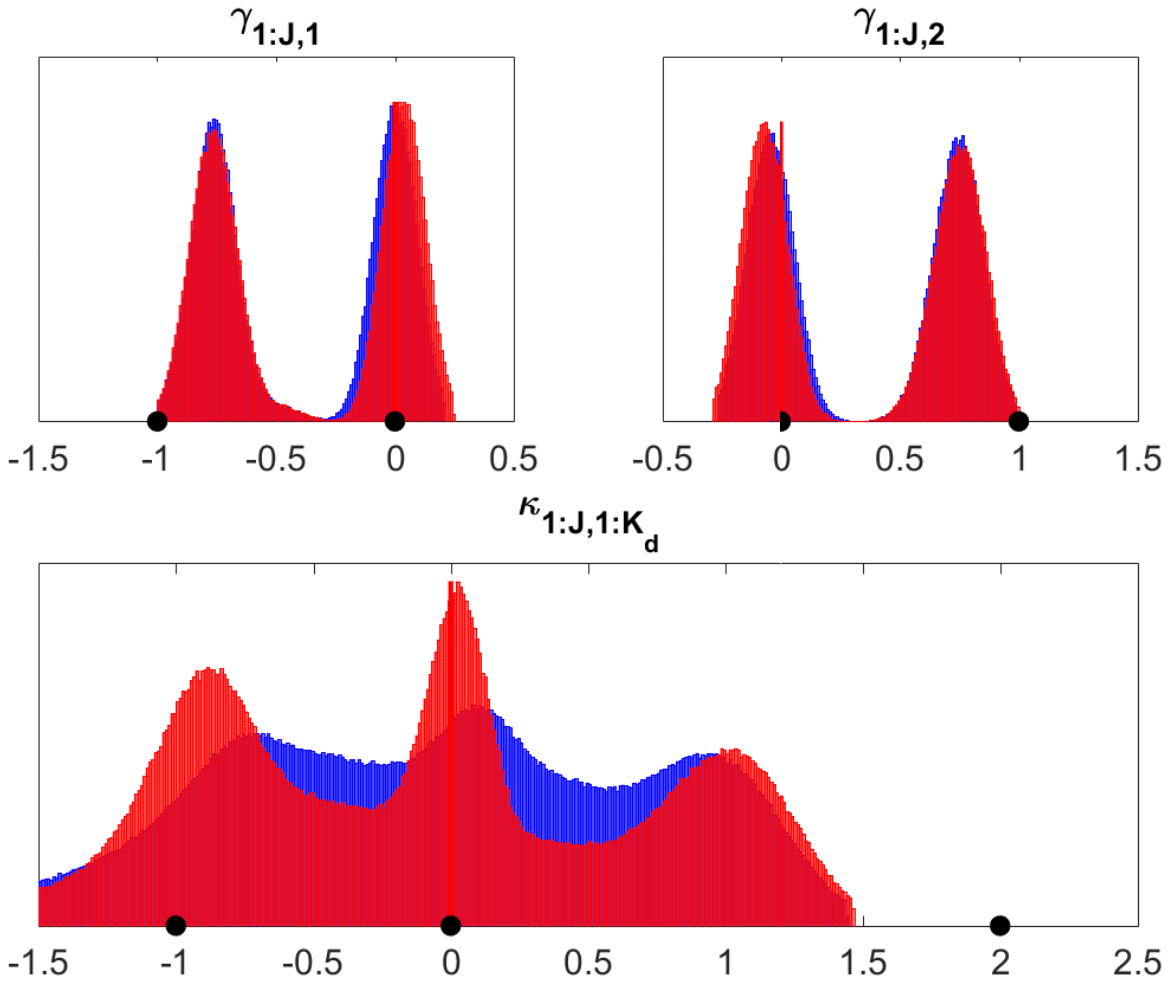
where the two-way mixture model concentrates the posterior probability mass at two. The right panel of Figure 6.5 shows that five categories have their own parameter cluster, while in the data generating process the category space is distributed over two equally sized clusters. The explanatory category mixture model only mixes the categories one, three, and five in a cluster, and category six and nine share also with posterior probability one a cluster. Although the categories are inefficiently assigned to clusters, the mixture model does not cluster categories with different parameter values in the data generating process.

The posterior parameter distributions of the dummy mixture model are almost indistinguishable from the posterior parameter distributions of the standard multinomial choice model. Figure 6.7 shows the probability mass of the parameters of the dummy variables, which is more concentrated around the posterior modes than for the standard multinomial choice model. However, the distributions are both biased towards zero and most of the probability mass is overlapping. This observation stands in stark contrast with the findings for the two-way mixture model and the one-way mixture over outcome categories in Figure 6.3 and 6.6, respectively. Since this simulation study has fifty choice categories and only ten dummy categories, separately estimating parameters for all choice categories evidently results in a greater accumulation of parameter uncertainty.

The performance measures on the posterior parameter distributions jointly agree that the two-way and choice category mixture models outperform the dummy mixture model. The diagnostics of the dummy mixture model in the third panel of Table 6.1 are close to the standard multinomial choice model in the fourth panel, and in some cases even worse. The third row and fourth row of Table 6.2 show that the dummy mixture model also approaches the performance of the standard multinomial choice model on hit rates. However, the model still outperforms the standard multinomial choice model in terms of root mean squared prediction error, for both category and set predictions, and both in-sample and out-of-sample.

To conclude, the two-way mixture model is most appropriate in this particular simulation study, which mimics the empirical study hereafter. We find that mixing over only one dimension yields substantially smaller gains over a standard multinomial choice model than clustering parameters over categories of choice alternatives and explanatory variables. However, the worse performance of one-way clustering relative to two-way seems to be caused by the high-dimensions in both directions. When the category space over which we do not cluster is large, the small signal-to-noise ratio prevents the mixture model to mix the other

**Figure 6.7:** Posterior parameter distributions explanatory category clustering




This figure shows the posterior parameter distributions of a standard multinomial probit model (fat, in blue) and the explanatory category mixture model (thin, in red). For additional information, see the note following Figure 6.3.

dimension well. We expect one-way clustering to perform well when the category space over which we do not cluster is small. Appendix 6.A confirms this conjecture by estimating the models on simulated data from alternative data generating processes.

## 6.5 Empirical Application

In this application we estimate the effect of household composition on holiday destinations. We acquired survey data from a Dutch market research company. The company observes that an important driver of holiday destination choice is the household composition. A single person under 35 and a family with three teenagers have different holiday preferences. Since

decision processes of households differ between short breaks and long vacations, we focus on destinations of holidays of more than seven days. The market research company is interested how preferences for these holidays differ across household types.

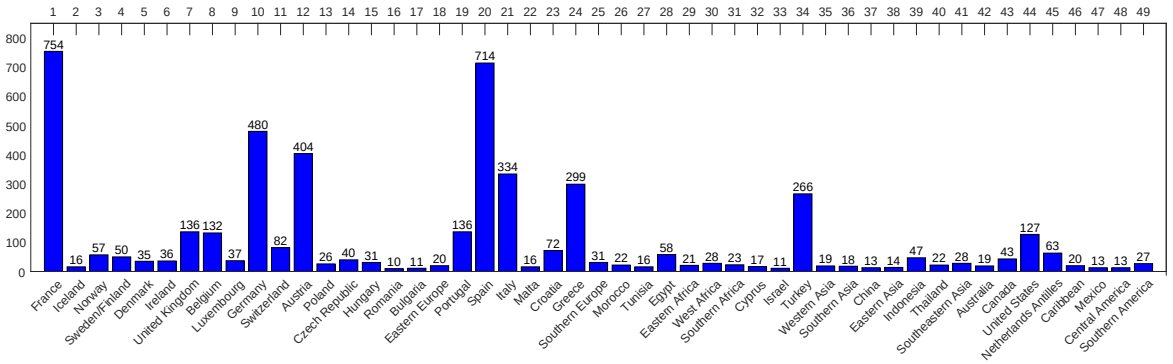
Due to the large number of choice categories and explanatory categories, this question is well-suited to be analyzed by the two-way mixture model proposed in this chapter. Depending on the level of detail of the geographical data, Dutch holidaymakers visit tens or hundreds of different holiday destinations each year. Although this application focuses on holiday destinations, we can also analyze holiday preferences over for example accommodations, transport, activities, or climate. Households can be grouped in different categories like, for instance, single households, couples with children of different age groups, and families of adults. Since the household composition is a categorical variable, it enters the analysis as a set of dummy variables. Estimating the effects of these household dummies on the holiday destination categories in a conventional choice model, results in a large number of parameter estimates. The large amount of parameter uncertainty and the large number of parameter estimates to be interpreted make it difficult to extract an answer to our research question.

### 6.5.1 Data

The data set consists of details of all reported holidays undertaken in 2015 by 6512 Dutch respondents and the individual characteristics of these respondents. Among other things, respondents were asked to which country or region they have been for holidays and for how long. We analyze the 4907 holidays with a foreign destination of more than seven days. Jointly analyzing the decision process for the 1881 domestic holidays and the 4907 foreign holidays asks for some kind of baseline inflated choice model, which is outside the scope of this chapter.

The respondents could select their foreign holiday destination from 77 categories in the survey, from which the market research company grouped countries of certain regions into one category. We group some categories in the survey answers again to end up with categories with a minimum of ten observations. Categories which are never chosen by respondents are deleted. Appendix 6.B shows the countries per holiday destination choice category. We set the most frequent chosen holiday destination, which is France, as the base category. Figure 6.8 shows the frequency counts for the 49 categories in the resulting dependent vari-

**Figure 6.8:** Frequency counts choice categories



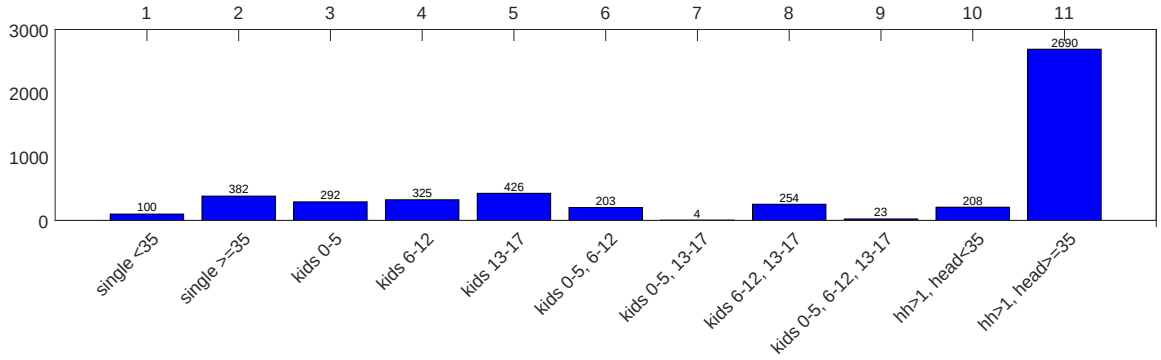

This figure shows the frequency counts for the categorical dependent variable. The categories represent destinations of foreign holidays of more than seven days of Dutch households.

able. The base category France is chosen 754 times, while Romania got only 10 visits of the respondents. The median number of observations within a category is 31 and the mean equals 100. The overall pattern in the survey answers is representative for the Dutch holiday market. France is the most popular destination, more generally Europe, and Turkey, Egypt, and the United States are favorites outside of Europe.

The survey asked the respondents to select their household composition out of eleven categories. The first two categories distinguish singles under 35 from singles above 35. The third till ninth category describe households with children. Kids are divided among the age groups 0-5, 6-12, and 13-17, and four categories describe all possible combinations of these age groups in a family. The final two categories contain households of two or more persons in which everyone is 18 years or older, with the head of the household under 35 or older than 35. Figure 6.9 shows the frequency counts for the dummy categories. Most of the households belong to the last category; 2690 out of the 4907 holidays are undertaken by households consisting of two or more persons of 18 years or older, with the head of household older than 35. Only four holidays are reported by families who have only children between the age of 0-5 and in the age group 13-17. The median number of holidays within a dummy category is 254 and the mean equals 446.

In addition to the set of household composition dummy variables, we have ten control variables. We control for the income of the household, which is measured as a categorical variable, by the standardized logarithm of the maximum of the income category of the household in a continuous variable. A dummy variable corrects for respondents who do not want to say or do not know their income. The set of controls is completed by dummies indicating

**Figure 6.9:** Frequency counts household categories




This figure shows the frequency counts for the categorical explanatory variable. The categories represent the household compositions of the survey respondents for each holiday.

respondents who are retired, are student, own a moving holiday accommodation, own a fixed holiday accommodation, and are in a specific social class. Moving holiday accommodations include tents, caravans, campers, and cabin boats. Fixed holiday accommodations are defined as holiday homes or a mobile home with a fixed location. The sample is divided in five social classes, captured by four dummy variables. Appendix 6.C explains the control variables in more detail and provides descriptive statistics.

Although there are reasons to suspect endogeneity in household decompositions, we do not believe this to be an issue in this application. Since the survey about holiday destinations in 2015 was conducted in 2016, reverse causality between household composition and holiday destination is highly unlikely. However, an omitted variable as "being adventurous" may affect both the preference for living together and the preference for holiday destination. Although we do not provide evidence that formally modeling endogeneity does not affect the results, we believe that the comprehensive set of controls addresses this issue.

### 6.5.2 Modelling choices

We estimate the parameters $\beta$ in the multinomial probit model defined in (6.1) and (6.2) on the first 4000 holidays and use the remaining 907 holidays for out-of-sample analysis. In the standard multinomial probit model, all parameters in the $J \times K$ matrix $\beta$ are unique, which amounts to $49 \times (10 + 11) = 1029$ parameters with only 4000 observations on a nominal scale. To decrease parameter uncertainty and increase interpretability of the results, we cluster over both dimensions of the parameter matrix $\beta$ in the two-way mixture model (6.17). The parameters are sampled as discussed in Section 6.3.

The truncation level of the number of potential choice category clusters is set equal to $L_J = 25$. Since the number of choice categories is $J = 49 > L_J$, we are charged with an approximation error. The expectation and the variance of the aggregated higher order probabilities equal $0.038$ and $6.926 \times 10^{-4}$ for the sampled $\alpha_J$ in the last iteration of the Gibbs sampler. We do not truncate the number of potential dummy category clusters, $L_D = K_d = 11$, which means that we specify a full Dirichlet process for the explanatory categories. Therefore, we conclude that the two-way mixture approximation error is sufficiently small in the posterior simulation.

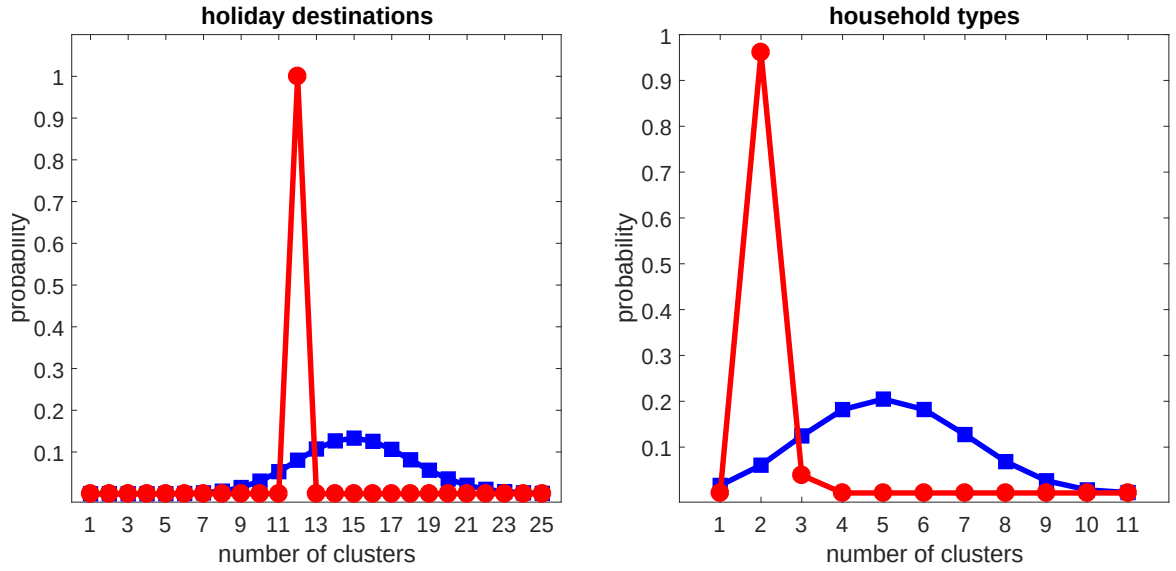
We follow Section 6.3.2 in choosing the parameter values in the prior distributions for the concentration parameters. Our prior belief about the mode of unique parameter values over holiday destinations is 15 and over household compositions 5. The prior distributions that match these beliefs are $\alpha_J \sim \text{Gamma}(7.15 \times 20, 20)$ with $\text{var}(L_J^*) = 8.852$ and $\alpha_D \sim \text{Gamma}(3.47 \times 1, 1)$ with $\text{var}(L_D^*) = 3.352$. Just as in the simulation study in Section 6.4, we allow for a wide range of plausible values for the model parameters within a multinomial choice model, and set the prior variance of the model parameters equal to $\sigma_\beta^2 = 1$, .

Posterior results are based on 100,000 iterations of the Gibbs sampler, from which the first 50,000 are discarded, and we use a thinning value of 10. Appendix 6.D shows by means of convergence diagnostics that this number of retained draws is sufficient for posterior inference.

### 6.5.3 Results

Figure 6.10 shows that the two-way mixture model substantially reduces the dimensions of both the choice categories and the explanatory categories. Instead of estimating unique parameter values over 49 holiday destinations and 11 household compositions, the Dirichlet process prior clusters the categories to a maximum of respectively twelve or three unique parameter values. The left panel of Figure 6.10 shows that after convergence all posterior probability mass is concentrated at twelve clusters of holiday destinations. The posterior mode in the right panel is located at two clusters of households compositions, with the remaining 4 percent of probability mass at three clusters. For both dimensions, a large variance of the prior distributions on the concentration parameters is employed, and the posterior distribution over the number of clusters makes a considerable move to the left relative to the

**Figure 6.10:** Application: Distribution number of unique parameter values



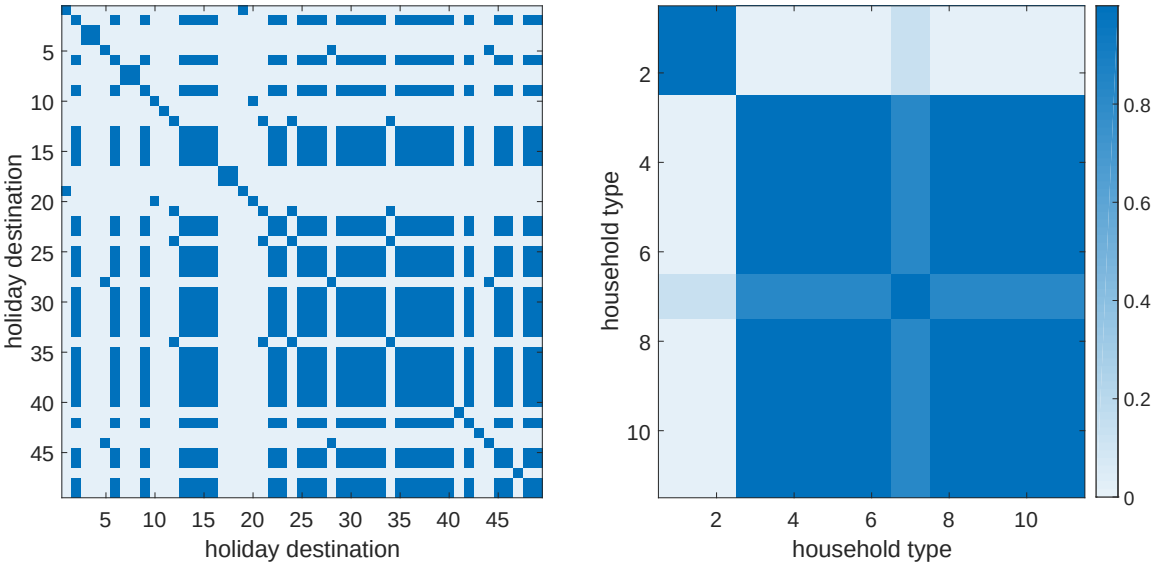

This figure shows the prior (□ in blue) and posterior (◯ in red) distribution over the number of unique parameter values $L^*$ over the holiday destinations (left panel) and the household compositions (right panel).

prior distributions. This observation suggests that the shift to a more parsimonious model is driven by the data, instead of a prior specification encouraging a small number of clusters.

Figure 6.11 shows which holiday destinations and which household compositions tend to cluster together. The left panel of Figure 6.11 shows that most holiday destinations share a cluster with multiple other destinations. Moreover, the posterior probabilities of cluster memberships of all holiday destinations converge to zero or one. The right panel shows the cluster assignment of the household composition dummies. We find that the single households, in the first and second household category, share a cluster with each other. The posterior probability is fourteen percent that this cluster also includes the seventh explanatory category. This seventh category shares in 82 percent of the sample iterations its cluster with the remaining categories, from three to eleven, and has in five percent its own cluster. However, the seventh category contains households with kids aged between zero and five and kids between 13-17 and we observe only four holidays for this household category. We conclude that single households have different holiday preferences than households of two or more persons.

Since the posterior probabilities of cluster memberships of all holiday destinations converge to zero or one, we are able to circumvent the label-switching problem and infer which destinations share the same clusters. The left panel of Figure 6.11 shows that there are four

**Figure 6.11:** Application: Posterior probabilities cluster memberships

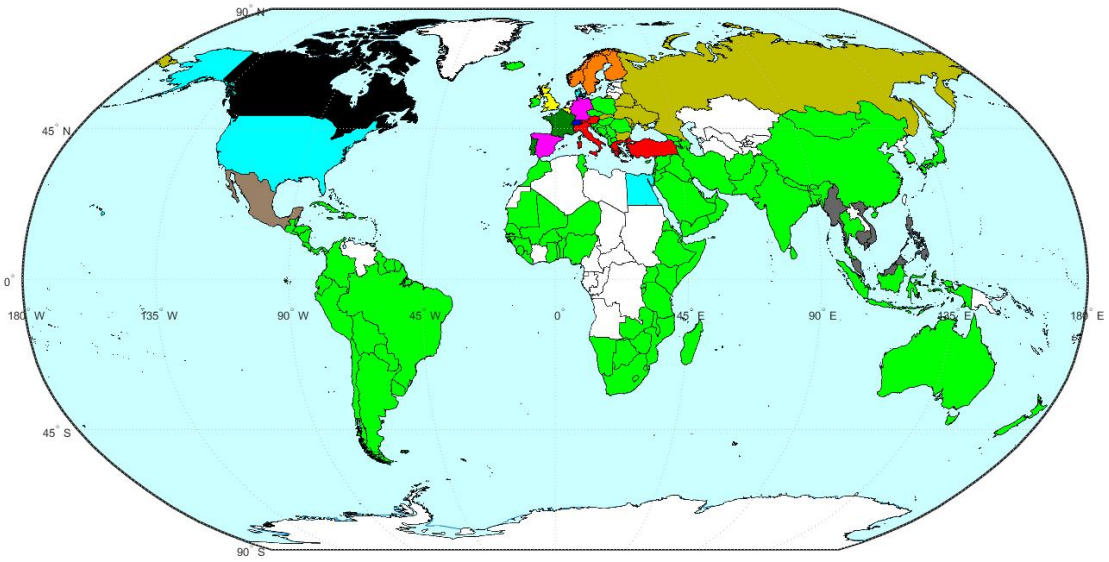



This figure shows the posterior probabilities that the holiday destination at a specific row is in the same cluster as the holiday destination at a specific column (left panel) and the posterior probabilities that household compositions at the rows and columns are in the same cluster (right panel) in the two-way mixture model. The posterior probabilities range from zero (light blue) to one (dark blue).

categories with their own cluster. That is, for category 11, 41, 43, and 47 no other category than itself has a positive posterior probability of being a cluster member. Figure 6.8 shows that these categories correspond to Switzerland, Southeastern Asia, Canada, and Mexico, respectively. Furthermore, we find five clusters with only two holiday destinations, one clustering of three, and one clustering of four. Finally, one large cluster is shared by no less than the 28 remaining choice categories.

Figure 6.12 shows the parameter clustering over the world's holiday destinations of Dutch households, according to the clustering of the choice categories in the left panel of Figure 6.11. The sets of destinations with the same color have the same conditional probability of being chosen by a household. Conditional on the household characteristics, the households have an expected preference ranking across countries with different colors, but the expected utility is identical between countries with the same color. To illustrate, households have the same probability of going to Greece or Turkey, which are in the same cluster. The probability differs between Canada and the United States, which are in different clusters. The estimated parameter clustering is very different from an ad hoc grouping based on, for instance, geographical location. Only in Europe, we already find nine different clusters.

**Figure 6.12:** Application: Clustering holiday destinations

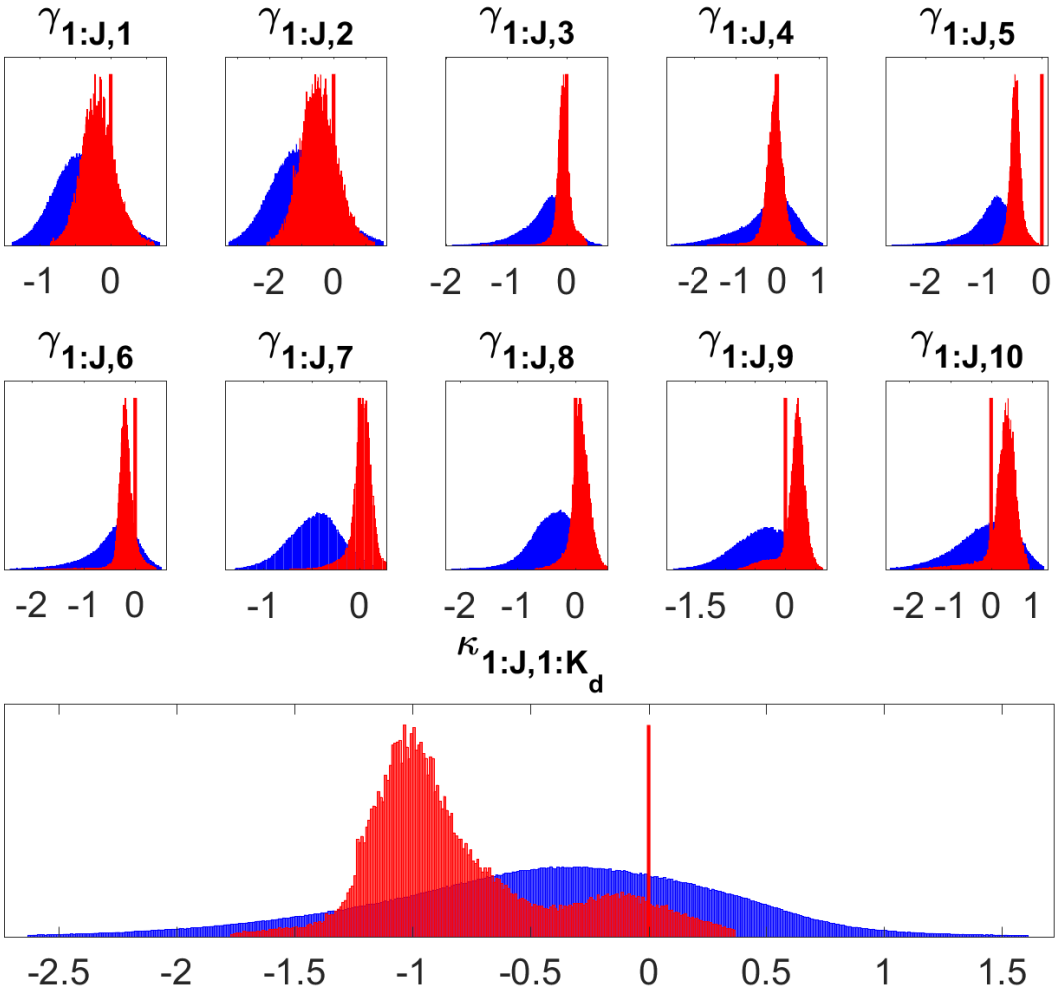



This figure shows the cluster assignments of the holiday destinations of Dutch households. Destinations with the same color are in the same parameter cluster according to Figure 6.11, and we do not have observations about white regions. The two-way mixture model estimates twelve clusters. Appendix 6.B shows the countries in each of the 49 holiday destination choice categories.

When the explanatory variables have modest predictive power, the conditional clustering is driven by base preferences. This may explain that the clustering in Figure 6.12 is strongly correlated with the observed number of visits per destination category. Frequently visited countries tend to cluster together. Popular destinations within Europe share clusters, such as Germany and Spain. The red cluster includes other popular destinations of Dutch travelers; Austria, Italy, Greece, and Turkey. There is one large cluster which includes Southern America, a large part of Africa, Eastern Europe and Asia.

Comparing the posterior densities of the standard multinomial probit model with the two-way mixture model, we find the densities of the latter model to be more precise. Figure 6.13 shows the posterior parameter distributions of the explanatory variables over all holiday destinations. The first two rows of panels show the parameter distributions of the control variables, and the last row shows the parameter distributions over all explanatory categories in one window. The mixture model accounts for the uncertainty about the number of clusters, the cluster assignments, and parameter uncertainty. However, sampling separate parameter values for each destination in the standard multinomial choice model results in much more

**Figure 6.13:** Application: Posterior parameter distributions

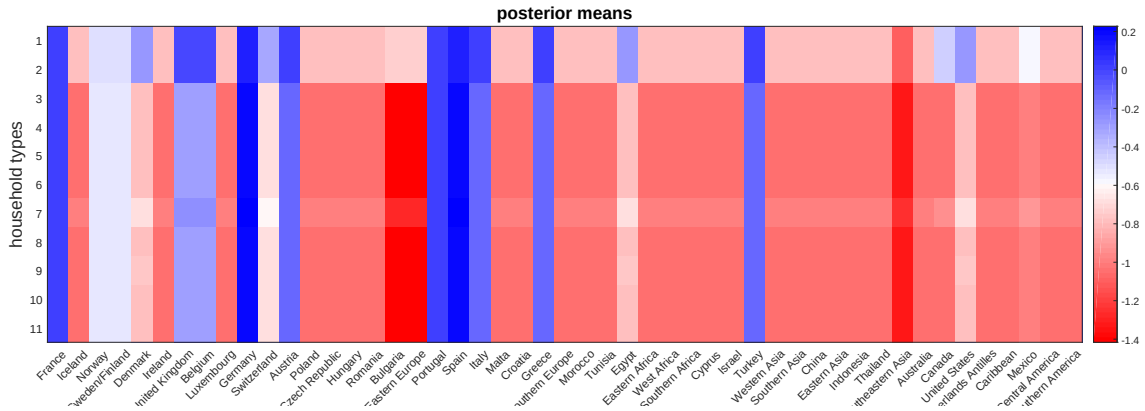



This figure shows the posterior parameter distributions of a standard multinomial probit model (fat in blue) and the two-way mixture model (thin in red). The first two rows show the $K_w$ parameter distributions of the control variables $\gamma_{jk}$, for $j = 1, \ldots, J$. The third row shows the parameter distributions of $\kappa_{jk}$ for $j = 1, \ldots, J$ and $k = 1, \ldots, K_d$.

noise. The shapes of the posterior parameter distributions show, except for width, also other differences. Since the mixture model clusters the base category destination France with other destinations, more probability mass is allocated to zero. The posterior distributions of the standard multinomial choice model over all choice categories approximate for most parameters a bell shape. Due to the mixing of parameters over different holiday destinations, the mixture model distributions are more often skewed and show for several parameters multiple modes.

Figure 6.13 shows the posterior parameter distributions of all choice categories and dummy categories altogether. However, the differences between the posterior parameter means over choice and explanatory categories provide insights in the estimated effects of the

**Figure 6.14:** Application: Posterior parameter means for all variables categories




The upper panel of this figure shows in each row the posterior parameter mean for a control variable over all choice categories in Figure 6.8. Descriptions of the ten control variables indicated by their index can be found in Appendix 6.C. The lower panel shows the posterior parameter mean for each household composition category, as showed in Figure 6.9, over all choice categories. The values of the posterior means are indicated by the color bars with colors ranging from dark red (strongly negative) to dark blue (strongly positive).

variables on the holiday destination choice. Figure 6.14 shows these posterior means for each parameter separately. We illustrate the interpretation of the posterior parameter means for a set of control variables. The third variable controls for retired respondents, and has a positive effect on going to Germany and Spain and a strong negative effect on Mexico. Students are not eager to travel to Switzerland but are more inclined to go to Eastern Europe. The fifth variable controls for households with a moving holiday accommodation. Not surprisingly, these households have a higher probability of staying close to home and travel to France. Households with fixed holiday accommodations have their holiday relatively more often in Switzerland.

Since each household category has its own dummy variable, the household composition parameters can be interpreted as the base preferences of the households relative to the base category. Figure 6.14 shows strong negative base preferences of the non single households, household category three till eleven, for Eastern Europe (holiday destination 17 and 18) and Southeastern Asia (41). On the other hand, these households are strongly inclined to travel to Germany or Spain (10 and 20). Relative to the other household compositions, the singles in the first two household categories have small estimated parameters for Germany and Spain, but relatively higher posterior parameter means for all other holiday destinations. Based on the base preferences, singles are less inclined to visit Germany and Spain, and the countries

**Table 6.3:** Application: Hit rates choice category predictions

| two-way | | standard | | naive | |
|---|---|---|---|---|---|
| in | out | in | out | in | out |
| 0.072 | 0.071 | 0.086 | 0.082 | 0.149 | 0.175 |

This table shows in-sample and out-of sample performance for predicting actual category choices measured by hit rates. The performance of the two-way mixture model is compared to a standard multinomial probit model and a naive method in which the category probabilities are calculated as percentage observed in the data, and the category with the largest probability is always chosen.

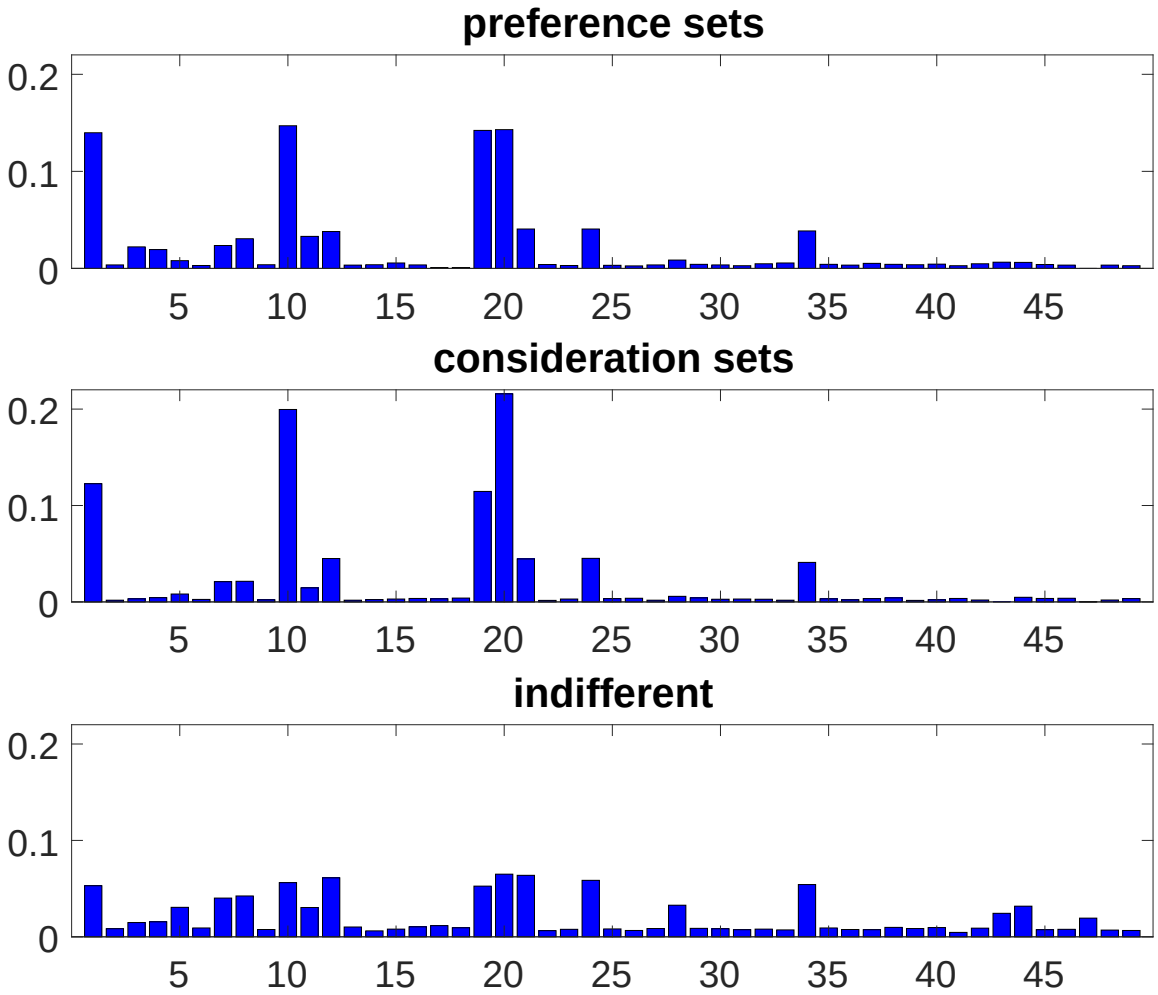
in the base category, France and Portugal, but more adventurous to explore countries further away from home.

Figure 6.13 suggest that the two-way mixture model is more efficient in estimating the model parameters than the standard multinomial choice model. However, Table 6.3 shows that the hit rates of the latter outperform the mixture model. Moreover, the naive prediction method performs best due to the skewed distribution of observations over the choice categories. The simulation study in Section 6.4 shows that the two-way mixture model does not so much improve upon hit rates of category predictions, but shows excellent performance on hit rates of preference sets and root mean squared prediction errors. Unfortunately, both of these measures can only be evaluated on simulated data.

The implied choice category probabilities by the two-way mixture model are conditional on the household characteristics and allow for a wide variety of different preference sets. Figure 6.15 shows the preference sets of three households with different characteristics in the sample. The implied posterior choice probabilities are calculated as the percentage predictions for each category over the sample iterations, which means that choice probabilities for categories with identical parameter values can be slightly different.

The first panel illustrates a typical preference set. The respondent is retired, owns a moving holiday accommodation, and lives in a household of two or more persons, with the head of the household older than 35, in the upper class. The household is in expectation indifferent within three sets of holiday destinations, while there is a preference relation across the sets. The categories 1, 10, 19, 20 are equally preferred over all other destinations. This household prefers the usual Dutch holiday destinations France, Germany, Portugal, and Spain. Next, the household has a second preference set with other conventional destinations of Dutch hol-

**Figure 6.15:** Application: Posterior holiday probabilities




This figure shows the implied posterior choice probabilities for each holiday destination in Figure 6.8. The posterior probabilities are calculated as the percentage predictions for each category over the sample iterations.

idaymakers in Europe. The last preference set, which contains mostly countries in Eastern Europe and outside Europe, has a probability of being chosen very close to zero.

The concept of preference sets also allows for consideration sets. The second panel of Figure 6.15 shows the implied choice probabilities of a retired respondent, with both moving and fixed holiday accommodations, in a household of two or more persons in the middle class. This household seems to construct a consideration set in the decision-making process for holiday destinations. The probabilities of almost every holiday destination is almost zero and not considered as a serious option. The remaining holiday destinations vary in their choice probabilities. The third panel shows the preferences of an upper class single

household under 35. Since the estimated choice probabilities are very close to each other, we conclude that this household is almost indifferent between all destinations.

## 6.6 Conclusion

With choice data, the number of model parameters typically becomes large. Categorical characteristics of the decision makers enter the model as sets of dummy variables, in which each variable has its own choice alternative specific parameter. The two-way Dirichlet process mixture model clusters parameters over the choice categories and the explanatory dummy categories, while taking the relation between the dependent and independent variables into account. The parameter clusters distinguish which categories have a different effect, and the clustered outcome categories also have an interpretation similar to consideration sets.

We find on simulated data that the mixture model substantially reduces the number of estimated parameters relative to a standard multinomial choice model. This increase in parsimony results in an improvement over a range of performance measures on posterior parameter distributions, in-sample fit, and out-of-sample predictions.

In the high-dimensional empirical application, we examine how preferences for holiday destinations differ across household types. The mixture model, with potentially more than a thousand parameters, provides clear insights in the holiday choice behavior. On average, we find that singles are more inclined to visit holiday destinations far away from home than households of two or more persons. On household level, we use the choice probabilities implied by the posterior parameter distributions to distinguish households who form a consideration set, several preference sets, or are more or less indifferent over the choice set.

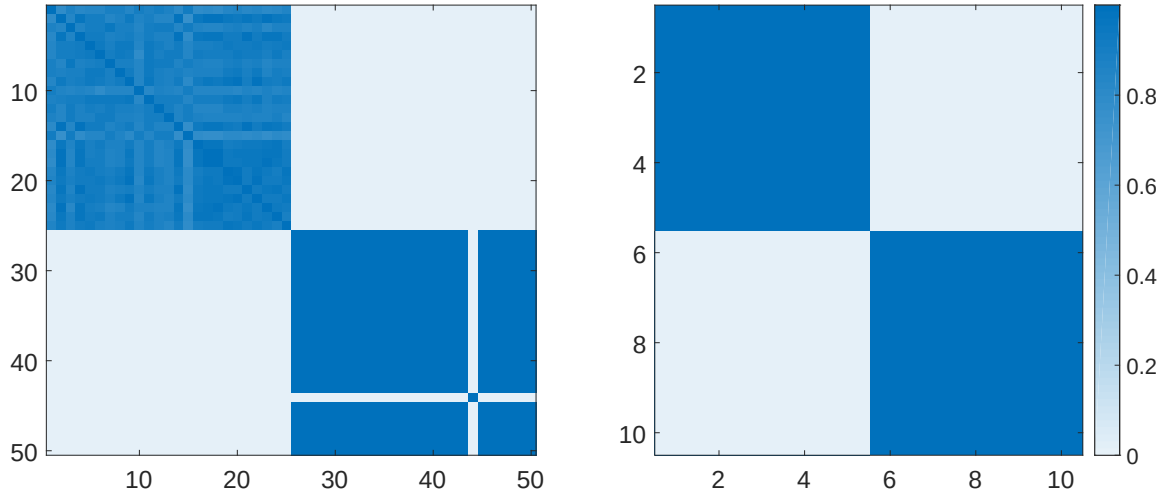
## 6.A Additional simulation studies

This appendix shows that the worse performance of one-mode clustering relative to two-mode seem to be caused by the large number of categories in both the choice set and the explanatory variables. When the category space over which we do not cluster is small, one-way clustering substantially increases in performance.

First, we draw from the same data generating process (6.34) as in Section 6.4, but we decrease the number of explanatory categories by setting $K_d = 4$. Subsequently, we estimate

the parameters in the one-way clustering model over choice alternatives. The left panel of Figure A1 shows that the posterior clustering is closer to the data generating process than we found in the left panel of Figure 6.5, where the dimension over which we are not clustering, $K_d = 10$, is relatively large.

**Figure A1:** One-way: Posterior probabilities cluster memberships




This figure shows the posterior probabilities that the outcome category at a specific row is in the same cluster as the outcome category at a specific column in the outcome category clustering model (left panel) and the posterior probabilities that explanatory categories at the rows and columns are in the same cluster in the explanatory category clustering model (right panel). The posterior probabilities range from zero (light blue) to one (dark blue).

Second, we again draw from the same data generating process (6.34), but now we decrease the number of choice alternatives to $J = 10$. Parameter estimation in the one-way clustering model over the explanatory categories results in a posterior clustering as in the right panel of Figure A1. The posterior clustering is identical to what is expected from the data generating process. This is a substantial improvement relative to the inefficient cluster assignment in the left panel of Figure 6.5, where the model clusters over 10 explanatory categories while $J = 50$.

## 6.B Application: Holiday destinations

This appendix shows the countries within each holiday destination choice category in Figure 6.8 in Section 6.5.

| | | | |
|---|---|---|---|
| **Eastern Europe** | Benin | **Southern Asia** | Haiti |
| Belarus | Burkina Faso | Afghanistan | Jamaica |
| Moldova | Cape Verde | Bangladesh | Martinique |
| Ukraine | Cote dIvoire | Bhutan | Montserrat |
| Slovakia | Ghana | Iran | Puerto Rico |
| Russia | Guinea | Maldives | Saint Barthelemy |
| **Southern Europe** | Guinea-Bissau | Nepal | Saint Kitts and Nevis |
| Slovenia | Liberia | Pakistan | Saint Lucia |
| Albania | Mali | India | Saint Martin |
| Bosnia and Herzegovina | Mauritania | Sri Lanka | Saint Vincent and the Grenadines |
| Gibraltar | Niger | **Eastern Asia** | Trinidad and Tobago |
| Vatican City | Nigeria | Hong Kong | Turks and Caicos Islands |
| Montenegro | Saint Helena | Japan | United States Virgin Islands |
| San Marino | Senegal | Korea | **Central America** |
| Serbia | Sierra Leone | Macau | Belize |
| Macedonia | Togo | Mongolia | Costa Rica |
| **Eastern Africa** | **Southern Africa** | **Southeastern Asia** | El Salvador |
| Kenya | South Africa | Brunei | Guatemala |
| Burundi | Botswana | Burma | Honduras |
| Comoros | Lesotho | Cambodia | Mexico |
| Djibouti | Namibia | Laos | Nicaragua |
| Eritrea | Swaziland | Philippines | Panama |
| Ethiopia | **Western Asia** | Singapore | **Southern America** |
| Madagascar | Jordan | Timor-Leste | Brazil |
| Malawi | Armenia | Viet Nam | Argentina |
| Mauritius | Azerbaijan | Malaysia | Bolivia |
| Mayotte | Bahrain | **Caribbean** | Chile |
| Mozambique | Georgia | Anguilla | Colombia |
| Reunion | Iraq | Antigua and Barbuda | Ecuador |
| Rwanda | Kuwait | Aruba | Falkland Islands |
| Seychelles | Lebanon | Bahamas | French Guiana |
| Somalia | Oman | Barbados | Guyana |
| Uganda | Palestine | British Virgin Islands | Paraguay |
| Tanzania | Qatar | Cayman Islands | Peru |
| Zambia | Saudi Arabia | Cuba | Suriname |
| Zimbabwe | Syrian | Dominica | Uruguay |
| **West Africa** | United Arab Emirates | Grenada | |
| Gambia | Yemen | Guadeloupe | |

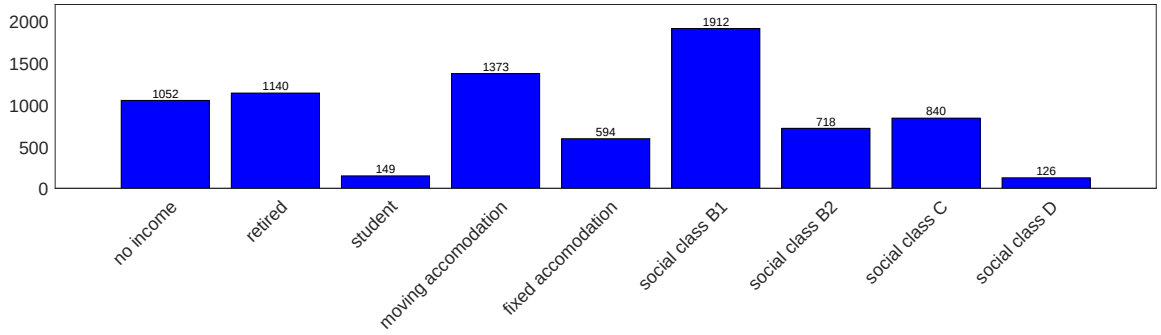
# 6.C Application: Control variables

**Table C1:** Gross annual income of household categories

| | | | |
|---|---|---|---|
| $<$ 4.600 | 14.300 - 15.400 | 38.800 - 51.300 | 181.300 - 206.400 |
| 4.600 - 6.300 | 15.400 - 17.100 | 51.300 - 65.000 | 206.400 - 232.600 |
| 6.300 - 8.000 | 17.100 - 20.000 | 65.000 - 77.500 | 232.600 - 258.900 |
| 8.000 - 9.100 | 20.000 - 23.400 | 77.500 - 103.800 | 258.900 - 284.500 |
| 9.100 - 10.800 | 23.400 - 26.200 | 103.800 - 129.400 | 284.500 - 310.700 |
| 10.800 - 12.500 | 26.200 - 32.500 | 129.400 - 155.100 | 310.700 $<$ |
| 12.500 - 14.300 | 32.500 - 38.800 | 155.100 - 181.300 | no response |

This table shows the 28 categories of gross annual income of a household. The last category, no response, includes the households which do not know or do not want to say what their income is. The income categories are included in the models as the standardized log mean of each income group. We correct for the no responses by including a dummy in the model.

**Figure C1:** Frequency counts dummy control variables




This figure shows the frequency counts for the explanatory control variables. The last four variables classify the Dutch households according to social class. The upper social class A is the reference category, B and C represent the middle class, and D is the lower social class.

# 6.D Application: Sampler convergence

**Table D1:** Summary of simulation convergence tests and inefficiency factors

| | Convergence test | | | Inefficiency factors | | |
|---|---|---|---|---|---|---|
| | 10% | 5% | 1% | Mean | Min | Max |
| control variables | | | | | | |
| income | 0.021 | 0.021 | 0.000 | 4.160 | 0.918 | 5.833 |
| income dummy | 0.000 | 0.000 | 0.000 | 3.635 | 0.750 | 5.158 |
| retired | 0.896 | 0.896 | 0.875 | 22.282 | 6.048 | 45.806 |
| student | 0.021 | 0.021 | 0.000 | 13.741 | 2.482 | 17.627 |
| moving accomodation | 0.063 | 0.000 | 0.000 | 16.500 | 5.249 | 48.434 |
| fixed accomodation | 0.000 | 0.000 | 0.000 | 17.724 | 4.384 | 28.004 |
| social class B1 | 0.000 | 0.000 | 0.000 | 10.301 | 2.742 | 13.554 |
| social class B2 | 0.021 | 0.000 | 0.000 | 14.932 | 4.878 | 19.143 |
| social class C | 0.000 | 0.000 | 0.000 | 13.422 | 6.506 | 16.150 |
| social class D | 0.083 | 0.000 | 0.000 | 14.973 | 6.221 | 18.680 |
| household categories | | | | | | |
| single<35 | 0.000 | 0.000 | 0.000 | 8.036 | 2.288 | 10.100 |
| single>=35 | 0.000 | 0.000 | 0.000 | 8.036 | 2.288 | 10.100 |
| kids 0-5 | 0.021 | 0.021 | 0.000 | 4.987 | 1.522 | 6.444 |
| kids 6-12 | 0.021 | 0.021 | 0.000 | 4.987 | 1.522 | 6.444 |
| kids 13-17 | 0.021 | 0.021 | 0.000 | 4.987 | 1.522 | 6.444 |
| kids 0-5, 6-12 | 0.021 | 0.021 | 0.000 | 4.987 | 1.522 | 6.444 |
| kids 0-5, 13-17 | 0.000 | 0.000 | 0.000 | 5.655 | 1.993 | 18.955 |
| kids 6-12, 13-17 | 0.021 | 0.021 | 0.000 | 4.987 | 1.522 | 6.444 |
| kids 0-5, 6-12, 13-17 | 0.021 | 0.021 | 0.000 | 5.247 | 1.458 | 6.618 |
| hh>1, head<35 | 0.021 | 0.021 | 0.000 | 4.987 | 1.522 | 6.444 |
| hh>1, head>=35 | 0.021 | 0.021 | 0.000 | 4.987 | 1.522 | 6.444 |

This table shows the percentage rejections per significance level on the convergence tests, and statistics of the inefficiency factors, over draws for all outcome categories. Parameters for which all draws are equal to the base category, which parameter values are identical to zero, are not included in this analysis. The diagnostics are discussed in Section 6.3.8.

# Nederlandse Samenvatting (Summary in Dutch)

Als een onderzoeker tegenwoordig een economische vraag wil beantwoorden heeft hij steeds vaker een grote hoeveelheid data tot zijn beschikking. Veel datasets kenmerken zich door een groot aantal variabelen ten opzichte van het aantal observaties. De economische groei van verschillende landen kan worden verklaard met een groot aantal economische indicatoren, zoals variabelen die de kwaliteit van onderwijs, gezondheidszorg en rechtspraak beschrijven. Het aantal observaties voor deze variabelen is echter gelimiteerd door het aantal landen in de wereld. Voor het voorspellen van inflatie over tijd is er misschien nog wel meer data beschikbaar. Denk bijvoorbeeld aan productiecijfers, werkloosheid, of inflatie in het verleden. Veel van deze economische variabelen worden maar eens per kwartaal geobserveerd. Als we naar het keuzegedrag van mensen kijken observeren we een groot aantal karakteristieken, zoals inkomen, leeftijd en woonplaats, en bevatten keuzesets vaak veel alternatieven, zoals alle mogelijke vakantiebestemmingen.

Om iets te leren van deze datasets schatten onderzoekers de relatie tussen de variabelen en hetgeen ze geïnteresseerd in zijn in econometrische modellen. Met andere woorden, voor elke variabele wordt een aparte parameter geschat. Veel schattingsmethoden werken echter niet wanneer het aantal parameters groter is dan het aantal observaties. Zelfs als het schatten van parameters nog wel mogelijk is, wordt het vaak heel moeilijk om conclusies te trekken op basis van de grote hoeveelheid parameterschattingen. Bovendien daalt de precisie van schattingsmethoden naarmate meer parameters tegelijk worden geschat. Meer data betekent dus niet automatisch dat er meer inzichten zijn te verkrijgen.

Vaak is er geen informatie beschikbaar over welke variabelen relevant zijn voor het beantwoorden van een onderzoeksvraag. Met behulp van slimme aannames over de parameters

is het mogelijk toch iets te leren van een groot aantal parameters. Dit proefschrift onderzoekt twee verschillende aannames. De eerste neemt aan dat parameters geclusterd kunnen worden in groepen met identieke parameter waarden. De tweede legt een restrictie op de grootte van de parameter waarden. Dit proefschrift gebruikt deze twee aannames voor de volgende twee doelstellingen: voorspellende en bijschrijvende analyses mogelijk maken wanneer er een grote hoeveelheid data beschikbaar is.

Het volgende hoofdstuk van dit proefschrift, Hoofdstuk 2, onderzoekt de voorspellingen van professionele voorspellers voor macro-economische variabelen. In plaats van te voorspellen met een econometrisch model dat zo goed mogelijk alle beschikbare informatie gebruikt, kan een onderzoeker ook vertrouwen op de voorspellingen van experts. Hoofdstuk 2 concludeert echter dat professionele voorspellers vooral de trend en de conjunctuurcyclus in deze variabelen voorspellen, twee componenten die ook goed met modellen zijn te voorspellen. Over onregelmatige gebeurtenissen, die lastig door modellen zijn te vangen, hebben professionals ook weinig informatie. Hoofdstuk 3 richt zich vervolgens op het verbeteren van econometrische voorspel modellen: Hoe kan een econometrisch model de instabiliteit in de economie over tijd meenemen in voorspellingen? Hoofdstuk 3 clustert parameters in regimes over de tijd. Binnen de regimes zijn de parameter waarden identiek, maar de waarden zijn verschillend tussen de regimes. Dit model werkt goed maar is wel complex. Hoofdstuk 4 laat zien dat ook met hele simpele methoden, het willekeurig selecteren of wegen van een kleine set van variabelen uit een grote set van beschikbare variabelen, de voorspel prestaties niet onder doen voor complexe econometrische methoden wanneer de parameter waarden niet te groot zijn.

Het tweede deel van het proefschrift richt zich minder op voorspellen maar meer op het analyseren van de relaties tussen een groot aantal variabelen. Hoofdstuk 5 analyseert een simpel lineair regressie model waarin het aantal variabelen, en dus ook het aantal parameters, groter is dan het aantal observaties. Het ontwikkelt methoden om alle parameter waarden te schatten samen met de onzekerheid over de parameter waarden. Hoofdstuk 6 bekijkt een simpel keuze model. In dit model neemt het aantal parameters toe met het aantal keuzecategorieën en met het aantal karakteristieken van de beslissingsmakers. Het aantal parameters benadert dus gemakkelijk het aantal observaties. Het hoofdstuk laat zien hoe onderzoekers op basis van de data kunnen schatten welke keuzecategorieën bij elkaar geclusterd kunnen worden in groepen van keuzes met identieke parameter waarden.

# Bibliography


Ahlswede, R., Winter, A., 2002. Strong converse for identification via quantum channels. IEEE Transactions on Information Theory 48 (3), 569–579.

Albert, A., 1972. Regression and the Moore-Penrose pseudoinverse. Elsevier.

Allenby, G. M., Rossi, P. E., 1998. Marketing models of consumer heterogeneity. Journal of Econometrics 89 (1), 57–78.

Altug, S., Çakmaklı, C., 2016. Forecasting inflation using survey expectations and target inflation: Evidence for Brazil and Turkey. International Journal of Forecasting 32 (1), 138–153.

Andrews, R. L., Srinivasan, T., 1995. Studying consideration effects in empirical choice models using scanner panel data. Journal of Marketing Research 32 (1), 30.

Ang, A., Bekaert, G., Wei, M., 2007. Do macro variables, asset markets, or surveys forecast inflation better? Journal of Monetary Economics 54 (4), 1163–1212.

Antoniak, C. E., 1974. Mixtures of Dirichlet processes with applications to Bayesian non-parametric problems. The Annals of Statistics 2 (6), 1152–1174.

Bai, J., Ng, S., 2006. Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions. Econometrica 74 (4), 1133–1150.

Bai, J., Ng, S., 2008. Forecasting economic time series using targeted predictors. Journal of Econometrics 146 (2), 304–317.

Bańbura, M., Giannone, D., Reichlin, L., 2010. Large Bayesian vector auto regressions. Journal of Applied Econometrics 25 (1), 71–92.

Barro, R. J., Lee, J.-W., 1993. International comparisons of educational attainment. Journal of Monetary Economics 32 (3), 363–394.

Bauwens, L., Carpantier, J.-F., Dufays, A., 2017. Autoregressive moving average infinite hidden Markov-switching models. Journal of Business & Economic Statistics 35 (2), 162–182.

Baxter, M., 1994. Real exchange rates and real interest differentials: Have we missed the business-cycle relationship? Journal of Monetary Economics 33 (1), 5–37.

Baxter, M., King, R. G., 1999. Measuring business cycles: Approximate band-pass filters for economic time series. Review of Economics and Statistics 81 (4), 575–593.

Belloni, A., Chernozhukov, V., Hansen, C., 2010. Inference for high-dimensional sparse econometric models. Advances in Economics and Econometrics 3.

Belloni, A., Chernozhukov, V., et al., 2013. Least squares after model selection in high-dimensional sparse models. Bernoulli 19 (2), 521–547.

Bickel, P., Ritov, Y., Tsybakov, A., 2009. Simultaneous analysis of Lasso and Dantzig selector. Annals of Statistics 37, 1705–1732.

Billio, M., Casarin, R., Ravazzolo, F., Van Dijk, H. K., 2013. Time-varying combinations of predictive densities using nonlinear filtering. Journal of Econometrics 177 (2), 213–232.

Bondell, H. D., Reich, B. J., 2009. Simultaneous factor selection and collapsing levels in ANOVA. Biometrics 65 (1), 169–177.

Bonhomme, S., Manresa, E., 2015. Grouped patterns of heterogeneity in panel data. Econometrica 83 (3), 1147–1184.

Boot, T., Nibbering, D., 2017a. Forecasting using random subspace methods, Working paper.

Boot, T., Nibbering, D., 2017b. Inference in high-dimensional linear regression models, Working paper.

Breiman, L., 1996. Bagging predictors. Machine learning 24 (2), 123–140.

Bronnenberg, B. J., Vanhonacker, W. R., 1996. Limited choice sets, local price response, and implied measures of price competition. Journal of Marketing Research 33 (2).

Bühlmann, P., Kalisch, M., Meier, L., 2014. High-dimensional statistics with a view toward applications in biology. Annual Review of Statistics and Its Application 1, 255–278.

Bühlmann, P., Van De Geer, S., 2011. Statistics for high-dimensional data: methods, theory and applications. Springer Science & Business Media.

Bühlmann, P., Yu, B., 2002. Analyzing bagging. Annals of Statistics 30 (4), 927–961.

Bühlmann, P., et al., 2013. Statistical significance in high-dimensional linear models. Bernoulli 19 (4), 1212–1242.

Bunch, D. S., 1991. Estimability in the multinomial probit model. Transportation Research Part B: Methodological 25 (1), 1–12.

Burda, M., Harding, M., Hausman, J., 2008. A Bayesian mixed logit–probit model for multinomial choice. Journal of Econometrics 147 (2), 232–246.

Burgette, L. F., Nordheim, E. V., 2012. The trace restriction: An alternative identification strategy for the Bayesian multinomial probit model. Journal of Business & Economic Statistics 30 (3), 404–410.

Candes, E., Tao, T., 2007. The Dantzig selector: Statistical estimation when p is much larger than n. Annals of Statistics, 2313–2351.

Caner, M., Kock, A. B., 2014. Asymptotically honest confidence regions for high dimensional parameters by the desparsified conservative lasso. arXiv preprint arXiv:1410.4208.

Carriero, A., Clark, T. E., Marcellino, M., 2015a. Bayesian VARs: specification choices and forecast accuracy. Journal of Applied Econometrics 30 (1), 46–73.

Carriero, A., Clark, T. E., Marcellino, M., 2015b. Common drifting volatility in large Bayesian VARs. Journal of Business & Economic Statistics (forthcoming).

Carson, R. T., Louviere, J. J., 2014. Statistical properties of consideration sets. Journal of Choice Modelling 13, 37–48.

Chernozhukov, V., Hansen, C., Spindler, M., 2015. Valid post-selection and post-regularization inference: An elementary, general approach. Annual Review of Economics 7 (1), 649–688.

Chiang, J., Chib, S., Narasimhan, C., 1998. Markov chain Monte Carlo and models of consideration set and parameter heterogeneity. Journal of Econometrics 89 (1), 223–248.

Chib, S., 1998. Estimation and comparison of multiple change-point models. Journal of Econometrics 86 (2), 221–241.

Chib, S., Nardari, F., Shephard, N., 2006. Analysis of high dimensional multivariate stochastic volatility models. Journal of Econometrics 134 (2), 341–371.

Chikuse, Y., 1990. The matrix angular central Gaussian distribution. Journal of Multivariate Analysis 33 (2), 265–274.

Chikuse, Y., 2012. Statistics on special manifolds. Vol. 174. Springer Science & Business Media.

Chiong, K. X., Shum, M., 2016. Random projection estimation of discrete-choice models with large choice sets. USC-INET Research Paper 2016 (16-14).

Chopin, N., Pelgrin, F., 2004. Bayesian inference and state number determination for hidden Markov models: an application to the information content of the yield curve about inflation. Journal of Econometrics 123 (2), 327–344.

Christiano, L. J., Fitzgerald, T. J., 2003. The band pass filter. International Economic Review 44 (2), 435–465.

Claeskens, G., Hjort, N. L., February 2008. Model Selection and Model Averaging. Cambridge University Press.

Clark, T. E., 2012. Real-time density forecasts from Bayesian vector autoregressions with stochastic volatility. Journal of Business & Economic Statistics 29 (3), 327–341.

Clark, T. E., Ravazzolo, F., 2015. Macroeconomic forecasting performance under alternative specifications of time-varying volatility. Journal of Applied Econometrics 30 (4), 551–575.

Clements, M. P., 2015. Are professional macroeconomic forecasters able to do better than forecasting trends? Journal of Money, Credit and Banking 47 (2-3), 349–382.

Cogley, T., Primiceri, G. E., Sargent, T. J., 2010. Inflation-gap persistence in the US. American Economic Journal: Macroeconomics 2 (1), 43–69.

Cogley, T., Sargent, T. J., 2002. Evolving post-world WWII US inflation dynamics. NBER Macroeconomics Annual 2001 16, 331–388.

Cogley, T., Sargent, T. J., 2005. Drifts and volatilities: monetary policies and outcomes in the post WWII US. Review of Economic Dynamics 8 (2), 262–302.

Coibion, O., Gorodnichenko, Y., 2012. What can survey forecasts tell us about information rigidities? Journal of Political Economy 120 (1), 116–159.

Coibion, O., Gorodnichenko, Y., 2015. Information rigidity and the expectations formation process: A simple framework and new facts. The American Economic Review 105 (8), 2644–2678.

Colander, D., Goldberg, M., Haas, A., Juselius, K., Kirman, A., Lux, T., Sloth, B., 2009. The financial crisis and the systemic failure of the economics profession. Critical Review 21 (2-3), 249–267.

Conley, T. G., Hansen, C. B., McCulloch, R. E., Rossi, P. E., 2008. A semi-parametric Bayesian approach to the instrumental variable problem. Journal of Econometrics 144 (1), 276–305.

Cramer, J. S., Ridder, G., 1991. Pooling states in the multinomial logit model. Journal of Econometrics 47 (2-3), 267–272.

Croushore, D., Stark, T., 2001. A real-time data set for macroeconomists. Journal of econometrics 105 (1), 111–130.

D'Agostino, A., Gambetti, L., Giannone, D., 2013. Macroeconomic forecasting and structural change. Journal of Applied Econometrics 28 (1), 82–101.

Dasgupta, S., Hsu, D., Verma, N., 2012. A concentration theorem for projections. arXiv preprint arXiv:1206.6813.

Diebold, F. X., Mariano, R. S., 1995. Comparing predictive accuracy. Journal of Business & Economic Statistics 13 (3), 253–263.

Doan, T., Litterman, R., Sims, C., 1984. Forecasting and conditional projection using realistic prior distributions. Econometric reviews 3 (1), 1–100.

Durbin, J., Koopman, S. J., 2012. Time series analysis by state space methods. No. 38. Oxford University Press.

Elliott, G., Gargano, A., Timmermann, A., 2013. Complete subset regressions. Journal of Econometrics 177 (2), 357–373.

Elliott, G., Gargano, A., Timmermann, A., 2015. Complete subset regressions with large-dimensional sets of predictors. Journal of Economic Dynamics and Control 54, 86–110.

Engle, R. F., Granger, C., 1987. Co-integration and error correction: Representation, estimation, and testing. Econometrica 55 (2), 251–276.

Escobar, M. D., West, M., 1995. Bayesian density estimation and inference using mixtures. Journal of the American Statistical Association 90 (430), 577–588.

Fan, J., Lv, J., 2008. Sure independence screening for ultrahigh dimensional feature space. Journal of the Royal Statistical Society: Series B 70 (5), 849–911.

Fan, J., Peng, H., et al., 2004. Nonconcave penalized likelihood with a diverging number of parameters. Annals of Statistics 32 (3), 928–961.

Ferguson, T. S., 1973. A Bayesian analysis of some nonparametric problems. The annals of statistics 1 (2), 209–230.

Fernandez, C., Ley, E., Steel, M. F., 2001. Model uncertainty in cross-country growth regressions. Journal of Applied Econometrics 16 (5), 563–576.

Fox, E. B., Sudderth, E. B., Jordan, M. I., Willsky, A. S., 2011. A sticky HDP-HMM with application to speaker diarization. The Annals of Applied Statistics 5 (2A), 1020–1056.

Francq, C., Zakoıan, J.-M., 2001. Stationarity of multivariate Markov–switching ARMA models. Journal of Econometrics 102 (2), 339–364.

Franses, P. H., Kranendonk, H. C., Lanser, D., 2011. One model and various experts: Evaluating Dutch macroeconomic forecasts. International Journal of Forecasting 27 (2), 482–495.

Franses, P. H., Legerstee, R., 2010. Do experts' adjustments on model-based SKU-level forecasts improve forecast quality? Journal of Forecasting 29 (3), 331–340.

Frieze, A., Kannan, R., Vempala, S., 2004. Fast Monte-Carlo algorithms for finding low-rank approximations. Journal of the Association for Computing Machinery 51 (6), 1025–1041.

Frühwirth-Schnatter, S., 2001. Markov chain Monte Carlo estimation of classical and dynamic switching and mixture models. Journal of the American Statistical Association 96 (453), 194–209.

Frühwirth-Schnatter, S., 2006. Finite mixture and Markov switching models. Springer Science & Business Media.

Fuentes, J., Poncela, P., Rodríguez, J., 2015. Sparse partial least squares in time series for macroeconomic forecasting. Journal of Applied Econometrics 30 (4), 576–595.

Gertheiss, J., Tutz, G., et al., 2010. Sparse modeling of categorial explanatory variables. The Annals of Applied Statistics 4 (4), 2150–2180.

Geweke, J., 1992. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In: Bayesian Statistics. University Press, pp. 169–193.

Geweke, J., 2007a. Interpretation and inference in mixture models: Simple MCMC works. Computational Statistics & Data Analysis 51 (7), 3529–3550.

Geweke, J., 2007b. Interpretation and inference in mixture models: Simple MCMC works. Computational Statistics & Data Analysis 51 (7), 3529–3550.

Ghysels, E., Wright, J. H., 2009. Forecasting professional forecasters. Journal of Business & Economic Statistics 27 (4), 504–516.

Giannone, D., Lenza, M., Primiceri, G. E., 2015. Prior selection for vector autoregressions. Review of Economics and Statistics 97 (2), 436–451.

Gil-Alana, L., Moreno, A., Pérez de Gracia, F., 2012. Exploring survey-based inflation forecasts. Journal of Forecasting 31 (6), 524–539.

Gilbride, T. J., Allenby, G. M., 2004. A choice model with conjunctive, disjunctive, and compensatory screening rules. Marketing Science 23 (3), 391–406.

Giordani, P., Kohn, R., van Dijk, D., 2007. A unified approach to nonlinearity, structural change, and outliers. Journal of Econometrics 137 (1), 112–133.

Gneiting, T., Raftery, A. E., 2007. Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association 102 (477), 359–378.

Golden, S., 1965. Lower bounds for the Helmholtz function. Physical Review 137 (4B), B1127.

Groen, J. J., Kapetanios, G., 2016. Revisiting useful approaches to data-rich macroeconomic forecasting. Computational Statistics & Data Analysis 100, 221–239.

Groen, J. J., Paap, R., Ravazzolo, F., 2013. Real-time inflation forecasting in a changing world. Journal of Business & Economic Statistics 31 (1), 29–44.

Guhaniyogi, R., Dunson, D. B., 2015. Bayesian compressed regression. Journal of the American Statistical Association 110 (512), 1500–1514.

Hamilton, J. D., 1989. A new approach to the economic analysis of nonstationary time series and the business cycle. Econometrica 57 (2), 357–384.

Hansen, B. E., 2008. Least-squares forecast averaging. Journal of Econometrics 146 (2), 342–350.

Harvey, A. C., 1985. Trends and cycles in macroeconomic time series. Journal of Business & Economic Statistics 3 (3), 216–227.

Harvey, A. C., 1990. Forecasting, structural time series models and the Kalman filter. Cambridge University Press.

Harvey, A. C., Trimbur, T. M., 2003. General model-based filters for extracting cycles and trends in economic time series. Review of Economics and Statistics 85 (2), 244–255.

Hastie, T., Tibshirani, R., Wainwright, M., 2015. Statistical learning with sparsity: the lasso and generalizations. CRC press.

Hirano, K., 2002. Semiparametric Bayesian inference in autoregressive panel data models. Econometrica 70 (2), 781–799.

Hirano, K., Wright, J. H., 2017. Forecasting with model uncertainty: Representations and risk reduction. Econometrica 85 (2), 617–643.
URL http://dx.doi.org/10.3982/ECTA13372

Ho, T.-H., Chong, J.-K., 2003. A parsimonious model of stockkeeping-unit choice. Journal of Marketing Research 40 (3), 351–365.

Hoerl, A. E., Kennard, R. W., 1970. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 12 (1), 55–67.

Hou, C., 2017. Infinite hidden Markov switching VARs with application to macroeconomic forecast. International Journal of Forecasting 33 (4), 1025–1043.

Hu, X., Munkin, M. K., Trivedi, P. K., 2015. Estimating incentive and selection effects in the Medigap insurance market: An application with Dirichlet process mixture model. Journal of Applied Econometrics 30 (7), 1115–1143.

Hyndman, R. J., Koehler, A. B., 2006. Another look at measures of forecast accuracy. International Journal of Forecasting 22 (4), 679–688.

Imai, K., Van Dyk, D. A., 2005. A Bayesian analysis of the multinomial probit model using marginal data augmentation. Journal of Econometrics 124 (2), 311–334.

Inoue, A., Kilian, L., 2008. How useful is bagging in forecasting economic time series? A case study of US consumer price inflation. Journal of the American Statistical Association 103 (482), 511–522.

Ishwaran, H., James, L. F., 2002. Approximate Dirichlet process computing in finite normal mixtures: Smoothing and prior information. Journal of Computational and Graphical Statistics 11 (3), 508–532.

Ishwaran, H., Zarepour, M., 2000. Markov chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models. Biometrika 87 (2), 371–390.

Ishwaran, H., Zarepour, M., 2002. Exact and approximate sum representations for the Dirichlet process. Canadian Journal of Statistics 30 (2), 269–283.

Isiklar, G., Lahiri, K., Loungani, P., 2006. How quickly do forecasters incorporate news? Evidence from cross-country surveys. Journal of Applied Econometrics 21 (6), 703–725.

Jacobs, B. J., Donkers, B., Fok, D., 2016. Model-based purchase predictions for large assortments. Marketing Science 35 (3), 389–404.

Javanmard, A., Montanari, A., 2014. Confidence intervals and hypothesis testing for high-dimensional regression. Journal of Machine Learning Research 15 (1), 2869–2909.

Jochmann, M., 2015. Modeling US inflation dynamics: A Bayesian nonparametric approach. Econometric Reviews 34 (5), 537–558.

Johnson, W. B., Lindenstrauss, J., 1984. Extensions of Lipschitz mappings into a Hilbert space. Contemporary mathematics 26 (189-206), 1.

Kabán, A., 2014. New bounds on compressive linear least squares regression. Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics, PMLR 33, 448–456.

Keane, M. P., 1992. A note on identification in the multinomial probit model. Journal of Business & Economic Statistics 10 (2), 193–200.

Kim, S., Shephard, N., Chib, S., 1998. Stochastic volatility: Likelihood inference and comparison with ARCH models. The Review of Economic Studies 65 (3), 361–393.

Kock, A. B., Callot, L., 2015. Oracle inequalities for high dimensional vector autoregressions. Journal of Econometrics 186 (2), 325–344.

Koop, G., Korobilis, D., 2013. Large time-varying parameter VARs. Journal of Econometrics 177 (2), 185–198.

Koop, G., Korobilis, D., Pettenuzzo, D., 2016. Bayesian compressed vector autoregressions. Available at SSRN 2754241.

Koop, G., Potter, S. M., 2007. Estimation and forecasting in models with multiple breaks. The Review of Economic Studies 74 (3), 763–789.

Koop, G. M., 2013. Forecasting with medium and large Bayesian VARs. Journal of Applied Econometrics 28 (2), 177–203.

Kozicki, S., Tinsley, P. A., 2012. Effective use of survey information in estimating the evolution of expected inflation. Journal of Money, Credit and Banking 44 (1), 145–169.

Lan, W., Zhong, P.-S., Li, R., Wang, H., Tsai, C.-L., 2016. Testing a single regression coefficient in high dimensional linear models. Journal of Econometrics 195 (1), 154–168.

Leitch, G., Ernesttanner, J., 1995. Professional economic forecasts: Are they worth their costs? Journal of Forecasting 14 (2), 143–157.

Liu, P., Mumtaz, H., Theodoridis, K., Zanetti, F., 2017. Changing macroeconomic dynamics at the zero lower bound. Journal of Business & Economic Statistics (just-accepted).

Liu, Q., Arora, N., 2011. Efficient choice designs for a consider-then-choose model. Marketing Science 30 (2), 321–338.

Ma, P., Mahoney, M. W., Yu, B., Jan. 2015. A statistical perspective on algorithmic leveraging. Journal of Machine Learning Research 16 (1), 861–911.

MacKinnon, J. G., 1996. Numerical distribution functions for unit root and cointegration tests. Journal of Applied Econometrics 11 (6), 601–618.

Mahoney, M. W., Drineas, P., 2009. CUR matrix decompositions for improved data analysis. Proceedings of the National Academy of Sciences 106 (3), 697–702.

Maillard, O., Munos, R., 2009. Compressed least-squares regression. In: Advances in Neural Information Processing Systems. Vol. 22. pp. 1213–1221.

Manzini, P., Mariotti, M., 2014. Stochastic choice and consideration sets. Econometrica 82 (3), 1153–1176.

Marzetta, T. L., Tucci, G. H., Simon, S. H., 2011. A random matrix-theoretic approach to handling singular covariance estimates. IEEE Transactions on Information Theory 57 (9), 6256–6271.

McCracken, M. W., Ng, S., 2016. FRED-MD: A monthly database for macroeconomic research. Journal of Business & Economic Statistics 34 (4), 574–589.

McCulloch, R., Rossi, P. E., 1994. An exact likelihood analysis of the multinomial probit model. Journal of Econometrics 64 (1), 207–240.

McCulloch, R. E., Polson, N. G., Rossi, P. E., 2000. A Bayesian analysis of the multinomial probit model with fully identified parameters. Journal of Econometrics 99 (1), 173–193.

Mehra, Y. P., 2002. Survey measures of expected inflation: Revisiting the issues of predictive content and rationality. Economic Quarterly-Federal Reserve Bank of Richmond 88 (3), 17–36.

Mehta, N., Rajiv, S., Srinivasan, K., 2003. Price uncertainty and consumer search: A structural model of consideration set formation. Marketing Science 22 (1), 58–84.

Meinshausen, N., Yu, B., 2009. Lasso-type recovery of sparse representations for high-dimensional data. The Annals of Statistics, 246–270.

Mertens, E., 2016. Measuring the level and uncertainty of trend inflation. Review of Economics and Statistics 98 (5), 950–967.

Murphy, K. M., Topel, R. H., 2002. Estimation and inference in two-step econometric models. Journal of Business & Economic Statistics 20 (1), 88–97.

Naik, P., Wedel, M., Bacon, L., Bodapati, A., Bradlow, E., Kamakura, W., Kreulen, J., Lenk, P., Madigan, D. M., Montgomery, A., 2008. Challenges and opportunities in high-dimensional choice data analyses. Marketing Letters 19 (3-4), 201–213.

Newey, W. K., West, K. D., 1987. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. Econometrica 55 (3), 703–708.

Ng, S., 2013. Variable selection in predictive regressions. Handbook of Economic Forecasting 2 (Part B), 752–789.

Ng, S., 2015. Opportunities and challenges: Lessons from analyzing terabytes of scanner data. Tech. rep., Columbia University.

Nibbering, D., 2017. A high-dimensional multinomial choice model with an application to holiday destinations, Working paper.

Nibbering, D., Paap, R., van der Wel, M., 2017a. A Bayesian infinite hidden Markov vector autoregressive model, Working paper.

Nibbering, D., Paap, R., van der Wel, M., 2017b. What do professional forecasters actually predict?, forthcoming International Journal of Forecasting.

Park, J. Y., Phillips, P. C., 1989. Statistical inference in regressions with integrated processes: Part 2. Econometric Theory 5 (01), 95–131.

Pearson, K., 1901. On lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 2 (11), 559–572.

Pesaran, M. H., Pettenuzzo, D., Timmermann, A., 2006. Forecasting time series subject to multiple structural breaks. The Review of Economic Studies 73 (4), 1057–1084.

Pollock, D., 2000. Trend estimation and de-trending via rational square-wave filters. Journal of Econometrics 99 (2), 317–334.

Primiceri, G. E., 2005. Time varying structural vector autoregressions and monetary policy. The Review of Economic Studies 72 (3), 821–852.

Reid, S., Tibshirani, R., Friedman, J., 2016. A study of error variance estimation in lasso regression. Statistica Sinica 26, 35–67.

Rietveld, C. A., Medland, S. E., Derringer, J., Yang, J., Esko, T., Martin, N. W., Westra, H.-J., Shakhbazov, K., Abdellaoui, A., Agrawal, A., et al., 2013. GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. Science 340 (6139), 1467–1471.

Rossi, B., 2013. Advances in forecasting under instability. Handbook of Economic Forecasting 2 (Part B), 1203–1324.

Rossi, P. E., Allenby, G. M., Robert, M., 2005. Bayesian Statistics and Marketing. John Wiley & Sons, Ltd.

Sala-i-Martin, X. X., 1997. I just ran two million regressions. The American Economic Review, 178–183.

Schneider, M. J., Gupta, S., 2016. Forecasting sales of new and existing products using consumer reviews: A random projections approach. International Journal of Forecasting 32 (2), 243–256.

Serfling, R. J., 2006. Multivariate symmetry and asymmetry. Encyclopedia of Statistical Sciences.

Sethuraman, J., 1994. A constructive definition of Dirichlet priors. Statistica Sinica 4, 639–650.

Sims, C. A., 1980. Macroeconomics and reality. Econometrica: Journal of the Econometric Society, 1–48.

Sims, C. A., Zha, T., 2006. Were there regime switches in US monetary policy? The American Economic Review, 54–81.

Song, Y., 2014. Modelling regime switching and structural breaks with an infinite hidden Markov model. Journal of Applied Econometrics 29 (5), 825–842.

Stock, J. H., Watson, M. W., 1998. Median unbiased estimation of coefficient variance in a time-varying parameter model. Journal of the American Statistical Association 93 (441), 349–358.

Stock, J. H., Watson, M. W., 2002. Forecasting using principal components from a large number of predictors. Journal of the American Statistical Association 97 (460), 1167–1179.

Stock, J. H., Watson, M. W., 2005. Implications of dynamic factor models for VAR analysis. Tech. rep., National Bureau of Economic Research.

Stock, J. H., Watson, M. W., 2006. Forecasting with many predictors. Handbook of Economic Forecasting 1, 515–554.

Stock, J. H., Watson, M. W., 2012. Generalized shrinkage methods for forecasting using many predictors. Journal of Business & Economic Statistics 30 (4), 481–493.

Su, L., Shi, Z., Phillips, P. C., 2016. Identifying latent structures in panel data. Econometrica 84 (6), 2215–2264.

Sun, T., Zhang, C.-H., 2012. Scaled sparse linear regression. Biometrika 99 (4), 879. URL +http://dx.doi.org/10.1093/biomet/ass043

Teh, Y. W., Jordan, M. I., Beal, M. J., Blei, D. M., 2012. Hierarchical Dirichlet processes. Journal of the American Statistical Association 101 (476), 1566–1581.

Terui, N., Ban, M., Allenby, G. M., 2011. The effect of media advertising on brand consideration and choice. Marketing Science 30 (1), 74–91.

Thanei, G.-A., Heinze, C., Meinshausen, N., 2017. Random projections for large-scale regression. In: Big and Complex Data Analysis. Springer, pp. 51–68.

Thomas, L. B., et al., 1999. Survey measures of expected United States inflation. Journal of Economic Perspectives 13 (4), 125–144.

Thompson, C. J., 1965. Inequality with applications in statistical mechanics. Journal of Mathematical Physics 6 (11), 1812–1813.

Tibshirani, R., 1996. Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society. Series B 58 (1), 267–288.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., Knight, K., 2005. Sparsity and smoothness via the fused lasso. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67 (1), 91–108.

Tutz, G., Gertheiss, J., 2016. Regularized regression for categorical data. Statistical Modelling 16 (3), 161–200.

van de Geer, S., Bühlmann, P., Ritov, Y., Dezeure, R., et al., 2014. On asymptotically optimal confidence regions and tests for high-dimensional models. Annals of Statistics 42 (3), 1166–1202.

Van de Geer, S. A., 2008. High-dimensional generalized linear models and the lasso. The Annals of Statistics, 614–645.

Van den Hauwe, S., 2015. Topics in applied macroeconometrics. Ph.D. thesis, Erasmus School of Economics.

Van Hasselt, M., 2011. Bayesian inference in a sample selection model. Journal of Econometrics 165 (2), 221–232.

Van Nierop, E., Bronnenberg, B., Paap, R., Wedel, M., Franses, P. H., 2010. Retrieving unobserved consideration sets from household panel data. Journal of Marketing Research 47 (1), 63–74.

Vershynin, R., 2010. Introduction to the non-asymptotic analysis of random matrices. arXiv preprint arXiv:1011.3027.

Wang, X., Leng, C., 2015. High dimensional ordinary least squares projection for screening variables. Journal of the Royal Statistical Society: Series B.

White, H., 1980. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. Econometrica 48 (4), 817–838.

White, H., 1984. Asymptotic Theory for Econometricians. Academic Press.

Wiesenfarth, M., Hisgen, C. M., Kneib, T., Cadarso-Suarez, C., 2014. Bayesian nonparametric instrumental variables regression based on penalized splines and Dirichlet process mixtures. Journal of Business & Economic Statistics 32 (3), 468–482.

Wigderson, A., Xiao, D., 2008. Derandomizing the Ahlswede-Winter matrix-valued Chernoff bound using pessimistic estimators, and applications. Theory of Computing 4 (1), 53–76.

Wold, H., 1982. Soft modelling: the basic design and some extensions. Systems under indirect observation, Part II, 36–37.

Zanutto, E. L., Bradlow, E. T., 2006. Data pruning in consumer choice models. Quantitative Marketing and Economics 4 (3), 267–287.

Zhang, C.-H., Zhang, S. S., 2014. Confidence intervals for low dimensional parameters in high dimensional linear models. Journal of the Royal Statistical Society: Series B 76 (1), 217–242.

Zou, H., 2006. The adaptive lasso and its oracle properties. Journal of the American Statistical Association 101 (476), 1418–1429.

The Tinbergen Institute is the Institute for Economic Research, which was founded in 1987 by the Faculties of Economics and Econometrics of the Erasmus University Rotterdam, University of Amsterdam and VU University Amsterdam. The Institute is named after the late Professor Jan Tinbergen, Dutch Nobel Prize laureate in economics in 1969. The Tinbergen Institute is located in Amsterdam and Rotterdam. The following books recently appeared in the Tinbergen Institute Research Series:

666. J. JI, *Three Essays in Empirical Finance*

667. H. SCHMITTDIEL, *Paid to Quit, Cheat, and Confess*

668. A. DIMITROPOULOS, *Low Emission Vehicles: Consumer Demand and Fiscal Policy*

669. G.H. VAN HEUVELEN, *Export Prices, Trade Dynamics and Economic Development*

670. A. RUSECKAITE, *New Flexible Models and Design Construction Algorithms for Mixtures and Binary Dependent Variables*

671. Y. LIU, *Time-varying Correlation and Common Structures in Volatility*

672. S. HE, *Cooperation, Coordination and Competition: Theory and Experiment*

673. C.G.F. VAN DER KWAAK, *The Macroeconomics of Banking*

674. D.H.J. CHEN, *Essays on Collective Funded Pension Schemes*

675. F.J.T. SNIEKERS, *On the Functioning of Markets with Frictions*

676. F. GOMEZ MARTINEZ, *Essays in Experimental Industrial Organization: How Information and Communication affect Market Outcomes*

677. J.A. ATTEY, *Causes and Macroeconomic Consequences of Time Variations in Wage Indexation*

678. T. BOOT, *Macroeconomic Forecasting under Regime Switching, Structural Breaks and High-dimensional Data*

679. I. TIKOUDIS, *Urban Second-best Road Pricing: Spatial General Equilibrium Perspectives*

680. F.A. FELSŐ, *Empirical Studies of Consumer and Government Purchase Decisions*

681. Y. GAO, *Stability and Adaptivity: Preferences over Time and under Risk*

682. M.J. ZAMOJSKI, *Panta Rhei, Measurement and Discovery of Change in Financial Markets*

683. P.R. DENDERSKI, *Essays on Information and Heterogeneity in Macroeconomics*

684. U. TURMUNKH, *Ambiguity in Social Dilemmas*

685. U. KESKIN, *Essays on Decision Making: Intertemporal Choice and Uncertainty*

686. M. LAMMERS, *Financial Incentives and Job Choice*

687. Z. ZHANG, *Topics in Forecasting Macroeconomic Time Series*

688. X. XIAO, *Options and Higher Order Risk Premiums*

689. D.C. SMERDON, *'Everybody's doing it': Essays on Trust, Norms and Integration*

690. S. SINGH, *Three Essays on the Insurance of Income Risk and Monetary Policy*

691. E. SILDE, *The Econometrics of Financial Comovement*

692. G. DE OLIVEIRA, *Coercion and Integration*

693. S. CHAN, *Wake Me up before you CoCo: Implications of Contingent Convertible Capital for Financial Regulation*

694. P. GAL, *Essays on the role of frictions for firms, sectors and the macroeconomy*

695. Z. FAN, *Essays on International Portfolio Choice and Asset Pricing under Financial Contagion*

696. H. ZHANG, *Dealing with Health and Health Care System Challenges in China: Assessing Health Determinants and Health Care Reforms*

697. M. VAN LENT, *Essays on Intrinsic Motivation of Students and Workers*

698. R.W. POLDERMANS, *Accuracy of Method of Moments Based Inference*

699. J.E. LUSTENHOUWER, *Monetary and Fiscal Policy under Bounded Rationality and Heterogeneous Expectations*

700. W. HUANG, *Trading and Clearing in Modern Times*

701. N. DE GROOT, *Evaluating Labor Market Policy in the Netherlands*

702. R.E.F. VAN MAURIK, *The Economics of Pension Reforms*

703. I. AYDOGAN, *Decisions from Experience and from Description: Beliefs and Probability Weighting*

704. T.B. CHILD, *Political Economy of Development, Conflict, and Business Networks*

705. O. HERLEM, *Three Stories on Influence*

706. J.D. ZHENG, *Social Identity and Social Preferences: An Empirical Exploration*

707. B.A. LOERAKKER, *On the Role of Bonding, Emotional Leadership, and Partner Choice in Games of Cooperation and Conflict*

708. L. ZIEGLER, *Social Networks, Marital Sorting and Job Matching. Three Essays in Labor Economics*

709. M.O. HOYER, *Social Preferences and Emotions in Repeated Interactions*

710. N. GHEBRIHIWET, *Multinational Firms, Technology Transfer, and FDI Policy*

711. H.FANG, *Multivariate Density Forecast Evaluation and Nonparametric Granger Causality Testing*

712. Y. KANTOR, *Urban Form and the Labor Market*

713. R.M. TEULINGS, *Untangling Gravity*

714. K.J.VAN WILGENBURG, *Beliefs, Preferences and Health Insurance Behavior*

715. L. SWART, *Less Now or More Later? Essays on the Measurement of Time Preferences in Economic Experiments*