Macroeconomic Forecasting under Regime Switching, Structural Breaks and High-dimensional Data

Macroeconomic Forecasting under Regime Switching, Structural Breaks and High-dimensional Data

Macro-economische voorspellingen bij wisselende economische omstandigheden, structurele veranderingen en hoog-dimensionale data

Thesis

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

prof.dr. H.A.P. Pols

and in accordance with the decision of the Doctorate Board.

The public defense shall be held on

Friday, February 17, 2017 at 11:30

by

TOM BOOT

born in Zevenaar, The Netherlands

Erasmus University Rotterdam

Doctorate Committee

Promotor:	Prof.dr. R. Paap
Other members:	Prof.dr. H.P. Boswijk
	Prof.dr. D.J. van Dijk
	Prof.dr. G. Kapetanios
Copromotor:	Dr. A. Pick

ISBN: 978 90 3610 466 1

Cover design: Crasborn Graphic Designers bno, Valkenburg a.d. Geul

This book is no. 678 of the Tinbergen Institute Research Series, established through cooperation between Rozenberg Publishers and the Tinbergen Institute. A list of books which already appeared in the series can be found in the back.

Acknowledgments

Meer dan vier jaar geleden begon ik met mijn promotieonderzoek: bergen artikelen lezen, seminars bijwonen, huiswerkopdrachten en tentamens nakijken, supercomputeruren verbruiken, en één onnavolgbaar boek van een Russische wiskundige proberen te doorgronden. Van Rotterdam, Londen, Toulouse, Genève, San Diego en Kaapstad uiteindelijk terug naar Groningen. Het was een fantastische ontdekkingsreis, waarbij ik me ook wel eens afvroeg hoe ik dit traject ooit tot een goed eind zou brengen. Met de steun, relativering en afleiding gebracht door familie, vrienden en collega's is het nu gelukt! Ik ben blij dat ik jullie hier op de eerste pagina's van dit proefschrift voor kan bedanken.

Andreas, thank you for all our inspiring conversations. I never felt for a moment that you doubted whether these four years would end well. Needless to say, I am looking forward to continue working together in the coming years!
Richard, jouw colleges waren een belangrijke reden om te solliciteren voor een promotieplek. Het was vanaf het begin duidelijk dat je me zou helpen als ik het nodig had. Ontzetend bedankt voor al je hulp in de totstandkoming van dit proefschrift.

I greatly appreciate the time and energy the doctorate committee, consisting of Dick van Dijk, Peter Boswijk and George Kapetanios, has spend on reading this thesis and providing helpful comments. Part of the work on this thesis was carried out at during my stay at UCSD. Many thanks to Allan Timmermann for his hospitality during this visit.

To all my colleagues: a big thank you for sharing our experiences, ideas, drinks and occasional frustrations over the past years. Bruno, voor H7-33, "één-twee-EIPC!", en voor het financieren van een research visit naar Maryland; Bart, voor het aanhoren van de meest onsamenhangende verhalen over mijn onderzoek in het Paviljoen; Francine, voor alle gesprekken over de universiteit, maar vooral ook over alles daarbuiten; Aiste, for all the laughs we shared over funny and occasionally strange experiences in teaching and research; Didier, voor alle inspirerende telefoontjes over onderzoek. Om 23:00. Op zaterdag; Koen, voor al je (toen nog gratis) adviezen; Sander, voor de broodnodige Amsterdamse flair; Gertjan, voor het redden van een berg uren op LISA die ik toen alsnog heb verbruikt; Max (Maaaaax!) en Bert, Dennis, Elio, Matthijs, Myrthe en Victor voor alle leuke lunches en conferenties.

Ik heb altijd geluk gehad met mijn docenten. Dit begon al op de middelbare school met de mooie[1] en áltijd-áltijd-áltijd drie regels lange bewijzen van Thijs van der Velden. Toen ik zelf begon met lesgeven heb ik veel plezier gehad van de inspirerende blik en het enthousiasme van Martyn Mulder en Jan Brinkhuis. In het managementteam heb ik vervolgens veel geleerd over de organisatie achter al die colleges, waarvoor dank Albert Wagelmans, Patrick Groenen en Erik Kole.

Wonen in Groningen en werken in Rotterdam is een logistieke uitdaging. Ik ben heel blij dat ik het afgelopen jaar bij zoveel vrienden en vriendinnen welkom was om een nacht, een week of zelfs een maand te blijven slapen! Thijs, die een eindeloze reeks couscous-maaltijden op de koop toe nam; Frank-Jan en Lizebeth, die zelfs speciaal in *Den Haag* gingen wonen; Daan en Milou die mij hun loft toevertrouwden inclusief de ~~duurste~~ meest overprijsde fietswielen die ik ooit heb gezien; Eef: van fantastische barbecues tot net iets minder fantastische crèpes, van popquiz (wow!) tot pubquiz (au!) en alle onmisbare gesprekken daar tussenin; Tom met een ontzettend klein bed, waar je dus zo vanaf valt; Mijke en Ivar tegenover het beste restaurant van Rotterdam; Francine en Victor waar ik een hele maand heb genoten van het gaafste uitzicht; Rob, een gouden oom met dito stem; en Hilde met de meest paradijselijke slaapplek, alhoewel de reistijd vanaf Aruba wel een beetje onpraktisch was.

Sinds de oprichting in 1634 is er geen fijnere generatie DSB'ers geweest: Jakko, Jasper, Hedde, Mark, Pelle, Peynacker, Rudy en Sander. Fier, halm, heaugh! En bedankt alvast voor het jasje.
Hidde, Klaas-Jan en Tom, het is jammer dat de camera het niet altijd goed vast kan leggen, maar bedankt voor het onwaarschijnlijk hoge niveau op (en naast!) het kunstgrasgraveltapijt. "En, is je werkstuk al af?", bedankt voor de relativering BJ, Stoeltie, Hermsen en Andries.

Pap en mam, jullie zijn de beste, liefste en meest betrokken ouders die ik me kan wensen. Als ik thuis ben, is het altijd goed. Rens en Tomas, mijn lieve zus en zwager, bedankt dat jullie me zijn komen redden in Amerika! Oma, bedankt voor al uw aandacht, vragen en luisterend oor (nu ook via WhatsApp!). Opa, bedankt voor alle mooie verhalen, en de pretogen waarmee u nog steeds over de hekjes van de tuin klimt. Lieve Ernie, gelukkig is er ook na het oppassen nog veel gezelligheid en af en toe een dikke knuffel! Klaas en Baukje, Marije en Bas, Ilse, en Durk, bedankt voor een tweede thuis in Rinsumageest. Het is altijd een feest om er te zijn (en zeker niet alleen vanwege de appeltaart)!

Lieve Fem, met jou beleef ik de allermooiste avonturen, samen $\overset{\infty}{\underset{\smile}{\geqslant}} 1000\times$ leuker! En wanneer gaan we eigenlijk weer eens op vakantie?

[1] Wiskunde is vaak mooi: "Deze matrix vind ik zo mooi, daar zet ik even een hartje bij" (Lineaire Algebra, Professor Trentelman). Voor natuurkundigen en econometristen is de wereld helaas soms iets minder mooi, "We shuffle these infinities under the carpet" (General Relativity, Professor Bergshoeff)

Contents

Chapter 1

Introduction

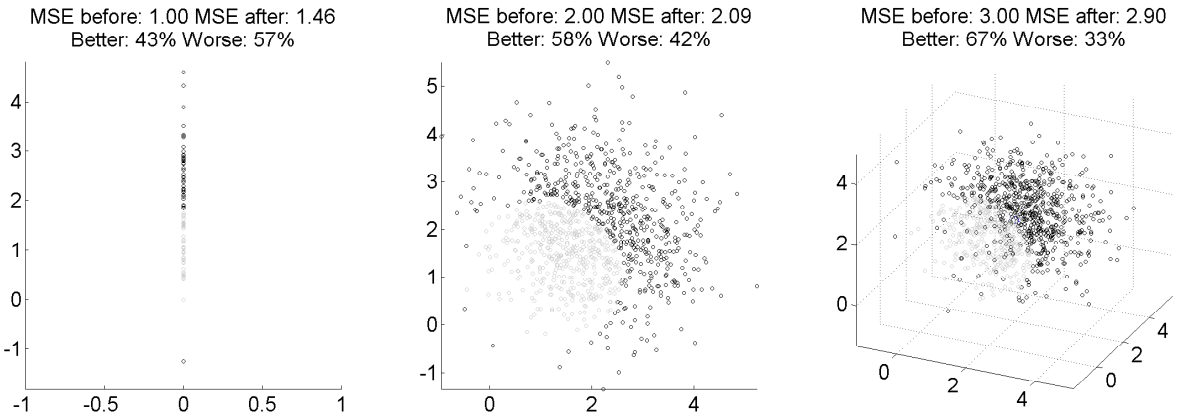
In order to make informed policy decisions, it is essential for central banks, government planning agencies, pension funds, investors and other economic agents to understand how the economy will develop in future periods. However, forecasting economic time series is by no means an easy task. A wealth of often unstable interactions makes it impossible to exactly identify what the future will look like. Instead, we have to settle for estimates. Estimates carry uncertainty, and a forecaster needs to be aware of all sources of uncertainty to construct an accurate forecast.

To measure forecast accuracy, we use a loss function that assigns a weight to each forecast error. By far the most widely used loss function to measure forecast accuracy is mean squared error. This widespread use is in part due to ease of interpretation in terms of Euclidean distance and its mathematical convenience. Since the function is everywhere differentiable, it is easily optimized either analytically or by means of numerical techniques. Hansen (2016) provides a more fundamental argument for expected mean squared error loss by showing that it is the asymptotic limit under local alternatives for a much wider class of differentiable loss functions.

Expected mean squared error is the sum of two components: squared bias, which measures the size of the average forecast error, and variance, measuring the variation of the forecasts. Most frequently used statistical estimators are unbiased under a correctly specified model. Hence, their expected mean squared forecast error consists only of the variance of the estimator. Complex models, lack of data and outliers, all potentially increase the variance and render the unbiased forecast inaccurate in many empirical settings.

As an alternative to using complex models that yield unbiased forecasts, we can use much simpler models, which yield forecasts with a smaller variance. These simple models for example omit some or even all explanatory variables, or they neglect time variation in the interactions. Due to misspecification, the resulting forecasts can be heavily biased. However, there are many examples where these simplistic forecasts are difficult to beat. When predicting excess equity returns for example, Welch and Goyal (2008) find that it is hard

Figure 1.1: Graphical illustration of the Stein effect



Note: this figure shows $N = 1,000$ points $x_{ij} \sim N(\mu_j, 1)$ where $\mu_j = 2$, $i = 1, \ldots, 1000$ and $j = 1, \ldots, d$ with the dimension $d = \{1, 2, 3\}$ for the left, center and right panel respectively. If we observe a single point $\boldsymbol{x}_i$ in a d-dimensional space, the unbiased estimator is $\hat{\mu}_j = x_{ij}$. Shrinkage is applied by multiplying each point with a factor $\max\left[0, 1 - \frac{1}{\boldsymbol{x}_i'\boldsymbol{x}_i}\right]\boldsymbol{x}_i$. Points for which the mean squared error compared to the true mean $\boldsymbol{\mu}$ increases are depicted in gray. Points for the the mean squared error decreases are depicted in black. The average mean squared error over $N = 100,000$ points, scaled by the dimension, is shown above the figure. Below the percentage of points for which the accuracy increases/decreases.

to outperform a simple historical average with more complex models. A similar conclusion is reached by Meese and Rogoff (1983) who show that the accuracy of the random walk forecast is often superior to structural exchange rate models.

The question is whether somewhere in between unbiased and simplistic, we can find forecasts that optimally trade off bias and variance, and as a result are more precise than both. For unbiased, normally distributed estimators, Stein (1956) and James and Stein (1961) show that a uniform improvement is possible when expected mean squared error is measured on average over more than two parameter estimates. The improved estimators shrink the unbiased estimator towards the simplistic estimator by a data-dependent amount. Figure 1.1 illustrates this effect and the dependence on the number of parameters over which mean squared error is measured. For the dimension $d = 1, 2, 3$, a point is generated from a normal distribution, $\boldsymbol{x}_i \sim N(\boldsymbol{\mu}, \boldsymbol{I}_d)$ where $\boldsymbol{\mu} = 2 \cdot \boldsymbol{\iota}_d$ with $\boldsymbol{\iota}_d$ a d-dimensional vector of ones. Given such a point, the goal is to estimate the unknown mean $\boldsymbol{\mu}$ such that the expected mean squared error $\mathrm{E}\left[(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})'(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})\right]$ is small. The unbiased estimator simply takes $\hat{\boldsymbol{\mu}} = \boldsymbol{x}_i$ which yields an expected mean squared error equal to the dimension d . Alternatively, we shrink the point towards the origin by a data-dependent factor $\hat{\boldsymbol{\mu}}^{JS} = \max\left[0, 1 - \frac{1}{\boldsymbol{x}_i'\boldsymbol{x}_i}\right]\boldsymbol{x}_i$. This factor implies that shrinkage is stronger for points close to the origin, than for points that are far away.

For $d = 1$, Figure 1.1 shows the shrinkage approach is ineffective. However, as the number of parameters grows, the average accuracy, as well as that of an increasingly large

fraction of the estimates, increases. When $d = 3$, as anticipated based on the results by James and Stein (1961), we observe an increase in accuracy using the adjusted estimator.

Since a wide class of statistical estimators is (asymptotically) normally distributed, the above results imply that from an expected mean squared error perspective, these estimators can be improved. This surprising finding has spawned a literature on alternative estimation techniques, ranging from empirical Bayes methods introduced by Efron and Morris (1973), to the currently popular lasso by Tibshirani (1996). Perhaps ironically, the James-Stein estimator itself is much less frequently used in empirical work.

In the first two chapters of this thesis, we consider the bias-variance trade-off in models subject to regime switches and structural breaks. The non-linearities pose a challenge to find an optimal trade-off as the variance term is not easily tractable. The final two chapters discuss the trade-off in (high-dimensional) linear models, in which we consider estimators for which explicit bounds on their performance can be provided.

Part I: Forecasting under regime switches and structural breaks

An inherent feature of economic dynamics is that they can change substantially over a short period of time. As an example, consider the recent outcome of the British referendum on their EU membership, which had an immediate, large effect on the pound/dollar exchange rate. To accommodate this feature, models have been developed that allow for breaks in the parameters between subsequent time periods.

Consider a series where a single break occurs at some point in time. When the break date is known, an unbiased forecast uses only post-break data. However, if the break occurred only months ago, the unbiased forecast is based on a very limited amount of data. As a result, the variance is large, which leads to a large mean squared forecast error. To increase the amount of data and thereby reduce the variance, Pesaran and Timmermann (2007) propose to include some periods before the break. Alternatively, Pesaran et al. (2013) assign a non-zero weight to all pre-break observations.

Not always is the break with the past as evident as in the UK scenario described above. A new president of the central bank can have only a minor effect on the economy, as can changes in legislation, or a small change in available natural resources. The precise timing of the change is then uncertain, and enters the forecasters mean squared error loss function in a non-standard way. Standard inference in such scenarios is shown by Elliott and Müller (2014) to be unreliable. This complicates a successful implementation of the methods by Pesaran and Timmermann (2007) and Pesaran et al. (2013). One solution is to construct forecasts that are robust to the exact timing of the break as proposed by Pesaran et al. (2013).

An alternative to robust forecasts is to quantify the uncertainty around the break date and find a method that optimally incorporates this uncertainty when constructing a forecast. A useful starting point is the class of Markov switching models, popularized by Hamilton

(1989). These models capture switches between economic regimes, for example between extended periods of growth and often much shorter lived periods of recession. As a by-product of the estimation procedure, regime probabilities are produced, which provide insight into regime uncertainty.

Empirically, forecasts from Markov switching models have been found to be less accurate than those from simple linear models in a number of applications, for example when forecasting exchange rates by Engel (1994) and US GNP growth by Clements and Krolzig (1998). In Chapter 2, based on Boot and Pick (2016b), we find that the lack of forecasting accuracy from Markov switching models is in part due to the fact that the Markov switching forecasts do not optimally incorporate the regime uncertainty. Based on the provided estimates of this uncertainty, we can develop a weighting scheme for the observations that reduces the variance of the forecasts. This reduction is achieved by emphasizing the switching nature of the model. The weighting scheme is found to be especially effective in situations frequently encountered in empirical work, where the regimes are well apart, yet the overall state uncertainty remains relatively large.

The weights are derived for an arbitrary number of regimes and possibly concurrent breaks in the variance, and hence, can be applied to a wide variety of Markov switching models. We show that by properly weighting the observations, the Markov switching model is able to outperform linear alternatives in an application to forecasting U.S. GNP. Interestingly, the weighting scheme hardly increases the bias of the forecasts, indicating that the bias-variance trade-off in these non-linear models is not as straightforward as in their linear counterparts.

Whereas in Chapter 2 we analyze how to construct an accurate forecast, given that we are forecasting using a Markov switching model, Chapter 3, based on Boot and Pick (2016a), takes a different approach. Instead of treating the model as given, we take a step back and discuss how to decide between a simple linear model and a model that allows for a break. A range of existing tests analyzes the presence of breaks in the parameters of a model. Under a known break date, the test by Chow (1960), is uniformly most powerful. When the break date is unknown, one has to search over all possible break dates which requires adjusted testing procedures as developed by Andrews (1993), Andrews and Ploberger (1994) and Bai and Perron (1998).

Existing tests focus on the question whether a break in the parameters of the model has occurred, regardless of its size. However, from a forecasting perspective, we are only interested in breaks to the extend that these adversely affect the mean squared forecast error. It can happen that a break occurs in the parameters, but does not translate into a break in the forecast, which is a weighted average of the parameters. Second, if the break is small and the timing uncertain, the variance of the forecast might increase dramatically upon modeling the

break. This can reduce the mean squared forecast error compared models that simply ignore the non-linear characteristics of the model.

The above discussion implies that a different test is needed to determine the relevance of structural breaks for accurate forecasting. In Chapter 3, we develop such a test that differs from existing tests in two respects. First, instead of testing the parameters of the model, we test whether there is a significant difference between the forecasts from a linear model and a break model. Second, due to the bias-variance trade-off describe before, small breaks are allowed under the null hypothesis. This will increase the critical values as can be anticipated based on results by Toro-Vizcarrondo and Wallace (1968) who consider testing under mean squared error loss.

Conditional on the timing of the break, our test statistic is a Wald statistic where the parameters are weighted by the value of the regressors at the forecast horizon. Nuder the null hypothesis of equal predictive accuracy, this statistic has a simple asymptotic distribution that is non-central chi-squared with one degree of freedom and non-centrality parameter equal to one. In addition, we show that even when the break date is unknown a powerful test can be derived for the null of equal predictive accuracy between a post-break forecast and a full-sample forecast. The additional uncertainty introduced by not knowing the break date substantially increases the break size up to which a linear model is preferred. In an empirical example on 130 monthly macroeconomic time series, we find that far fewer breaks are relevant for forecasting than indicated by existing tests.

Part II: estimation and forecasting in linear models

In Chapter 4, based on Boot (2015), instead of forecasting we now consider estimation under mean squared error loss in the context of a simple linear model. As mentioned in the introduction, Stein (1956) showed that estimators exist that uniformly dominate the unbiased least squares estimator. However, this holds when we measure accuracy on average over more than two parameters. For individual parameters, the accuracy might very well be worse as the left panel of Figure 1.1 shows. In fact, for individual coefficients the unbiased estimator is the unique minimax expected mean squared error estimator, see for example Magnus and Durbin (1996).

Given that we cannot uniformly outperform the unbiased estimator, Chapter 4 introduces an estimator that provides a lower bound on the probability at which the estimates improve. The maximum risk of this estimator turns out to be relatively close to that of the least squares estimator. This offers a reliable shrinkage technique for individual parameter estimates as in the left panel of Figure 1.1. In addition, we show that when the dimension of the estimator is larger than two, in line with the James-Stein estimator, the expected mean squared error of the derived estimator is uniformly lower than that of the unbiased estimator.

The estimator is compared to the lasso introduced by Tibshirani (1996) through Monte Carlo simulations, which show that it is more robust to different realizations of the regressors. The gains over the standard least squares estimator are smaller then observed for the lasso when the signal is weak, but the maximum loss is much better under control. When we measure accuracy on average over all parameters, the performance of the estimator is equivalent to that of the estimator by James and Stein (1961).

In light of the increasing dimensions of recent macroeconomic databases, for example McCracken and Ng (2015), Chapter 5 based on Boot and Nibbering (2016), focuses on accurate forecasts from high-dimensional models. In these models the number of parameters is close to the sample size. As the variance of the unbiased least squares estimator scales with the ratio of the number of parameters over the sample size, dimension reduction techniques are often used to increase forecast performance. One example is principal component regression, which tries to capture the information in the data in a small number of factors, which are then used as predictor variables (Stock and Watson, 2002).

In contrast to principal component regression, the construction of the low-dimensional space we consider in Chapter 5, is fully random. We introduce two randomized strategies. Random subset regression, randomly samples subsets of predictors based on which forecasts are constructed. Instead of sampling all possible subsets as in Elliott et al. (2013), we show that no accuracy is lost when only a small fraction is estimated. The second strategy, random projection regression, constructs small sets of artificial predictors formed by randomly weighting the original predictors. Theoretical results by Johnson and Lindenstrauss (1984) have recently inspired several applications of this technique 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).

We obtain tight bounds on the expected mean squared forecast error for both randomization strategies. These bounds illustrate when the methods are expected to perform well. When the eigenvalues are roughly equal, the methods provide equally accurate forecasts. If on the other hand the data has a factor structure, than random projection regression works well when the dominant factors drive the variable of interest. Random subset regression is more suited to cases where lower order factors are important. An empirical application on 130 US monthly macroeconomic series shows that for a majority of the series both randomized methods outperform widely applied alternatives: principal component regression (Pearson, 1901), partial least squares (Wold, 1982), ridge regression (Hoerl and Kennard, 1970) and the lasso (Tibshirani, 1996).

Chapter 2

Optimal forecasts from Markov switching models

2.1 Introduction

Markov switching models have long been recognized to suffer from a discrepancy between in-sample and out-of-sample performance. In-sample analysis of Markov switching models often leads to appealing results, for example, the identification of business cycles. Out-of-sample performance, in contrast, is frequently inferior to simple benchmark models for standard loss functions. Examples include forecasting exchange rates by Engel (1994), Dacco and Satchell (1999) and Klaassen (2005), forecasting US GNP growth by Clements and Krolzig (1998) and Perez-Quiros and Timmermann (2001), forecasting US unemployment by Deschamps (2008), and forecasting house prices by Crawford and Fratantoni (2003). Additionally, Guidolin (2011) and Rapach and Zhou (2013) provide reviews of the use of Markov switching models in finance.

In this chapter, we derive minimum mean square forecast error (MSFE) forecasts for Markov switching models by means of optimal weighting schemes for observations. We provide simple, analytic expressions for the weights when the model has an arbitrary number of states and exogenous regressors. We find that forecasts using optimal weights substantially increase forecast precision and, in our application, are more precise than linear alternatives. Additionally, optimal weights lead to insights that help explain why standard Markov switching forecasts are often less precise than linear forecasts.

We start our discussion assuming that the states of the Markov switching model are known and, in a second step, we relax this assumption. When conditioning on the states, the intuition for the optimal weights can easily be seen: a forecast obtained from optimal weights pools all observations and places different weights on observations from different states. This reduces the variance of the forecast but introduces a bias. Optimally weighting all observations ensures that the trade-off is optimal in the MSFE sense. The usual Markov

switching forecasts, in contrast, assign non-zero weights only to observations from the state that will govern the forecast period. Conditional on the states of the Markov switching model, the weights mirror those obtained by Pesaran et al. (2013), emphasizing a correspondence with the structural break model. The weights depend on the number of observations per regime and the relative differences of the parameter between the regimes.

In the case of three regimes, the weights have interesting properties. For some parameter values, optimal weighting corresponds to equal weighting of observations. For other parameter values, observations from the state prevailing in the forecast period will not be most heavily weighted. However, conditional on the states of the Markov switching model, the optimal weights can be written as $\mathcal{O}(1/T)$ corrections to the usual Markov switching weights, which implies that, conditional on the states, standard Markov switching weights asymptotically achieve the minimum MSFE.

In practice, the states of the Markov switching model are not known with certainty. We therefore relax the assumption that the states are known and derive weights conditional on state probabilities, which is the information used in standard Markov switching forecasts. This results in optimal weights that no longer correspond to those for the structural break model. Contrasting weights conditional on states with those conditional on state probabilities yields insights into the effect that uncertainty around states has on forecasts. Our findings explain the deterioration of forecast accuracy of the optimal weights in the application of Pesaran et al. (2013) because plug-in estimates of the break date substantially shrink optimal weights towards equal weights. Weights conditional on states and the weights implicit in standard Markov switching forecasts downplay the Markov switching nature of the data when estimates of states are plugged in. Weights conditional on state probabilities, in contrast, retain the emphasis on the Markov switching nature of the data. This implies that the forecast accuracy from optimal weights conditional on state probabilities relative to that implied by standard Markov switching forecasts increases in the difference between the states in terms of their parameters and in the variance of the smoothed probabilities. The forecast improvements from using optimal weights do not vanish as the sample size increases as the standard weights and the optimal weights conditional on the state probabilities are not asymptotically equivalent.

We perform Monte Carlo experiments to evaluate the performance of the optimal weights. The results confirm the theoretically expected improvements. The weights that are derived conditional on the states and use the estimated probabilities as plug-in values improve over standard forecasts only for small differences in parameters, which are unlikely to lead to applications of Markov switching models in practice. The weights based on state probabilities, in contrast, produce substantial gains for large differences in parameters between states, uncertainty over the states, and large samples. These settings are likely to be found in many applications, including the one in this chapter.

We apply our methodology to forecasting quarterly US GNP. Out-of-sample forecasts are constructed for 124 quarters and a range of Markov switching models. At each point, forecasts are made with the Markov switching model that has the best forecasting history using standard weights. With this model, we calculate forecasts based on the standard Markov switching weights and the optimal weights developed in this chapter. The results suggest that the forecasts using optimal weights significantly outperform the standard Markov switching forecast. We also find that our forecasting schemes lead to improved forecasts compared to a range of linear alternatives. We analyze the sensitivity of the results to the choice of the out-of-sample forecast evaluation period using the tests of Rossi and Inoue (2012), which confirm our findings.

The outline of this chapter is as follows. Section 2.2 introduces the model and the standard forecast. In Section 2.3 we derive the optimal weights for a simple location model and in Section 2.4 for a model with exogenous regressors. Monte Carlo experiments are presented in Section 2.5 and an application to US GNP in Section 2.6. Finally, Section 2.7 concludes the chapter. Additional details are presented in the appendix.

2.2 Markov switching models and their forecasts

Consider the following m-state Markov switching model

$$y_t = (\mathbf{B}'\mathbf{s}_t)'\mathbf{x}_t + \boldsymbol{\sigma}'\mathbf{s}_t\varepsilon_t, \quad \varepsilon_t \sim iid(0,1) \tag{2.1}$$

where $\mathbf{B} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_m)'$ is an $m \times k$ matrix, $\boldsymbol{\beta}_i$ is a $k \times 1$ parameter vector for $i = 1, 2, \ldots, m$, $\mathbf{x}_t$ is a $k \times 1$ vector of exogenous regressors, $\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \ldots, \sigma_m)'$ is an $m \times 1$ vector of error standard deviations, and $\mathbf{s}_t = (s_{1t}, s_{2t}, \ldots, s_{mt})'$ is an $m \times 1$ vector of binary state indicators, such that $s_{it} = 1$ and $s_{jt} = 0$, $j \neq i$, if the process is in state i at time t.

This is the standard Markov switching model introduced by Hamilton (1989). The model is completed by a description of the stochastic process governing the states, where $\mathbf{s}_t$ is assumed to be an ergodic Markov chain with transition probabilities

$$\mathbf{P} = \begin{bmatrix} p_{11} & p_{21} & \cdots & p_{m1} \\ p_{12} & p_{22} & \cdots & p_{m2} \\ \vdots & \vdots & & \vdots \\ p_{1m} & p_{2m} & \cdots & p_{mm} \end{bmatrix}$$

where $p_{ij} = \mathrm{P}(s_{jt} = 1 | s_{i,t-1} = 1)$ is the transition probability from state i to state j.

The standard forecast, in this context, would be to estimate $\boldsymbol{\beta}_i$, $i = 1, 2, \ldots, m$, as

$$\hat{\boldsymbol{\beta}}_i = \left(\sum_{t=1}^{T} \hat{\xi}_{it} \mathbf{x}_t \mathbf{x}_t' \right)^{-1} \sum_{t=1}^{T} \hat{\xi}_{it} \mathbf{x}_t y_t \tag{2.2}$$

where $\hat{\xi}_{it}$ is the estimated probability that observation at time t is from state i using, for example, the smoothing algorithm of Kim (1994). The forecast is then constructed as $\hat{y}_{T+1} = \sum_{i=1}^{m} \hat{\xi}_{i,T+1} \mathbf{x}_{T+1}' \hat{\boldsymbol{\beta}}_i$, where $\hat{\xi}_{i,T+1}$ is the predicted probability of state i in the forecast period, and $\mathbf{x}_{T+1}$ is the vector of regressors in the forecast period, which we assume known at time T. See Hamilton (1994) for an introduction to the Markov switching modeling and forecasting.

In this chapter, we derive the minimum MSFE forecast for finite samples and different assumptions about the information set that the forecast is based on. We replace the estimated probabilities by general weights w_t for the forecast $\hat{y}_{T+1} = \mathbf{x}_{T+1}' \hat{\boldsymbol{\beta}}(\mathbf{w})$, so that

$$\hat{\boldsymbol{\beta}}(\mathbf{w}) = \left(\sum_{t=1}^{T} w_t \mathbf{x}_t \mathbf{x}_t' \right)^{-1} \sum_{t=1}^{T} w_t \mathbf{x}_t y_t$$

subject to the restriction $\sum_{t=1}^{T} w_t = 1$. The weights are restricted to sum to one as an identifying restriction is required, and we will see in the next section that this is the restriction of the standard Markov switching weights. We do, however, not restrict the weights to be positive. In fact, in Section 2.3.1 we will see that negative weights are a common feature in models with more than two states as they allow the cancellation of biases. The resulting forecasts are then optimal in the sense that the weights will be chosen such that they minimize the expected MSFE.

2.3 Optimal forecasts for a simple model

Initially, consider a simple version of model (2.1) with $k = 1$ and $x_t = 1$ such that

$$y_t = \boldsymbol{\beta}' \mathbf{s}_t + \boldsymbol{\sigma}' \mathbf{s}_t \varepsilon_t, \quad \varepsilon_t \sim iid(0, 1) \tag{2.3}$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_m)'$. We use this simple model for ease of exposition but will return to the full model (2.1) in Section 2.4 below.

We can derive the optimal forecast by using a weighted average of the observations with weights that minimize the MSFE. The forecast from weighted observations for (2.3) is

$$\hat{y}_{T+1} = \sum_{t=1}^{T} w_t y_t \tag{2.4}$$

subject to $\sum_{t=1}^{T} w_t = 1$.

Note, that the standard forecast can be expressed as (2.4) with weights

$$w_{\text{MS},t} = \sum_{i=1}^{M} \frac{\hat{\xi}_{i,T+1}\hat{\xi}_{it}}{\sum_{t=1}^{T}\hat{\xi}_{it}} \tag{2.5}$$

which only depend on the smoothed and predicted probabilities and have the property that $\sum_{t=1}^{T} w_{\text{MS},t} = 1$. We will call weights (2.5) the standard Markov switching weights.

In order to derive the optimal weights, consider the forecast error, which, without loss of generality, is scaled by the error standard deviation of regime m, is

$$\begin{aligned}\sigma_m^{-1}e_{T+1} &= \sigma_m^{-1}(y_{T+1} - \hat{y}_{T+1}) \\ &= \boldsymbol{\lambda}'\tilde{\mathbf{s}}_{T+1} + \mathbf{q}'\mathbf{s}_{T+1}\varepsilon_{T+1} - \sum_{t=1}^{T} w_t\boldsymbol{\lambda}'\tilde{\mathbf{s}}_t - \sum_{t=1}^{T} w_t\mathbf{q}'\boldsymbol{s}_t\varepsilon_t\end{aligned}$$

where

$$\boldsymbol{\lambda} = \begin{pmatrix} (\beta_2 - \beta_1)/\sigma_m \\ (\beta_3 - \beta_1)/\sigma_m \\ \vdots \\ (\beta_m - \beta_1)/\sigma_m \end{pmatrix}, \ \mathbf{q} = \begin{pmatrix} \sigma_1/\sigma_m \\ \sigma_2/\sigma_m \\ \vdots \\ 1 \end{pmatrix} \text{ and } \tilde{\boldsymbol{s}}_t = \begin{pmatrix} s_{2t} \\ s_{3t} \\ \vdots \\ s_{mt} \end{pmatrix}$$

The scaled MSFE is

$$\begin{aligned}\mathrm{E}\left(\sigma_m^{-2}e_{T+1}^2\right) &= \mathrm{E}\left\{\left[\boldsymbol{\lambda}'\left(\tilde{\mathbf{s}}_{T+1} - \sum_{t=1}^{T} w_t\tilde{\mathbf{s}}_t\right)\right]^2\right\} + \sum_{t=1}^{T} w_t^2\mathrm{E}\left[(\mathbf{q}'\boldsymbol{s}_t)^2\right] + \mathrm{E}\left[(\mathbf{q}'\mathbf{s}_{T+1})^2\right] \\ &= \mathrm{E}\left(\tilde{\mathbf{s}}_{T+1}'\boldsymbol{\lambda}\boldsymbol{\lambda}'\tilde{\mathbf{s}}_{T+1}\right) - 2\mathbf{w}'\mathrm{E}\left(\tilde{\mathbf{S}}'\boldsymbol{\lambda}\boldsymbol{\lambda}'\tilde{\mathbf{s}}_{T+1}\right) \\ &\quad + \mathbf{w}'\mathrm{E}\left(\tilde{\mathbf{S}}'\boldsymbol{\lambda}\boldsymbol{\lambda}'\tilde{\mathbf{S}}\right)\mathbf{w} + \mathrm{E}\left[(\mathbf{q}'\mathbf{s}_{T+1})^2\right] + \mathbf{w}'\mathrm{E}(\mathbf{Q})\mathbf{w} \\ &= \mathbf{w}'\left[\mathrm{E}(\mathbf{Q}) + \mathrm{E}\left(\tilde{\mathbf{S}}'\boldsymbol{\lambda}\boldsymbol{\lambda}'\tilde{\mathbf{S}}\right)\right]\mathbf{w} - 2\mathbf{w}'\mathrm{E}\left(\tilde{\mathbf{S}}'\boldsymbol{\lambda}\boldsymbol{\lambda}'\tilde{\mathbf{s}}_{T+1}\right) \\ &\quad + \mathrm{E}\left(\tilde{\mathbf{s}}_{T+1}'\boldsymbol{\lambda}\boldsymbol{\lambda}'\tilde{\mathbf{s}}_{T+1}\right) + \mathrm{E}\left[(\mathbf{q}'\mathbf{s}_{T+1})^2\right]\end{aligned} \tag{2.6}$$

where $\tilde{\mathbf{S}} = (\tilde{\boldsymbol{s}}_1, \tilde{\boldsymbol{s}}_2, \ldots, \tilde{\boldsymbol{s}}_T)$, $\mathbf{S} = (\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_T)$ and $\mathbf{Q}$ is a diagonal matrix with typical (t,t)-element $Q_{tt} = \sum_{i=1}^{m} q_i^2 s_{it}$. The first line of (2.6) contains the squared bias as the first expression on the right hand side, the variance of the estimated parameters as the second term and, finally, the variance of the future disturbance term. The weights will trade off the first and second term on the right hand side to minimize the MSFE. The last term, in contrast, cannot be reduced.

Furthermore, define

$$\mathbf{M} = \mathrm{E}(\mathbf{Q}) + \mathrm{E}\left(\tilde{\mathbf{S}}'\boldsymbol{\lambda}\boldsymbol{\lambda}'\tilde{\mathbf{S}}\right) \tag{2.7}$$

and note that $\mathbf{M}$ is invertible as $\mathbf{Q}$ is a diagonal matrix with positive entries and $\mathrm{E}\left(\tilde{\mathbf{S}}'\boldsymbol{\lambda}\boldsymbol{\lambda}'\tilde{\mathbf{S}}\right)$ is positive semidefinite, so that $\mathbf{M}$ is the sum of a positive definite matrix and a positive semi-definite matrix and therefore itself positive definite.

Minimizing (2.6) subject to $\sum_{t=1}^{T} w_t = 1$ yields the optimal weights

$$\mathbf{w} = \mathbf{M}^{-1}\mathrm{E}\left(\tilde{\mathbf{S}}'\boldsymbol{\lambda}\boldsymbol{\lambda}'\tilde{\mathbf{s}}_{T+1}\right) + \frac{\mathbf{M}^{-1}\boldsymbol{\iota}}{\boldsymbol{\iota}'\mathbf{M}^{-1}\boldsymbol{\iota}}\left[1 - \boldsymbol{\iota}'\mathbf{M}^{-1}\mathrm{E}\left(\tilde{\mathbf{S}}'\boldsymbol{\lambda}\boldsymbol{\lambda}'\tilde{\mathbf{s}}_{T+1}\right)\right] \tag{2.8}$$

where $\boldsymbol{\iota} = (1, 1, \ldots, 1)'$ is a $T \times 1$ vector of ones. We will discuss the properties of the optimal weights in Sections 2.3.1 and 2.3.2 under different assumption about the information set. The MSFE given by (2.6) when applying the optimal weights (2.8) is

$$\begin{aligned}\mathrm{MSFE}(\mathbf{w}) = {} & \frac{\left[1 - \boldsymbol{\iota}'\mathbf{M}^{-1}\mathrm{E}\left(\tilde{\mathbf{S}}'\boldsymbol{\lambda}\boldsymbol{\lambda}'\tilde{\mathbf{s}}_{T+1}\right)\right]^2}{\boldsymbol{\iota}'\mathbf{M}^{-1}\boldsymbol{\iota}} + \mathrm{E}\left(\tilde{\mathbf{s}}_{T+1}'\boldsymbol{\lambda}\boldsymbol{\lambda}'\tilde{\mathbf{s}}_{T+1}\right) \\ & -\mathrm{E}\left(\tilde{\mathbf{S}}'\boldsymbol{\lambda}\boldsymbol{\lambda}'\tilde{\mathbf{s}}_{T+1}\right)'\mathbf{M}^{-1}\mathrm{E}\left(\tilde{\mathbf{S}}'\boldsymbol{\lambda}\boldsymbol{\lambda}'\tilde{\mathbf{s}}_{T+1}\right) + \mathrm{E}\left[(\mathbf{q}'\mathbf{s}_{T+1})^2\right]\end{aligned} \tag{2.9}$$

In order to proceed, we need to specify the information set that is available to calculate the expectations in (2.8) and (2.9). Initially, we will base the weights on the full information set of the DGP, including the state for each observation. Clearly, this information is not available in practice. However, the resulting analysis will prove to be highly informative. The intuition that is gained will prove useful when interpreting the forecast that we will obtain subsequently when allowing for uncertainty around the states. This second step will enable us to analyze the differences between the plug-in estimator for the weights that assume knowledge of the states and optimal weights that are derived under the assumption that the states are uncertain.

Note, that we condition on $\boldsymbol{\lambda}$ throughout our analysis. The reason is that, in a decomposition of the optimal weights for the structural break case, Pesaran et al. (2013) show that the time of the break enters the weights in a term that is of order $\mathcal{O}(1/T)$, whereas the size of the break, $\boldsymbol{\lambda}$, enters the weights in a term that is of order $\mathcal{O}(1/T^2)$. We will show below that the optimal weights for the Markov switching model conditional on the states are equivalent to the weights of Pesaran et al. (2013) and their argument therefore carries over to the Markov switching model.

2.3.1 Weights conditional on the states

Conditional on the states the expectation operator in (2.7), (2.8) and (2.9) can be omitted such that $\mathbf{M} = \mathbf{Q} + \tilde{\mathbf{S}}'\boldsymbol{\lambda}\boldsymbol{\lambda}'\tilde{\mathbf{S}}$ and $\mathrm{E}\left(\tilde{\mathbf{S}}'\boldsymbol{\lambda}\boldsymbol{\lambda}'\tilde{\mathbf{s}}_{T+1}\right) = \tilde{\mathbf{S}}'\boldsymbol{\lambda}\boldsymbol{\lambda}'\tilde{\mathbf{s}}_{T+1}$. Given the number of states, weights can now readily be derived.

Two-state Markov switching models

In the case of a two-state Markov switching model, $\tilde{\mathbf{s}} = (s_{21}, s_{22}, \ldots, s_{2T})'$ and therefore $\mathbf{M} = \mathbf{Q} + \lambda^2 \tilde{\mathbf{s}}\tilde{\mathbf{s}}'$ for which the inverse is given by

$$\begin{aligned}\mathbf{M}^{-1} &= \mathbf{Q}^{-1} - \frac{\lambda^2}{1 + \lambda^2 \tilde{\mathbf{s}}' \mathbf{Q}^{-1} \tilde{\mathbf{s}}} \mathbf{Q}^{-1} \tilde{\mathbf{s}}\tilde{\mathbf{s}}' \mathbf{Q}^{-1} \\ &= \mathbf{Q}^{-1} - \frac{\lambda^2}{1 + \lambda^2 T \pi_2} \tilde{\mathbf{s}}\tilde{\mathbf{s}}'\end{aligned}$$

where $\lambda^2 = \frac{(\beta_2 - \beta_1)^2}{\sigma_2^2}$ and $\pi_i = \frac{1}{T}\sum_{t=1}^{T} s_{it}$. The elements of the diagonal matrix $\mathbf{Q}$ are $Q_{tt} = q^2 s_{1t} + s_{2t}$ with $q = \frac{\sigma_1}{\sigma_2}$. This yields the following weights:

When $s_{1,T+1} = 1$,

$$w_{(1,1)} = \frac{1}{T} \frac{1 + T\lambda^2 \pi_2}{\pi_2 q^2 + \pi_1 (1 + T\pi_2 \lambda^2)} \quad \text{if } s_{1t} = 1 \tag{2.10}$$

$$w_{(1,2)} = \frac{1}{T} \frac{q^2}{\pi_2 q^2 + \pi_1 (1 + T\pi_2 \lambda^2)} \quad \text{if } s_{2t} = 1 \tag{2.11}$$

where $w_{(i,j)}$ is the weight for an observation when $s_{jt} = 1$ while $s_{i,T+1} = 1$.

When $s_{2,T+1} = 1$,

$$w_{(2,1)} = \frac{1}{T} \frac{1}{[\pi_2 q^2 + \pi_1 (1 + T\pi_2 \lambda^2)]} \quad \text{if } s_{1t} = 1 \tag{2.12}$$

$$w_{(2,2)} = \frac{1}{T} \frac{q^2 + T\lambda^2 \pi_1}{[\pi_2 q^2 + \pi_1 (1 + T\pi_2 \lambda^2)]} \quad \text{if } s_{2t} = 1 \tag{2.13}$$

Note that, conditional on the state of the future observation, the weights are symmetric under a relabeling of the states. Derivations are provided in Appendix 2.A.1.

The weights are equivalent to the weights for the break point process developed by Pesaran et al. (2013). This implies that, conditional on the states, a Markov switching model is equivalent to a break point model with known break point with the exception that the observations are ordered by the underlying Markov process.

Since the weights $w_{(1,2)}$ and $w_{(2,1)}$ are nonzero, the decrease in the variance of the optimal weights forecast should outweigh the increase in the squared bias that results from using all observations. The expected MSFE under the above weights is

$$\mathrm{E}\big[\sigma_2^{-2} e_{T+1}^2\big]_{\text{opt}} = \begin{cases} q^2(1 + w_{(1,1)}) & \text{if } s_{1,T+1} = 1 \\ 1 + w_{(2,2)} & \text{if } s_{2,T+1} = 1 \end{cases} \tag{2.14}$$

Table 2.1: Ratio between the expected MSFEs of optimal and standard MS weights

λ	$q = 1$			$q = 0.5$		
	$\pi_2 = 0.1$	0.2	0.5	0.1	0.2	0.5
0	0.8500	0.9273	0.9808	0.8500	0.9273	0.9808
0.5	0.9294	0.9758	0.9953	0.9268	0.9745	0.9949
1	0.9727	0.9919	0.9986	0.9724	0.9918	0.9985
2	0.9921	0.9978	0.9996	0.9921	0.9978	0.9996

Note: Reported are the ratio between (2.14) and (2.15) when $s_{2,T+1} = 1$ for different values of λ, the difference in means, and q, the ratio of standard deviations, and π_2, the proportion of observations in state 2. $T = 50$.

while the expected MSFE for standard Markov switching weights is

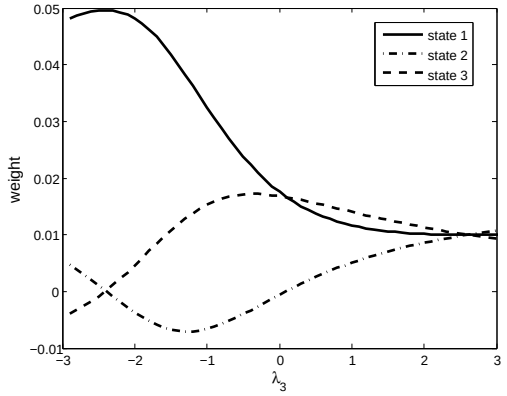
$$\mathrm{E}[\sigma_2^{-2} e_{T+1}^2]_{\mathrm{MS}} = \begin{cases} q^2 \left(1 + \frac{1}{T\pi_1}\right) & \text{if } s_{1,T+1} = 1 \\ 1 + \frac{1}{T\pi_2} & \text{if } s_{2,T+1} = 1 \end{cases} \tag{2.15}$$

It is easy to show that $\mathrm{E}[\sigma_2^{-2} e_{T+1}^2]_{\mathrm{opt}} < \mathrm{E}[\sigma_2^{-2} e_{T+1}^2]_{\mathrm{MS}}$.

Numerical examples of the magnitude of the improvement in MSFE are presented in Table 2.1, which shows that the improvements scale inversely with the differences in parameters. To gain intuition for these results, consider the case of $\lambda = 0$ and $q = 1$, that is, the case where the two states have identical means, and $s_{1,T+1} = 1$. The standard Markov switching model will use weights $w_{\mathrm{MS},(1,1)} = \frac{1}{T\pi_1}$ and $w_{\mathrm{MS},(1,2)} = 0$, which results in an MSFE of $1 + \frac{1}{T\pi_1}$. The optimal weights, in contrast, are $= w_{\mathrm{opt},(1,1)} = w_{\mathrm{opt},(1,2)} = 1/T$, and the MSFE is $1 + 1/T$. The usual Markov switching forecast disregards the information from the second state even though, in this case, it is highly informative, whereas the optimal weights forecast uses all the observation equally as one would suggest intuitively, given that the states have the same mean.

As λ increases, the usefulness of the observations in state 2 decreases because the bias introduced by these observations increases. This is reflected in the numbers in Table 2.1. The same intuition can be gained by increasing or decreasing q away from 1. The difference in MSFE also depends on π_1, that is, the fraction of observations in the state used for forecasting in the standard Markov switching forecast. The fewer observations are available for the standard Markov switching forecast the more valuable will the observation from the second state be. Finally, as T increases, for a fixed π_1, the parameter estimates will be more precise so that any further gains from using observation in the second state will be less important. In fact, we show below that, asymptotically, the optimal weights and the standard weights are identical. However, as we will show in Section 2.3.2, the asymptotic equivalence of optimal and standard weights relies on the fact that the states are known with certainty. With uncertainty around the states, the gain from using optimal weights will not disappear with large T.

Figure 2.1: Optimal weights for three state Markov switching model



Note: The graph depicts the optimal weights (2.16) for a representative observation in each state when $s_{1,T+1} = 1$, for $\lambda_2 = -2.5$, λ_3 over the range -3 to 3, $T = 100$, $\pi_1 = 0.2$, and $\pi_2 = \pi_3 = 0.4$. The solid line gives the weights for a representative observation where $s_{1t} = 1$, the dash-dotted line a representative observation where $s_{2t} = 1$, and the dashed line a representative observation for $s_{3t} = 1$.

Three-state Markov switching models

If $s_{j,T+1} = 1$, then define $q_i^2 = \sigma_i^2/\sigma_j^2$ and $\lambda_i^2 = (\beta_i - \beta_j)^2/\sigma_j^2$ where $i, j \in \{1, 2, 3\}$. The optimal weights are

$$\begin{aligned} w_{(j,j)} &= \frac{1}{T} \frac{1 + T\sum_{i=1}^{3} q_i^{-2}\lambda_i^2\pi_i}{\sum_{i=1}^{3} q_i^{-2}\pi_i + T\sum_{i=1}^{3}\sum_{m=1}^{3} q_i^{-2}q_m^{-2}\pi_i\pi_m\lambda_m(\lambda_m - \lambda_i)} \\ w_{(j,k)} &= \frac{1}{T} \frac{q_k^{-2} + Tq_k^{-2}\sum_{i=1}^{m} q_i^{-2}\lambda_i\pi_i(\lambda_i - \lambda_k)}{\sum_{i=1}^{3} q_i^{-2}\pi_i + T\sum_{i=1}^{3}\sum_{m=1}^{3} q_i^{-2}q_m^{-2}\pi_i\pi_m\lambda_i(\lambda_i - \lambda_m)} \\ w_{(j,l)} &= \frac{1}{T} \frac{q_l^{-2} + Tq_l^{-2}\sum_{i=1}^{m} q_i^{-2}\lambda_i\pi_i(\lambda_i - \lambda_l)}{\sum_{i=1}^{3} q_i^{-2}\pi_i + T\sum_{i=1}^{3}\sum_{m=1}^{3} q_i^{-2}q_m^{-2}\pi_i\pi_m\lambda_m(\lambda_i - \lambda_m)} \end{aligned} \tag{2.16}$$

where $j, k, l \in \{1, 2, 3\}$. Derivations are available in Appendix 2.A.1.

Figure 2.1 plots weights (2.16) for $s_{1,T+1} = 1$, that is, the future observation is known to be from the first state. The difference in mean between the first and second state relative to the variance of the first state is set to $\lambda_2 = -2.5$, and the difference in mean between the first and third state, λ_3, varies from -3 to 3. Furthermore, the proportions of observations for the states are $\pi_1 = 0.2$, $\pi_2 = \pi_3 = 0.4$, $T = 100$, and the ratio of variances is $q_1 = q_2 = 1$. Each line represents the weight for one representative observation in each state. As 20 observations are in state 1 and 40 in the other two states, it can easily be verified that the weights sum to one. Consider the weights at $\lambda_3 = -2.5$: each observation in state one is weighted with $w_{(1,1)} \approx 0.05$ and the remaining observations with a weight close to zero. As there are 20 observations in state one, the sum of the weights equals 1. Equally, at $\lambda_3 = 2.5$: all observations are equally weighted with a weight of 0.01. As 100 observations are in the sample, the sum of weights equals 1. The standard Markov switching weights are

independent of the parameters, $w_{\mathrm{MS},(1,1)} = 0.05$ and $w_{\mathrm{MS},(1,i)} = 0$ for $i \neq 1$, and are not included in Figure 2.1.

On the left of the graph, where $\lambda_3 = -3$, the observations from state 1 receive nearly all the weight, those from state 2 receive a small positive weight and those from state 3 a small negative weight. When $\lambda_3 = -2.5$ the weights for $s_{2t} = 1$ and $s_{3t} = 1$ are equal and close to zero. The intuition for the equal weights is that at $\lambda_2 = \lambda_3$ the DGP is essentially a two state Markov switching model and the observations for the states with equal mean receive the same weight. The large difference between the mean of state 1 and that of the other states induces a potentially large bias when using observations from the other states. As a result, the weights on observations with $s_{2t} = 1$ and $s_{3t} = 1$ are very small.

As λ_3 increases, weights for observations from state 3 increase until, at $\lambda_3 = 0$, they are equal to those for observations with $s_{1t} = 1$. That is, as the third state becomes increasingly similar to the first state, the observations are increasingly useful for forecasting. At $\lambda_3 = 0$, the first and the third state have identical means and the observations therefore receive equal weight. When λ_3 ranges between -2.5 and 0, the weights for the observations from the second state are negative. The intuition is that as the observations from the third state receive an increasingly higher weight they induce a larger bias, which is in the same direction as the bias due to the observations from the second state. By giving the observations from the second state negative weights, the biases of the observations from the second and third state are of opposite signs and can counteract each other.

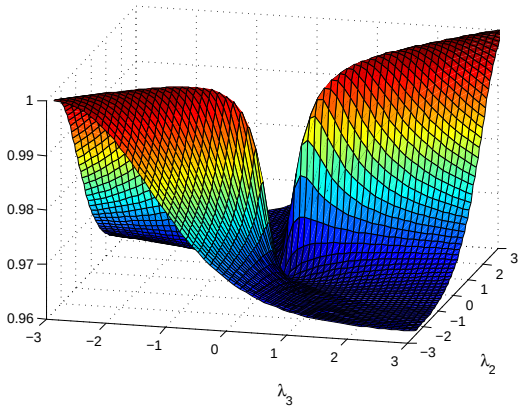
As λ_3 increases further and $0 < \lambda_3 < 2.5$, the observations from the third state are weighted heavier than the observations from the first state even though this is the future state. The reason for this at first sight surprising result is that, in this range, the means of observations from state 2 and state 3 have opposite signs. As the bias induced by the observations from the second state is, in absolute terms, larger than that from the third state, the weights on the observations from the third state receive a larger weight to counteract this bias.

At $\lambda_3 = 2.5 = -\lambda_2$ all observations receive the same weight of $\frac{1}{T}$. At this point, the mean of the observations with $s_{1t} = 1$ is between and equally distant to the means of observations with $s_{2t} = 1$ and $s_{3t} = 1$, which implies that with equal weight any biases arising from using observations of the other states cancel. In this case, the optimal weights effectively ignore the Markov switching structure of the model and forecast with equal weights, which is a very different weighting scheme from that suggested by the Markov switching model.

As in the two state case, when $s_{j,T+1} = 1$ the expected MSFE using the optimal weights is of the form

$$\mathrm{E}\left(\sigma_i^{-2} e_{T+1}^2\right)_{\text{opt}} = \frac{\sigma_j^2}{\sigma_i^2}(1 + w_{(j,j)})$$

Figure 2.2: MSFE of optimal weights relative to standard Markov switching weights



Note: The figure displays the ratio of the MSFE of the optimal weights relative to that of the standard MSFE forecast for $T = 100$, $\pi_1 = 0.2$, $\pi_2 = \pi_3 = 0.4$ for a range of values for λ_2 and λ_3.

with $w_{(j,j)}$ given in (2.16). For the Markov switching weights we have

$$\mathrm{E}\left(\sigma_i^{-2} e_{T+1}^2\right)_{\mathrm{MS}} = \frac{\sigma_j^2}{\sigma_i^2}\left(1 + \frac{1}{T\pi_j}\right)$$

Figure 2.2 displays the ratio of MSFE of the optimal weights relative to that of the standard MSFE forecast for $T = 100$, $\pi_1 = 0.2$, $\pi_2 = \pi_3 = 0.4$ for a range of values for λ_2 and λ_3. At $\lambda_2 = \lambda_3 = \pm 3$ the gains from using optimal weights are very small. In this case, the model is essentially a two state model with a large difference in mean. When λ_2 and λ_3 are of opposite sign, the improvements are the largest. We can therefore expect most gains when the observation to be forecast is in the regime with intermediate location.

m-state Markov switching models

For $s_{j,T+1} = 1$ we set $\lambda_i = \frac{\beta_i - \beta_j}{\sigma_j}$ and $q_i = \frac{\sigma_i}{\sigma_j}$, which gives for the weights for observations with $s_{l,t} = 1$

$$w_{(j,l)} = \frac{1}{T}\frac{q_l^{-2}\left(1 + T\sum_{i=1}^m q_i^{-2}\lambda_i\pi_i(\lambda_i - \lambda_l)\right)}{\sum_{i=1}^m q_i^{-2}\pi_i + T\sum_{i=1}^m\sum_{k=1}^m q_i^{-2}q_k^{-2}\pi_i\pi_k\lambda_i(\lambda_i - \lambda_k)} \tag{2.17}$$

As in the previous cases, the expected MSFE when $s_{j,T+1} = 1$ is

$$\mathrm{E}\left(\sigma_i^{-2} e_{T+1}^2\right)_{\mathrm{opt}} = \frac{\sigma_j^2}{\sigma_i^2}(1 + w_{(j,j)})$$

The derivation of the weights and the MSFE is in Appendix 2.A.1. Maximizing the expected MSFE with respect to β_j yields

$$\beta_j = \frac{\sum_{k=1}^m q_k^{-2}\pi_k\beta_k}{\sum_{k=1}^m q_k^{-2}\pi_k}$$

Hence, the largest gain occurs when the regime to be forecast is located at the probability and variance weighted average of the other regimes. The minimum MSFE is then

$$\mathrm{E}\left(\sigma_i^{-2}e_{T+1}^2\right) = \frac{1}{\sigma_i^2}\left(\sigma_j^2 + \frac{1}{T}\frac{1}{\sum_{k=1}^m \sigma_k^{-2}\pi_k}\right)$$

and when the variances are equal this reduces to

$$\mathrm{E}\left(\sigma_i^{-2}e_{T+1}^2\right) = 1 + \frac{1}{T}$$

Thus, the maximum improvement is independent of the number of states when all variances are equal.

Large T approximation

Interesting results can be obtained when considering the large sample approximation of the two state weights. The optimal weight assigned to an observation is given by

$$\begin{aligned} Tw =& s_{1,T+1}\left[\frac{1+\lambda^2 T\pi_2}{\pi_2 q^2 + \pi_1(1+\lambda^2 T\pi_2)}s_{1t} + \frac{q^2}{\pi_2 q^2 + \pi_1(1+\lambda^2 T\pi_2)}s_{2t}\right] \\ &+ s_{2,T+1}\left[\frac{1}{\pi_2 q^2 + \pi_1(1+\lambda^2 T\pi_2)}s_{1t} + \frac{q^2+\lambda^2 T\pi_1}{\pi_2 q^2 + \pi_1(1+\lambda^2 T\pi_2)}s_{2t}\right] \end{aligned}$$

We approximate this expression using that $(1+\frac{\theta}{T})^{-1} = 1 - \frac{\theta}{T} + \mathcal{O}(T^{-2})$, where $\theta = (\pi_2 q^2 + \pi_1)/(\lambda^2 \pi_2 \pi_1)$. This yields

$$\begin{aligned} Tw =& \left(\frac{1}{\pi_1} - \frac{1}{T}\frac{q^2}{\lambda^2\pi_1^2}\right)s_{1t}s_{1,T+1} + \frac{1}{T}\frac{q^2}{\lambda^2\pi_1\pi_2}s_{2t}s_{1,T+1} + \\ &+ \frac{1}{T}\frac{1}{\lambda^2\pi_1\pi_2}s_{1t}s_{2,T+1} + \left(\frac{1}{\pi_2} - \frac{1}{T}\frac{1}{\lambda^2\pi_2^2}\right)s_{2t}s_{2,T+1} + \mathcal{O}(T^{-2}) \end{aligned} \tag{2.18}$$

Hence, the standard Markov switching weights are optimal up to a first order approximation in T. It is worth noting that this is equivalent to the result obtained by Pesaran et al. (2013) for the structural break case where the first order approximation gives zero weight to pre-break observations and equally weight the post-break observations. This result in (2.18) also suggests that, in a Markov switching model, accurate estimation of the proportions of the sample in each state is of first order importance, whereas the differences in means are of second order importance to obtain a minimal MSFE. This is the motivation for considering the uncertainty around the state estimates, which we turn to now.

2.3.2 Optimal weights when states are uncertain

We will now contrast the weights conditional on the states with weights that do not assume knowledge of the states. The expectations in (2.8) can be expressed in terms of the underlying Markov chain. However, it turns out that in this case analytic expressions for the inverse of $\mathbf{M}$ cannot be obtained. In Section 2.3.3, we will show how numerical values for the inverse can be used to calculate numerical values for the optimal weights.

In order to analyze the theoretical properties of the optimal weights, analytic expressions for the weights are required, which will allow us to contrast them with the weights that are derived conditional on the states. Such expressions can be obtained by making the simplifying assumption that a time dependent expectation is available for the states of the Markov chain. Estimates of the probabilities are available as output of the estimation of Markov switching models, and this information is also used for the standard forecast from Markov switching models in (2.2). Note that this is, in fact, more general than the Markov switching model and can accommodate state probabilities from other sources, such as surveys of experts or models other than those considered here.

Denote the probability of state i occurring at time t by ξ_{it}. We assume that the expectations in (2.8) and (2.9) can be approximated as

$$\mathrm{E}(s_{it}s_{j,t+m}) = \begin{cases} \xi_{it} & \text{if } i = j \\ \xi_{it}\xi_{j,t+m} & \text{if } i \neq j, m \geqslant 0 \end{cases}$$

We will, initially, focus on the two state case and, subsequently, on m states.

Two-state Markov switching models

In a two state model, we have $\tilde{\mathbf{S}} = \mathbf{s}_2 = (s_{21}, s_{22}, \ldots, s_{2T})'$. The matrix $\mathbf{M}$ in (2.8) is given by

$$\begin{aligned} \mathbf{M} &= \lambda^2\boldsymbol{\xi}\boldsymbol{\xi}' + \lambda^2\mathbf{V} + q^2\mathbf{I} + (1-q^2)\boldsymbol{\Xi} \\ &= \lambda^2\boldsymbol{\xi}\boldsymbol{\xi}' + \mathbf{D} \end{aligned}$$

with $\boldsymbol{\xi} = (\xi_{21}, \xi_{22}, \ldots, \xi_{2T})$, $\boldsymbol{\Xi} = \mathrm{diag}(\boldsymbol{\xi})$, $\mathbf{V} = \boldsymbol{\Xi}(\mathbf{I} - \boldsymbol{\Xi})$, and $\mathbf{D} = \lambda^2\mathbf{V} + q^2\mathbf{I} + (1-q^2)\boldsymbol{\Xi}$ and again $q = \sigma_1/\sigma_2$. The inverse of $\mathbf{M}$ is

$$\mathbf{M}^{-1} = \mathbf{D}^{-1} - \frac{\lambda^2}{1 + \lambda^2\boldsymbol{\xi}'\mathbf{D}^{-1}\boldsymbol{\xi}}\mathbf{D}^{-1}\boldsymbol{\xi}\boldsymbol{\xi}'\mathbf{D}^{-1} \tag{2.19}$$

Using (2.8) and (2.19) yields

$$\mathbf{w} = \lambda^2\xi_{2,T+1}\mathbf{M}^{-1}\boldsymbol{\xi} + \frac{\mathbf{M}^{-1}\boldsymbol{\iota}}{\boldsymbol{\iota}'\mathbf{M}^{-1}\boldsymbol{\iota}}\left(1 - \lambda^2\xi_{2,T+1}\boldsymbol{\iota}'\mathbf{M}^{-1}\boldsymbol{\xi}\right) \tag{2.20}$$

Denote the typical (t,t)-element of $\mathbf{D}^{-1}$ by d_t, where

$$d_t = \left[\lambda^2 \xi_{2,t}(1-\xi_{2,t}) + q^2 + (1-q^2)\xi_{2,t}\right]^{-1}$$

Then, the weight for the observation at time t is given by

$$w_t = \frac{d_t \left[1 + \lambda^2 \sum_{t'=1}^{T} d_{t'}(\xi_{2t} - \xi_{2t'})(\xi_{2,T+1} - \xi_{2t'})\right]}{\sum_{t'=1}^{T} d_{t'} + \lambda^2 \left[\left(\sum_{t'=1}^{T} d_{t'}\xi_{2t'}^2\right)\left(\sum_{t'=1}^{T} d_{t'}\right) - \left(\sum_{t'=1}^{T} d_{t'}\xi_{2t'}\right)^2\right]} \tag{2.21}$$

The expected MSFE can be calculated from (2.6) and reduces to

$$\mathrm{E}(\sigma_2^{-2} e_{T+1}^2) = \left[1 + \lambda^2 \xi_{2,T+1}(1-\xi_{2,T+1})\right](1 + w_{T+1}) \tag{2.22}$$

where w_{T+1} is given by (2.21).

When T is large, weights (2.21) can be written as

$$w_t = \tilde{d}_t \frac{\sum_{t'=1}^{T} \tilde{d}_{t'} (\xi_{2,T+1} - \xi_{2t'})(\xi_{2t} - \xi_{2t'})}{\sum_{t'=1}^{T} \tilde{d}_{t'} \left(\xi_{2t'} - \sum_{t''=1}^{T} \tilde{d}_{t''}\xi_{2t''}\right)^2} + \mathcal{O}(T^{-2}) \tag{2.23}$$

where $\tilde{d}_t = d_t / (\sum_{t'=1}^{T} d_t)$. Derivations are provided in Appendix 2.A.2. While the weights in (2.21) and (2.23) provide closed form solutions, interpretation can be aided by momentarily making the simplifying assumption of constant state variances.

Constant state variance The interpretation of (2.21) and (2.23) is complicated by the fact that ξ_{2t} is a continuous variable in the range $[0,1]$ – as opposed to the binary variable s_{2t} for the weights conditional on states – so that an infinite number of possible combinations of ξ_{2t} over t is possible. In order to simplify the interpretation of the weights, we will therefore, for a moment, assume that the variance of the states is constant and denoted as $\sigma_s^2 = \xi_{2t}(1-\xi_{2t})$.

Summing σ_s^2 over t and solving for σ_s^2 yields

$$\sigma_s^2 = \bar{\xi}_1 \bar{\xi}_2 - \frac{1}{T}\sum_t (\xi_{2t} - \bar{\xi}_2)^2 \tag{2.24}$$

where $\bar{\xi}_1 = \frac{1}{T}\sum_{t=1}^{T} \xi_{1t}$ and $\bar{\xi}_2 = \frac{1}{T}\sum_{t=1}^{T} \xi_{2t}$. Note that the maximum value of σ_s^2 is given by $\bar{\xi}_2\bar{\xi}_1$, which occurs when the probability vector is constant. In the case of a constant σ_s^2, $\tilde{d}_t$ simplifies to $1/T$. Hence, (2.21) can be written as

$$w_t = \frac{1}{T}\left[1 + \lambda^2 \frac{(\xi_{2,T+1} - \bar{\xi}_2)(\xi_{2t} - \bar{\xi}_2)}{(T\bar{d})^{-1} + \lambda^2(\bar{\xi}_1\bar{\xi}_2 - \sigma_s^2)}\right]$$

and the large T approximation (2.23) as

$$w_t = \frac{1}{T} + \frac{(\xi_{2,T+1} - \bar{\xi}_2)(\xi_{2t} - \bar{\xi}_2)}{T(\bar{\xi}_1\bar{\xi}_2 - \sigma_s^2)} \tag{2.25}$$

The standard Markov switching weights can be expressed as

$$w_{\text{MS},t} = \frac{1}{T} + \frac{(\xi_{2,T+1} - \bar{\xi}_2)(\xi_{2t} - \bar{\xi}_2)}{T\bar{\xi}_1\bar{\xi}_2} \tag{2.26}$$

see Appendix 2.A.2 From a comparison of (2.25) and (2.26) it is clear that the two weights differ by the factor σ_s^2 in the denominator and that this difference will not disappear asymptotically. Effectively, the Markov switching weights are more conservative as the optimal weights exploit the regime switching structure more strongly because of the smaller denominator in (2.25) compared to (2.26).

The MSFE for the optimal weights and for the standard Markov switching weights under constant state variance are

$$\begin{aligned} \mathrm{E}\left(\sigma_2^{-2}e_{T+1}^2\right)_{\text{opt}} &= \left[1 + \lambda^2\xi_{2,T+1}(1 - \xi_{2,T+1})\right] \\ &\quad \times \left(1 + \frac{1}{T} + \frac{\lambda^2(\xi_{2,T+1} - \bar{\xi}_2)^2}{1 + \lambda^2\sigma_s^2 + \lambda^2 T(\bar{\xi}_2(1 - \bar{\xi}_2) - \sigma_s^2)}\right) \end{aligned} \tag{2.27}$$

$$\begin{aligned} \mathrm{E}\left(\sigma_2^{-2}e_{T+1}^2\right)_{\text{MS}} &= 1 + \lambda^2\xi_{2,T+1}(1 - \xi_{2,T+1}) + \frac{1}{T}(\lambda^2\sigma_s^2 + 1) \\ &\quad + \left(\frac{\xi_{2,T+1} - \bar{\xi}_2}{\bar{\xi}_2(1 - \bar{\xi}_2)}\right)^2 \left[\frac{1}{T}(\bar{\xi}_2(1 - \bar{\xi}_2) - \sigma_s^2)\left(\lambda^2\sigma_s^2 + 1\right) + \lambda^2\sigma_s^4\right] \end{aligned} \tag{2.28}$$

The MSFE for the optimal weights is derived from (2.22) by substituting in the weights in (2.21) and using the fact that $\tilde{d}_t = 1/T$ and $d_t = d$, for $t = 1, 2, \ldots, T + 1$, which together with the MSFE for the standard Markov switching weights is derived in Appendix 2.A.2.

Table 2.2 displays the improvements in forecast performance expressed as the ratio of (2.27) over (2.28) for different values of $\bar{\xi}_2$, $\tilde{\sigma}_s^2 = \sigma_s^2/(\bar{\xi}_2\bar{\xi}_1)$ and λ for $T = 100$. The results indicate that the optimal weights lead to larger gains when λ is large and when $\bar{\xi}_2$ is closer to 0.5. The influence of σ_s^2 is U-shaped with the largest improvement when $\sigma_s^2 = 0.6$. The results in Table 2.2 show that the improvement can be as large as 11.3% for the range of parameter values considered here.

In this simplified framework, the increase in forecast accuracy does not disappear when the sample size increases. The asymptotic approximation to the MSFE under optimal weights is given by

$$\mathrm{E}\left(\sigma_0^2 e_{T+1}^2\right)_{\text{opt}} = 1 + \lambda^2\xi_{2,T+1}(1 - \xi_{2,T+1}) + \mathcal{O}(T^{-1}) \tag{2.29}$$

Table 2.2: Maximum improvements in a two state model with $T = 100$

		$\bar{\xi}_2$				
	$\tilde{\sigma}_s^2$	0.1	0.2	0.3	0.4	0.5
$\lambda = 2$	0	1.000	1.000	1.000	1.000	1.000
	0.2	0.993	0.986	0.981	0.979	0.978
	0.4	0.977	0.960	0.950	0.944	0.942
	0.6	0.967	0.946	0.934	0.927	0.926
	0.8	0.974	0.957	0.948	0.944	0.942
$\lambda = 3$	0	1.000	1.000	1.000	1.000	1.000
	0.2	0.982	0.969	0.962	0.958	0.957
	0.4	0.951	0.926	0.913	0.907	0.905
	0.6	0.935	0.908	0.895	0.889	0.887
	0.8	0.949	0.930	0.921	0.917	0.916

Note: The table reports the ratio of the MSFE of the optimal weights to that of the Markov switching weights conditional on a constant state variance σ_s^2. $\lambda = (\beta_2 - \beta_1)/\sigma$ denotes the scaled difference between means, $\bar{\xi}_2$ the average probability for state 2, and $\tilde{\sigma}_s^2$ is a negative function of the variance of the state 2 probability.

and that under standard Markov switching weights is

$$\mathrm{E}\left(\sigma_0^2 e_{T+1}^2\right)_{\mathrm{MS}} = 1 + \lambda^2 \xi_{2,T+1}(1 - \xi_{2,T+1}) + \left(\frac{\xi_{2,T+1} - \bar{\xi}_2}{\bar{\xi}_2 \bar{\xi}_1}\right)^2 \lambda^2 \sigma_s^4 + \mathcal{O}(T^{-1}) \tag{2.30}$$

The difference between (2.30) and (2.29) is positive and does not disappear asymptotically. The relative improvement is expected to be high when λ, σ_s^2, and the difference $\xi_{2,T+1} - \bar{\xi}_2$ are large.

m-state Markov switching models

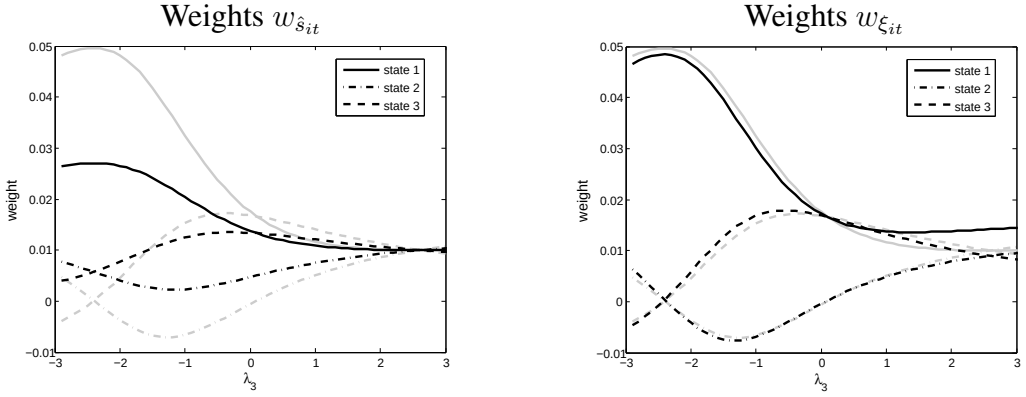
The derivations can be extended to an arbitrary number of states. Note that $\mathbf{M} = \mathrm{E}(\mathbf{Q}) + \mathrm{E}(\tilde{\mathbf{S}}'\boldsymbol{\lambda}\boldsymbol{\lambda}'\tilde{\mathbf{S}})$ and $\mathrm{E}(\tilde{\mathbf{S}}'\boldsymbol{\lambda}\boldsymbol{\lambda}'\tilde{\mathbf{S}}) = \mathrm{E}(\tilde{\mathbf{S}})'\boldsymbol{\lambda}\boldsymbol{\lambda}'\mathrm{E}(\tilde{\mathbf{S}}) + \mathbf{A}$ where, conditional on the state probabilities, ξ_{jt}, $j = 1, 2, \ldots, m$,

$$\mathbf{A} = \sum_{j=2}^{m} \lambda_j^2 \boldsymbol{\Xi}_j - \left(\sum_{j=2}^{m} \lambda_j \boldsymbol{\Xi}_j\right)^2$$

and $\boldsymbol{\Xi}_j$ is a $T \times T$ diagonal matrix with typical element ξ_{jt}. Define $\tilde{\boldsymbol{\xi}} = \mathrm{E}(\tilde{\mathbf{S}})'\boldsymbol{\lambda}$, which is a $T \times 1$ vector, and $\mathbf{D} = \mathrm{E}(\mathbf{Q}) + \mathbf{A}$. Then,

$$\mathbf{M}^{-1} = \mathbf{D}^{-1} - \frac{1}{1 + \tilde{\boldsymbol{\xi}}\mathbf{D}^{-1}\tilde{\boldsymbol{\xi}}}\mathbf{D}^{-1}\tilde{\boldsymbol{\xi}}\tilde{\boldsymbol{\xi}}'\mathbf{D}^{-1}$$

Figure 2.3: Optimal weights for a three state Markov switching model



Note: In both plots, the lighter, gray lines depict optimal weights (2.16), which are conditional on the states, for a representative observation in each state. In the left plot, the darker lines are the optimal weights (2.16) for a representative observation in each state where the probabilities are used as plug-in values for the states. In the right plot, the darker lines are the weights (2.31) that are derived conditional on the states under state probabilities $\hat{\boldsymbol{\xi}}_{T+1} = [0.8, 0.1, 0.1]'$ for $\lambda_2 = -2.5$, λ_3 over the range -3 to 3, $T = 100$, $\pi_1 = 0.2$, and $\pi_2 = \pi_3 = 0.4$. The dark, solid line gives the weights when $\hat{\boldsymbol{\xi}}_t = [0.8, 0.1, 0.1]'$, the dark, dash-dotted line when $\hat{\boldsymbol{\xi}}_t = [0.1, 0.8, 0.1]'$, and the dark, dashed line when $\hat{\boldsymbol{\xi}}_t = [0.1, 0.1, 0.8]'$.

We can use (2.8) to derive the weights similar to the case of the two-state weights

$$w_t = \frac{d_t^{(m)} \left\{ 1 + \left[\sum_{t'=1}^{T} d_{t'}^{(m)} (\tilde{\xi}_t - \tilde{\xi}_{t'})(\tilde{\xi}_{T+1} - \tilde{\xi}_{t'}) \right] \right\}}{\sum_{t'=1}^{T} d_{t'}^{(m)} + \left(\sum_{t'=1}^{T} d_{t'}^{(m)} \tilde{\xi}_{t'}^2 \right) \left(\sum_{t'=1}^{T} d_{t'}^{(m)} \right) - \left(\sum_{t'=1}^{T} d_{t'}^{(m)} \tilde{\xi}_{t'} \right)^2} \tag{2.31}$$

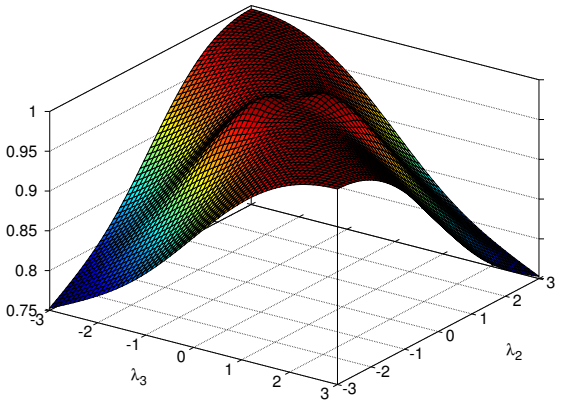
where

$$\begin{aligned} d_t^{(m)} &= \left[\sum_{j=1}^{m} q_j^2 \xi_{jt} + \sum_{j=2}^{m} \lambda_j^2 \xi_{jt} - \left(\sum_{j=2}^{m} \lambda_j \xi_{jt} \right)^2 \right]^{-1} \\ &= \left[\sum_{j=1}^{m} (q_j^2 + \lambda_j^2) \xi_{jt} - \left(\sum_{j=2}^{m} \lambda_j \xi_{jt} \right)^2 \right]^{-1} \\ \tilde{\xi}_t &= \sum_{j=2}^{m} \xi_{jt} \lambda_j \end{aligned}$$

given that $\lambda_1 = 0$.

Examples of weights for a three state Markov switching model when states are uncertain are plotted in Figure 2.3. Again, the difference in mean between the first and second state relative to the variance of the first state is set to $\lambda_2 = -2.5$, and the difference in mean between the first and third state, λ_3, varies from -3 to 3. Furthermore, $\pi_1 = 0.2$, $\pi_2 = \pi_3 = 0.4$, $T = 100$, and the ratio of variances is $q_1 = q_2 = 1$. For simplicity of exposition, we assume that the state probabilities are identical for each state in the sense that a prevailing

Figure 2.4: MSFE of optimal weights relative to standard weights when states are uncertain



Note: The figure displays the ratio of the MSFE of the optimal weights relative to that of the standard MSFE forecast. For details of the parameter settings see the footnote of Figure 2.3.

state is given probability $\xi_{it} = 0.8$ and other states $\xi_{jt} = 0.1$. The light gray lines represent the optimal weights (2.16) that are conditional on the states. The graph on the left plots weights (2.16) substituting the probabilities ξ_{it} for the states s_{it}, that is, the plug-in estimator of the weights as the black lines. The graph on the right plots the weights (2.31) as the black lines.

The graph on the left shows how the introduction of the probabilities brings the weights closer to equal weighting compared to the weights for known states. This contrasts with the weights that explicitly take the uncertainty around the states into account. In the plot on the right these weights are very close to the weights conditional on the states. Hence, using the uncertainty of the states in the derivation of the weights leads to weights that are similar to when the states are known.

An additional difference arises for positive λ_3, where the weights conditional on state probabilities for the future state increase over those conditional on states. The reason is that for λ_2 and λ_3 of opposite sign, the variance of $\boldsymbol{\iota}'\tilde{\boldsymbol{\xi}}$ increases relative to the case of λ's of equal sign, which affects $d_t^{(m)}$ in (2.31). Hence, the increase of uncertainty about the states leads to an increased reliance on the data that are likely from same state as the future observation.

The relative MSFE of optimal relative to standard weights is displayed in Figure 2.4. When λ_2 and λ_3 are large and of similar magnitude, optimal weights have a much smaller MSFE as the standard weights are compressed due to the uncertainty around the states. When λ_2 and λ_3 are of opposite signs, the gain is smaller as the compression of the standard weights brings them closer to the optimal weights, which for $\lambda_2 = -\lambda_3$ are equal weights.

2.3.3 Estimating state covariances from the data

Above, we derived weights conditional on the state probabilities, in which case we can write the expectation of the product of two states as $\mathrm{E}(s_{it}s_{j,t+m}) = \xi_{it}\xi_{j,t+m}$. While this assumption allows us to find an explicit inverse of the matrix $\mathbf{M}$ and to obtain analytic expressions for the weights, it does not use the Markov switching nature of the DGP. If one is willing to lose the convenience of explicit expressions for the weights, it is possible to estimate $\mathbf{M}$ directly from the data.

To estimate $\mathbf{M}$ directly from the data, we now condition on the information set up to time T, denoted Ω_T. Then $\mathrm{E}(s_{it}s_{j,t+m}|\Omega_T) = p(s_{j,t+m} = 1|\Omega_T)p(s_{it} = 1|s_{j,t+m} = 1, \Omega_T)$. The first term is the smoothed probability of being in state j at time $t + m$ as given by an EM-algorithm Hamilton (1994) or a MCMC sampler Kim and Nelson (1999). The second term can be written as

$$p(s_{it} = 1|s_{j,t+m} = 1, \Omega_T) = \frac{\xi^i_{t|t}}{\xi^j_{t+m|t+m-1}} \left[\left(\prod_{l=1}^{m-1} \mathbf{P}'\mathbf{A}_{t+l} \right) \mathbf{P}' \right]_{i,j} \tag{2.32}$$

where $\mathbf{A}_t$ is a $m \times m$ diagonal matrix with typical i, i-element $\xi_{it|t}/\xi_{it|t-1}$, and $\xi_{it|t}$ and $\xi_{it|t-1}$ denote the filtered and forecast probabilities of state i at time t. The derivation of (2.32) can be found in Appendix 2.A.2. Using these expressions we can calculate the expectations in (2.8). Define

$$\mathbf{\Xi}^* = \left[\left(\prod_{l=1}^{k-1} \mathbf{P}'\mathbf{A}_{t+l} \right) \mathbf{P}' \right]_{2:m,2:m}$$

Then we can write $m - 1 \times m - 1$ matrix of expectations

$$\mathrm{E}\left(\tilde{\mathbf{s}}_t\tilde{\mathbf{s}}'_{t+k} \middle| \mathbf{\Omega}_T \right) = \mathbf{\Xi}_{t|t}\mathbf{\Xi}^* \left(\mathbf{\Xi}_{t+k|T} \div \mathbf{\Xi}_{t+k|t+k-1} \right)$$

where $\mathbf{\Xi}_{t|t}$ is an $m - 1 \times m - 1$ matrix with typical i, i element $\hat{\xi}_{it|t}$ is, and $\div$ denotes element-by-element division. Recall $\mathbf{M} = \mathrm{E}(\mathbf{Q}) + \mathrm{E}(\tilde{\mathbf{S}}'\boldsymbol{\lambda}\boldsymbol{\lambda}'\tilde{\mathbf{S}})$. A typical element of the second matrix is given by

$$\begin{aligned} \mathrm{E}\left(\tilde{\mathbf{S}}'\boldsymbol{\lambda}\boldsymbol{\lambda}'\tilde{\mathbf{S}} \middle| \mathbf{\Omega}_T \right)_{t,t} &= \boldsymbol{\lambda}'\mathrm{diag}\left[\mathrm{E}\left(\tilde{\mathbf{s}}_t \middle| \mathbf{\Omega}_T \right) \right] \boldsymbol{\lambda} \\ \mathrm{E}\left(\tilde{\mathbf{S}}'\boldsymbol{\lambda}\boldsymbol{\lambda}'\tilde{\mathbf{S}} \middle| \mathbf{\Omega}_T \right)_{t,t+k} &= \boldsymbol{\lambda}'\mathrm{E}\left(\tilde{\mathbf{s}}_t\tilde{\mathbf{s}}'_{t+k} \middle| \mathbf{\Omega}_T \right) \boldsymbol{\lambda} \end{aligned} \tag{2.33}$$

Using (2.33) in (2.8) yields numerical solutions for the weights.

2.4 Markov switching models with exogenous regressors

So far, we have considered models that only contain a constant as the regressor. Now, we return to the model with regressors in (2.1). Rewrite this model as

$$\begin{aligned}\mathbf{y} &= \sum_{i=1}^{m} \mathbf{S}_i(\mathbf{X}\boldsymbol{\beta}_i + \sigma_i\boldsymbol{\varepsilon}) \\ &= \mathbf{X}\boldsymbol{\beta}_1 + \sum_{i=1}^{m} \mathbf{S}_i\mathbf{X}(\boldsymbol{\beta}_i - \boldsymbol{\beta}_1) + \sum_{i=1}^{m} \mathbf{S}_i\sigma_i\boldsymbol{\varepsilon}\end{aligned}$$

where $\mathbf{S}_i$ is a $T \times T$ matrix with as its j-th diagonal element equal to one if observation j belongs to state i and zero elsewhere, $\mathbf{X}$ a $T \times k$ matrix of exogenous regressors and $\boldsymbol{\beta}_i$ a $k \times 1$ vector of parameters, σ_i the variance of regime i, and we used the fact that $\sum_{i=1}^{m} \mathbf{S}_i = \mathbf{I}$. Also,

$$y_{T+1} = \mathbf{x}'_{T+1}\boldsymbol{\beta}_1 + \sum_{i=2}^{m} s_{i,T+1}\mathbf{x}'_{T+1}(\boldsymbol{\beta}_i - \boldsymbol{\beta}_1) + \sum_{i=1}^{m} s_{i,T+1}\sigma_i\varepsilon_{T+1}$$

As before, we define the optimally weighted estimator as follows

$$\boldsymbol{\beta}(\mathbf{w}) = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{y}$$

where $[\boldsymbol{W}]_{ii} = w_i$ and $[\boldsymbol{W}]_{ij} = 0$ if $i \neq j$. The optimal forecast is then given by $\hat{y}_{T+1} = \mathbf{x}'_{T+1}\boldsymbol{\beta}(\mathbf{w})$.

Define $\boldsymbol{\lambda}_i = (\boldsymbol{\beta}_i - \boldsymbol{\beta}_1)/\sigma_m$, $q_i = \sigma_i/\sigma_m$ and $\boldsymbol{\Lambda}_{ij} = \boldsymbol{\lambda}_i\boldsymbol{\lambda}'_j$. As in the case of structural breaks analyzed by Pesaran et al. (2013), large sample approximations to the MSFE are necessary to obtain analytical expressions for the weights. We make the following approximations: $\operatorname{plim}_{T\to\infty} \mathbf{X}'\mathbf{W}\mathbf{X} = \boldsymbol{\Omega}_{XX}$, $\operatorname{plim}_{T\to\infty} \mathbf{X}'\mathbf{S}_i\mathbf{W}\mathbf{X} = \boldsymbol{\Omega}_{XX}\mathbf{w}'\mathbf{s}_i$, $\operatorname{plim}_{T\to\infty} \mathbf{X}'\mathbf{W}^2\mathbf{S}_i\mathbf{X} = \boldsymbol{\Omega}_{XX}\mathbf{w}'\mathbf{S}_i\mathbf{w}$. Then, the MSFE is

$$\begin{aligned}\mathrm{E}\left(\sigma_m^{-2}e_{T+1}^2\right) = &\sum_{i=1}^{m} \mathrm{E}(s_{i,T+1})\mathbf{x}'_{T+1}\boldsymbol{\Lambda}_{ij}\mathbf{x}_{T+1} + \sum_{i=1}^{m} \mathrm{E}(s_{i,T+1})q_i^2\varepsilon_{T+1}^2 \\ &+ \sum_{i=1}^{m}\sum_{j=1}^{m} \mathbf{w}'\mathrm{E}(\mathbf{s}_i\mathbf{s}'_j)\mathbf{w}\boldsymbol{\Lambda}_{ij}\mathbf{x}_{T+1} + \mathbf{x}'_{T+1}\boldsymbol{\Omega}_{XX}^{-1}\sum_{i=1}^{m} q_i^2\mathbf{w}'\mathrm{E}(\mathbf{S}_i)\mathbf{w}\mathbf{x}_{T+1} \\ &- 2\mathbf{x}'_{T+1}\sum_{i=1}^{m}\sum_{j=1}^{m} \mathbf{w}'\mathrm{E}(\mathbf{s}_i s_{j,T+1})\boldsymbol{\Lambda}_{ij}\mathbf{x}_{T+1}\end{aligned}$$

Maximizing (2.4) subject to $\boldsymbol{\iota}'\mathbf{w} = 1$ leads to the following optimal weights

$$\mathbf{w} = \mathbf{M}^{-1}\mathrm{E}\left(\tilde{\mathbf{S}}'\boldsymbol{\phi}\boldsymbol{\phi}'\tilde{\mathbf{s}}_{T+1}\right) + \frac{\mathbf{M}^{-1}\boldsymbol{\iota}}{\boldsymbol{\iota}'\mathbf{M}^{-1}\boldsymbol{\iota}}\left[1 - \boldsymbol{\iota}'\mathbf{M}^{-1}\mathrm{E}\left(\tilde{\mathbf{S}}'\boldsymbol{\phi}\boldsymbol{\phi}'\tilde{\mathbf{s}}_{T+1}\right)\right] \tag{2.34}$$

where $\boldsymbol{\phi}_i = \mathbf{x}'_{T+1}\boldsymbol{\lambda}_i/(\mathbf{x}_{T+1}\boldsymbol{\Omega}_{XX}^{-1}\mathbf{x}_{T+1})^{1/2}$, $\mathbf{M} = \mathrm{E}(\mathbf{Q}) + \mathrm{E}\left(\tilde{\mathbf{S}}'\boldsymbol{\phi}\boldsymbol{\phi}'\tilde{\mathbf{S}}\right)$ and $\mathbf{Q}$ a diagonal matrix with typical (t,t)-element $Q_{tt} = \sum_{i=1}^{m} q_i^2 s_{it}$. The results derived for the location model above can, therefore, be straightforwardly extended to allow for exogenous regressors by replacing $\boldsymbol{\lambda}$ with $\boldsymbol{\phi}$.

2.5 Evidence from Monte Carlo experiments

2.5.1 Set up of the experiments

We analyze the forecast performance of the optimal weights in a series of Monte Carlo experiments. Data are generated according to (2.1) and we consider models with with $m = 2$ and $m = 3$ states. We set $\sigma_2^2 = 0.25$ and use a range of values for λ_i and q^2.

The states are generated by a Markov chain with transition probabilities $p_{ij} = \frac{1}{T\pi_i}$, for $i \neq j$, and ergodic probabilities $\pi_i = \pi = 1/m,\ \forall i$, where m is the number of states. The diagonal elements of the transition probability matrix are $p_{ii} = 1 - \sum_{j=1}^{m} p_{ij}$. This creates Markov chains with relatively high persistence. The first state is sampled from the ergodic probability vector, $\mathbf{s}_1 \sim \text{Binomial}(1, \boldsymbol{\pi})$ and subsequent states are drawn as $\mathbf{s}_t \sim \text{Binomial}(1, \mathbf{p}_t)$ where $\mathbf{p}_t = \mathbf{P}\mathbf{s}_{t-1}$. We restrict attention to draws of the data that would be identified as Markov switching models in an application: we require that each regime has at least 10 observations and that regimes are identified empirically in that $\sum_{t=1}^{T} \hat{\xi}_{t|T}^{i} \geqslant 5$, $\forall i$, which ensures identification of the parameters. The estimation uses the EM algorithm (Dempster et al. 1977) as outlined by Hamilton (1994).

The first set of the Monte Carlo experiments analyzes two state models with a constant only, that is, $k = 1$ and $x_t = 1$ for $T = 200$. A second set of experiments considers three state models for $T = 200$. We also ran experiments for a two-state model with an exogenous regressor. The results do not substantially differ from the results of the mean only model and can be found in Appendix 2.B.

Given the parameter estimates $\hat{\boldsymbol{\beta}}_i$, $\hat{\mathbf{P}}$, $\hat{\sigma}_i$ and the probability vectors with $\hat{\xi}_{t|T}$, $\hat{\xi}_{t|t}$, $\hat{\xi}_{t|t-1}$ we construct the usual Markov switching forecast as

$$\hat{y}_{T+1}^{\text{MS}} = \mathbf{x}'_{T+1} \sum_{i=1}^{m} \hat{\beta}_i \hat{\xi}_{T+1|T}^{i}$$

where $\hat{\beta}_i$ is given in (2.2).

The optimal weights are calculated as outlined in the sections above. The following notation is used to distinguish the different weights:

- $w_{\hat{s}}$: weights based on known states, operationalized by substituting the smoothed probability vector $\hat{\boldsymbol{\xi}}_{t|T}$ for the states as discussed in Section 2.3.1.

- $w_{\hat{\xi}}$: weights derived based on state probabilities, with the smoothed probability vector $\hat{\boldsymbol{\xi}}_{t|T}$ as the probabilities as discussed in Section 2.3.2.
- $w_{\hat{\mathbf{M}}}$: the weights based on state probabilities derived by directly estimating the matrix $\hat{\mathbf{M}}$ as detailed in Section 2.3.3.

Using these weights, the optimal forecast is constructed as

$$\hat{y}_{T+1}^{\text{opt}} = \mathbf{x}'_{T+1} \left(\mathbf{X}'\mathbf{W}\mathbf{X}\right)^{-1} \mathbf{X}'\mathbf{W}\mathbf{y}$$

where $\mathbf{W}$ is a diagonal matrix with typical diagonal element $w_{\hat{s},t}$, $w_{\hat{\xi},t}$, or $w_{\hat{\mathbf{M}},t}$.

We report ratios of the MSFE of optimally weighted forecasts to that of standard Markov switching forecasts. Additionally, we separated the results by the size of the regime difference, λ_i. Finally, we have seen above that the performance of the weights $w_{\hat{\xi}}$ depends on the variance of the smoothed probability vector. Thus, we separate the results based on the normalized variance of the smoothed probability vector

$$\tilde{\sigma}^2_{\hat{\xi}} = \frac{\frac{1}{T}\sum_{t=1}^{T} \hat{\xi}^{(i)}_{t|T}(1 - \hat{\xi}^{(i)}_{t|T})}{\frac{1}{T}\sum_{t=1}^{T} \hat{\xi}^{(i)}_{t|T} \frac{1}{T}\sum_{t=1}^{T}(1 - \hat{\xi}^{(i)}_{t|T})} \tag{2.35}$$

where i the state which has the minimum normalized variance. Note that in the case of two states for $\frac{1}{T}\sum_{t=1}^{T} \hat{\xi}^{(i)}_{t|T} = \frac{1}{T}\sum_{t=1}^{T}(1 - \hat{\xi}^{(i)}_{t|T}) = 0.5$, the measure $\tilde{\sigma}^2_{\hat{\xi}}$ is analogous to the regime classification measure (RCM) of Ang and Bekaert (2002). The Monte Carlo results are from 10,000 replications.

2.5.2 Monte Carlo results

The Monte Carlo results for the two-state model are reported in Table 2.3, where results for models with switches in mean and homoskedastic errors are in the left panel. The results in Section 2.3.1 suggest that forecasts from optimal weights conditional on states, $w_{\hat{s}}$, will show the largest gains when the difference between regimes, λ, is small. In contrast, the results in Section 2.3.2 suggest that the gains for the forecasts from optimal weights conditional on state probabilities, $w_{\hat{\xi}}$ and $w_{\hat{\mathbf{M}}}$, will be largest for large λ, which is the practically more relevant case.

The results from the simulation confirm the theoretical findings. For small λ, the forecasts from weights, $w_{\hat{s}}$, are more precise than those using standard weights and weights conditional on state probabilities. An additional effect that improves the forecasts using $w_{\hat{s}}$ is that the parameter estimates are biased upwards when $\lambda = 1$. In Section 2.3.1, we show that the weights $w_{\hat{s}}$ are shrunk towards equal weights. The upwards bias of $\hat{\lambda}$ will return the weights closer to the infeasible optimal weights based on the true DGP. The estimated

Table 2.3: Monte Carlo results: two states, mean only models

		$q^2 = 1$			$q^2 = 2$		
λ	$\tilde{\sigma}^2_{\hat{\xi}_{\vert T}}$	$w_{\hat{s}}$	$w_{\hat{\xi}}$	$w_{\hat{\mathbf{M}}}$	$w_{\hat{s}}$	$w_{\hat{\xi}}$	$w_{\hat{\mathbf{M}}}$
1	0.0-0.1	0.997	1.004	1.008	0.998	1.002	1.002
	0.1-0.2	1.000	1.007	1.023	1.000	1.004	1.012
	0.2-0.3	1.000	1.009	1.027	1.000	1.010	1.023
	0.3-0.4	1.001	1.009	1.030	1.001	1.008	1.022
2	0.0-0.1	1.000	1.000	1.024	1.000	1.001	1.018
	0.1-0.2	1.002	0.989	1.024	1.001	0.997	1.034
	0.2-0.3	1.003	0.966	0.998	1.003	0.984	1.011
	0.3-0.4	1.004	0.940	0.967	1.002	0.983	1.004
3	0.0-0.1	1.000	0.999	1.025	1.000	0.999	1.027
	0.1-0.2	1.004	0.959	0.990	1.003	0.975	1.013
	0.2-0.3	1.005	0.903	0.953	1.005	0.950	0.988
	0.3-0.4	1.003	0.845	0.921	1.006	0.889	0.918

Note: The table reports the ratio of the MSFE of the optimal weights to that of the Markov switching weights. $y_t = \beta_1 s_{1t} + \beta_2 s_{2t} + (\sigma_1 s_{1t} + \sigma_2 s_{2t})\varepsilon_t$ where $\varepsilon_t \sim \mathrm{N}(0,1)$, $\sigma_2^2 = 0.25$, $q^2 = \sigma_1^2/\sigma_2^2$. Column labels: $\lambda = (\beta_2 - \beta_1)/\sigma_2$, $\tilde{\sigma}^2_{\hat{\xi}_{\vert T}}$ is the normalized variance in of the smoothed probability vector (2.35). $w_{\hat{s}}$: forecasts from weights based on estimated parameters and state probabilities. $w_{\hat{\xi}}$: forecasts from weights conditional on state probabilities. $w_{\hat{\mathbf{M}}}$ are the weights based on numerically inverting $\hat{\mathbf{M}}$. The sample size is $T = 200$ and the results are from $R = 10000$ repetitions.

weights conditional on state probabilities are close to the infeasible optimal weights in the absence of a bias, and the bias in $\hat{\lambda}$ will increase them beyond the infeasible optimal weights. The case of $\lambda = 1$ may, however, not be recognized in a given time series as the switches are as large as the disturbance standard deviation. This setting is, therefore, of less practical relevance than those with larger λ.

For larger λ the ordering is reversed: the forecasts from optimal weights conditional on states, $w_{\hat{s}}$, are less precise than those of the standard weights. In contrast, the weights conditional on state probabilities, $w_{\hat{\xi}}$ and $w_{\hat{\mathbf{M}}}$, are substantially more precise. The reason is that in this settings there is a smaller, at times even downwards, bias in $\hat{\lambda}$ and the shrinking of the weights $w_{\hat{s}}$ towards equal weights deteriorates the forecasts, whereas the weights conditional on the states benefit from the fact that the weights are close to the infeasible optimal weights.

The theoretical results in Section 2.3.2 suggest that the relative performance of the weights based on state probabilities, $w_{\hat{\xi}}$ and $w_{\hat{\mathbf{M}}}$, will increase in the uncertainty around the states. This is because the standard weights and the plug-in weights, $w_{\hat{s}}$, are compressed towards

Table 2.4: Monte Carlo results: three states, intercept only models

$\{\lambda_{31}, \lambda_{21}\}$	$\tilde{\sigma}^2_{\hat{\xi}_{\vert T}}$	$w_{\hat{s}}$	$w_{\hat{\xi}}$	$w_{\hat{\mathbf{M}}}$
{2,1}	0.0-0.1	0.999	1.014	1.027
	0.1-0.2	1.000	1.010	1.024
	0.2-0.3	1.001	1.007	1.031
	0.3-0.4	1.001	0.999	1.019
{3,1}	0.0-0.1	1.000	1.004	1.025
	0.1-0.2	1.001	0.989	1.019
	0.2-0.3	1.002	0.958	0.969
	0.3-0.4	1.002	0.938	0.952
{3.5,2}	0.0-0.1	1.000	1.001	1.024
	0.1-0.2	1.001	0.983	1.021
	0.2-0.3	1.002	0.954	0.960
	0.3-0.4	1.003	0.902	0.918

Note: The table reports the ratio of the MSFE of the optimal weights to that of the Markov switching weights for $q^2 = 1$. For details see Table 2.3.

equal weights whereas the optimal weights retain the shape of the weights as if the states where known. Again, the results in Table 2.3 confirm the finding: the results for weights $w_{\hat{s}}$ are worse when the states are uncertain, the forecasts from the weights conditional on state probabilities improve substantially and lead to large gains. Our application will highlight the practical relevance of large λ and state uncertainty, so that we can expect large gains when using $w_{\hat{\xi}}$ and $w_{\hat{\mathbf{M}}}$ in practice.

The sample size is the final factor that influences the performance of the forecasts, where weights $w_{\hat{\xi}}$ and $w_{\hat{\mathbf{M}}}$ improve with the sample size while weights, $w_{\hat{s}}$, deteriorate in the samples size. However, this affect is relevant only for small T such as $T = 50$, which are unlikely to be relevant in practice. Results for $T = 50$ and 100 can be found in Appendix 2.B.

The right panel of Table 2.3 reports the results for a model with state dependent mean and variance, where the variance in regime 2 is the same as before but the variance in regime 1 is doubled. This should mute the improvements since the average difference in regimes standardized by the variance decreases. While this decrease is indeed observed, substantial improvements remain in the same parameter regions where the weights under constant variance perform well.

Finally, we investigate forecasts from three state models. The results in Table 2.4 suggest that the conclusions from two state models carry over to three state models. Sizable improvements are made when using $w_{\hat{\xi}}$ and $w_{\hat{\mathbf{M}}}$ when $\tilde{\sigma}^2_{\hat{\xi}}$ is large and both differences in parameters, λ_{21} and λ_{31}, are large.

Overall, the findings from the Monte Carlo experiments suggest that optimal weights conditional on states, $w_{\hat{s}}$, work well only for small differences in regimes and when states are estimated with great certainty, which are, arguably, less realistic assumptions in practice. In contrast, optimal weights conditional on state probabilities improve forecasts over standard weights when differences in regimes and uncertainty around states are large, which is the setting most likely found in applications, such as that in Section 2.6. Using weights that treat states as independent binary variables, $w_{\hat{\xi}}$, avoids the estimation uncertainty around covariances of the state, and in many settings leads to the most precise forecasts. Estimating the full matrix of second moments, $\mathbf{M}$, in the construction of the optimal weights, $w_{\hat{\mathbf{M}}}$, can, however, improve forecasts when the difference between regimes is large while the uncertainty about regimes remains large, too.

2.6 Application to US GNP

The US business cycle, which was analyzed by Hamilton (1989), arguably remains one of the most prominent application of Markov switching models. Different variants of such models have been used to model US GNP growth, see, for example, Clements and Krolzig (1998) and Krolzig (1997, 2000). These authors also show that the Markov switching model is frequently outperformed in terms of MSFE by AR models. We use a pseudo-out-of-sample forecast exercise to investigate whether optimal weights improve the forecast accuracy of Markov switching models for US GNP growth, and whether optimal weights improve the forecast accuracy of Markov switching models over that of linear alternatives.

The model by Hamilton (1989) is an example of a Markov Switching in mean model with non-switching autoregressive regressors. This class of models

$$y_t = \beta_{s_t} + \sum_{i=1}^{p} \phi_i (y_{t-i} - \beta_{s_{t-i}}) + \sigma \varepsilon_t$$

is denoted as MSM(m)-AR(p) by Krolzig (1997), where Hamilton's model takes $m = 2$ and $p = 4$. Here, y_t depends on the current state and on the previous p states. If, in addition, the model contains a state dependent variance, σ_{s_t}, it is denoted as MSMH(m)-AR(p).

Clements and Krolzig (1998) find that a three state model with switching intercept instead of switching mean and a state dependent variance does well in terms of business cycle description and forecast performance. This class of models

$$y_t = \beta_{s_t} + \sum_{i=1}^{p} \phi_i y_{t-i} + \sigma_{s_t} \varepsilon_t$$

is denoted as MSIH(m)-AR(p) by Krolzig (1997) and the model in Clements and Krolzig (1998) takes $m = 3$ and $p = 4$.

Note that, for both models, we can use the optimal weights of the intercept only model because, conditional on the estimated parameters, the state-independent autoregressive component can be moved to the left hand side. On the right hand side, only the constant remains and we can use the optimal weights of the intercept only model. We estimate the models using the EM algorithm suggested by Hamilton (1994) with the extensions discussed by Krolzig (1997). We have investigated the performance of optimal weights for such dynamic models in Monte Carlo experiments with details provided in Appendix 2.B. The results indicate that the insights gained from the intercept only model in Section 2.5 carry over to dynamic models.

In this exercise, we focus on pseudo-out-of-sample forecasts generated by a range of candidate Markov switching models: MSM(m)-AR(p) and MSMH(m)-AR(p) models with $m = 2$ and $p = 0, 1, 2, 3, 4$ and $m = 3$ with $p = 1, 2$, and MSI (m)-AR(p) and MSIH(m)-AR(p) models with $m = 2, 3$ and $p = 0, 1, 2, 3, 4$. We construct expanding window forecasts where for each forecast all models are re-estimated to include all available data at that point in time. We select the Markov switching model that, based on standard weights, delivers the lowest MSFE in a cross-validation sample. Using this model, we then compare the pseudo out-of-sample forecasts using standard weights and optimal weights.

We report the ratio of the MSFE of forecasts from optimal weights relative to those from standard weights together with the Diebold and Mariano (1995) test statistic of equal predictive accuracy. Additionally, we calculate the components of MSFE: the squared biases and variances. We report the differences between the squared bias of the standard weights forecasts and that of the optimal weight forecasts relative to the MSFE of the standard weight forecast, and the differences between the variance of the standard weights forecasts and that of the optimal weight forecasts relative to the MSFE of the standard weight forecast.

The data are (log changes in) US GNP series from 1947Q1 to 2014Q1, which we obtained from the Federal Reserve Economic Data (FRED). The data are seasonally adjusted. In total, the series consists of 269 observations. After accounting for the necessary pre-sample, we start the estimation sample in 1948Q2.

The out-of-sample forecast period is 1983Q2-2014Q1, which amounts to 124 observations and ensures that throughout the forecasting exercise all models are estimated on at least 100 observations. We start evaluating forecasts for model selection purposes with a training period 1973Q2-1983Q1 (40 observations). The model that has the minimum MSFE over this period (using standard weights) is selected as the forecasting model for the observation 1983Q2, and forecasts using the different weights are made with this model. In this way, no information is used that is not available to researchers in real time. Next, we add the next period to our estimation and cross-validation sample, select the minimum MSFE model, and

Table 2.5: GNP forecasts: forecasting performance

	w_{MS}	$w_{\hat{s}}$	$w_{\hat{\xi}}$	$w_{\hat{\mathbf{M}}}$
1983Q2-2014Q1	0.367	1.001	0.970^{**}	0.959^{***}
Subperiods				
1983Q2-1993Q1	0.225	1.002	0.875^{**}	0.898^{*}
1993Q2-2003Q1	0.306	1.000	1.021	0.989
2003Q2-2014Q1	0.553	1.000	0.980^{*}	0.965^{**}
Full sample: 1983Q2-2014Q1				
Square bias	0.008	0.000	0.003	0.005
Variance	0.359	−0.001	0.028	0.037

Note: The second column in the top two panels of the table reports the MSFE based on the best Markov switching model with standard weights. The remaining columns of the table reports the relative MSFE of the optimal weights compared with the Markov switching weights. Asterisks denote significance at the 10% (*), 5% (**), and 1% (***) level using the Diebold-Mariano test statistic. The second column of the last panel reports the square bias and variance of the best Markov switching model with standard weights. The remaining columns give the differences in squared biases and variances between the standard weights and optimal weights forecasts relative to the MSFE of the Markov switching model with standard weights. Positive numbers indicate lower bias/variance.

construct the next forecast. Remarkably, in our application, the MSM(3)-AR(1) model is selected throughout.

As mentioned above, the beginning of the out-of-sample forecast period is chosen such that a sufficient amount of observations is available to estimate all Markov switching models. Still, we need to ensure that our results do not critically depend on this choice. In a second step, we therefore check the robustness of our results using the forecast evaluation measures proposed by Rossi and Inoue (2012).

The forecasting performances of the standard and optimal weights are reported in Table 2.5. The column with heading w_{MS} reports the MSFE of the best Markov switching model using standard weights. The next three columns report the ratio of MSFE of the optimal weights forecast to the standard weights forecast for the same model. The results in the first line, which are over the full forecast period, show that optimal weights conditional on states, $w_{\hat{s}}$, do not improve forecasts but that, in contrast, weights conditional on state probabilities, $w_{\hat{\xi}}$ and $w_{\hat{\mathbf{M}}}$, substantially improve the forecast performance over standard weights and that these improvements are significant. The most precise forecasts result from using $w_{\hat{\mathbf{M}}}$. The three state models have an average estimated differences in mean (scaled by the standard deviation) $\hat{\lambda}_{21} = 2.28$ and $\hat{\lambda}_{31} = 4.23$. The average minimum normalized variance

of the smoothed probability vector is $\tilde{\sigma}^2_{\hat{\xi}_{|T}} = 0.20$. The size of the improvements over the Markov switching forecast is close to the improvements found in the Monte Carlo simulation for three state models as presented in Table 2.4.

It is interesting to also compare forecast performance in subsamples. In the first subsample, 1983Q2–1993Q1, forecasts based on the optimal weights conditional on state probabilities, $w_{\hat{\xi}}$ and $w_{\hat{\mathbf{M}}}$, improve significantly over the standard weights with gains of more than 10% in forecast accuracy. Forecasts based on the plug-in weights, $w_{\hat{s}}$, in contrast, cannot improve on the standard MS forecasts. In the second subsample, 1993Q2–2003Q1, which largely covers the great moderation, only $w_{\hat{\mathbf{M}}}$ offers a modest improvement. In the last subsample, 2003Q2–2014Q1, again all optimal weights conditional on the state probabilities lead to more precise forecasts than the standard weights and these improvements are again significant.

The optimal weights trade off bias and variance of the forecasts, and it is therefore interesting to consider the magnitude of the bias incurred. The bottom panel of Table 2.5 reports the squared bias and variance of the forecasts from the standard weights forecasts in the second column and, in the subsequent columns, the difference in squared biases and variances of the standard weights and the optimal weights forecasts relative to the MSFE of the standard weights forecasts. It can be seen that the squared bias of the standard weights forecast is very small and only a fraction of the size of the variance. The reduction in MSFE that the optimal weights (based on state probabilities) achieve is therefore for the most part via a reduction in variance. Yet, in this application there appears to be no trade-off in bias as the biases of the optimal weights forecasts are no larger and typically smaller than that of the standard weights forecasts. It appears that the model uncertainty around the Markov switching model induces a bias that the optimal weights mitigate, which leads to improvements of the forecasts in bias and variance.

Having established that the optimal weights improve on the Markov switching model with standard weights, the question remains how the optimal weights forecasts compare to forecasts from linear models, which here are AR(p) models with $p = 1, 2, 3, 4$ and a mean only model. We select the best linear model based on the historic forecast performance in line with the model selection for the Markov switching model. The AR(1) model is selected for the first 69 forecasts and the AR(2) model for the remaining forecasts. The resulting MSFE and relative performance of the different weighting scheme for the selected Markov switching model are reported in Table 2.6. Over the entire forecast period, the performance of the linear models is very similar to the Markov switching model with standard weights. The same is true for the weights conditional on states. This contrasts with the forecast based on optimal weights conditional on state probabilities that substantially beat the linear models, even if for the full forecast sample the difference is not significant at conventional levels.

Table 2.6: GNP forecasts: comparison to linear models

	AR_{dyn}	w_{MS}	$w_{\hat{s}}$	$w_{\hat{\xi}}$	$w_{\hat{\mathbf{M}}}$
1983Q2-2014Q1	0.368	0.999	1.000	0.970	0.958
Subperiods					
1983Q2-1993Q1	0.265	0.849^{**}	0.851^{**}	0.743^{**}	0.763^{**}
1993Q2-2003Q1	0.280	1.091	1.091	1.114	1.080
2003Q2-2014Q1	0.540	1.023	1.023	1.003	0.988

Note: The second column contains the MSFE of the best linear model. The remaining columns contain the MSFE of the best Markov switching model with different weights relative to that of the linear model. The best Markov switching model is selected based on standard weights. The linear model is the AR(1) model for the first 69 forecasts and AR(2) for the final 55 forecasts. Asterisks denote significance at the 10% (*), 5% (**), and 1% (***) level using the Diebold-Mariano test statistic.

The results for the three different subsamples reveal that, in the first subsample, all Markov switching forecasts significantly improve on the linear forecasts. The largest gains are made using the optimal weights conditional on the state probabilities. In the middle subsample, no Markov switching forecast is more precise than the linear model. In the final subsample, optimal weights, $w_{\hat{\mathbf{M}}}$ again yield forecasts with a lower MSFE than the linear model. Comparing these results to those in Table 2.5, suggests that the optimal weights improve forecasts over the standard weights the most when the data exhibit strong switching behavior. This ties in with the results from our theory in two ways. First, we showed above that the weights conditional on the states are tending towards equal weighting, that is in the direction of the linear models, whereas the optimal weights derived conditional on state probabilities emphasize the Markov switching nature of the data. Second, we demonstrated that, in a three state model, the optimal weights are around $1/T$ when the future regime is the middle regime. This appears to be a distinguishing feature of the subsamples: in the first subsample, the forecast observation is estimated to be, on average, in the middle regime with probability 0.65. In the second and third subsamples, in contrast, the average probabilities are 0.83 and 0.84. Hence, the linear model is more difficult to beat in the second and third subsample as, for many forecast observations, the forecast from the linear model is close to the optimal forecast from the Markov switching model.

In order to check the robustness of our results to the choice of forecast sample, we additionally use the forecast accuracy tests suggested by Rossi and Inoue (2012). The tests require the calculation of Diebold-Mariano test statistics over a range of possible out-of-sample forecast windows. From these different windows, two tests can be constructed: first, the $\mathcal{A}_T$ test, which is the average of the Diebold-Mariano test statistics, and, second, the $\mathcal{R}_T$ test, which is the supremum of the Diebold-Mariano test statistics. The application of

Table 2.7: Rossi and Inoue test of forecast accuracy

	w_{MS}	$w_{\hat{s}}$	$w_{\hat{\xi}}$	$w_{\hat{\mathbf{M}}}$
Test against MS weights				
$\mathcal{A}_T$		0.585	-0.356	-0.910
$\mathcal{R}_T$		-0.646	-1.803	-2.342**
Test against AR(1)				
$\mathcal{A}_T$	-0.223	-0.222	-0.208	-0.546
$\mathcal{R}_T$	-0.954	-0.951	-1.071	-1.575
Test against AR(2)				
$\mathcal{A}_T$	0.372	0.375	0.261	-0.027
$\mathcal{R}_T$	-0.469	-0.477	-0.621	-0.928

Note: The beginning of the out-of-sample forecast evaluation period is varied between $[\mu T, (1-\mu)T]$ with $\mu = 0.35$ and $T = 264$. $\mathcal{A}_T$ denotes the average and $\mathcal{R}_T$ the supremum of the Diebold-Mariano test statistics over the range of forecast periods. Asterisks denote significance at the 10% (*), 5% (**), and 1% (***) level.

these tests comes with two caveats in our application. First, the relative short first estimation window implied by these tests is problematic as various switches of the Markov chain are required for the estimation of Markov switching models. For the test by Rossi and Inoue (2012), the beginning of the out-of-sample forecast evaluation period is varied over the interval $[\mu T, (1-\mu)T]$ and we set μ to the maximum of 0.35. In contrast, in the baseline application above, the shortest estimation sample is $0.53\,T$. Early forecasts for the Rossi and Inoue test may suffer as a result of a short estimation window. Second, as a further consequence of the shortened estimation sample, we cannot use cross-validation as model selection procedure and therefore consider only the MSM(3)-AR(1) model, which has been selected in our baseline forecast procedure throughout, and for the linear model we use the AR(1) and AR(2) models, which are the models selected in the baseline forecasting exercise.

Table 2.7 reports the test statistics and associated significance levels. The top panel reports the test statistics of the optimal weights forecasts against the standard weights forecasts. It can be seen that the signs of the test statistics are as expected and that the $w_{\hat{\mathbf{M}}}$ weights provide significant improvements on the standard weights according to the $\mathcal{R}_T$ test. The lower two panels of Table 2.7 report the test statistics when the MSM(3)-AR(1) model is tested against a simple AR(1) and AR(2) model. For the AR(1) model the signs are as expected, although the test statistics do not exceed the critical values reported in Rossi and Inoue (2012). For the AR(2) model the $\mathcal{A}_T$ test statistic for $w_{\hat{\mathbf{M}}}$ weights remains negative. For these weights

the largest negative $\mathcal{R}_T$ test statistic is observed, which it is not significant at conventional levels. This reflects the fact that the linear model is a close approximation to the optimal weights Markov switching model as the forecast sample is dominated by observations that are most likely from the middle regime.

2.7 Conclusion

In this chapter, we have derived optimal forecasts for Markov switching models and analyzed the effect of uncertainty around states on forecasts based on optimal weights. The importance of uncertainty of the states of the Markov chain is highlighted in the comparison of forecasts from weights conditional on the states and those when the states are not known. The optimal weights for known states share the properties of the weights derived in Pesaran et al. (2013) and are asymptotically identical to the Markov switching weights. Improvements in forecasting performance are found when the ratio of the number of observations to the number of estimated parameters is small. This contrasts with the optimal weights for unknown states that are asymptotically different from the Markov switching weights and potential improvements in forecasting accuracy can be considerable for large differences in parameters even in large samples.

The results from theory and the application show that optimal forecasts can differ substantially from standard MS forecasts. Optimal weights emphasize the Markov switching nature of the DGP more than standard weights do. However, in the three state case, the optimal weights for forecasts in the middle regime lead to weights that effectively ignore the Markov switching nature of the data. This is the case for the GNP forecasts from the great moderation where the vast majority of observations are from the middle regime. This explains the difficulty of Markov switching forecasts to beat linear models, as the optimal forecast from the Markov switching model is essentially the same as the forecast from the linear model.

For practitioners two messages emerge. First, when the observation in the forecast period could likely be from any regime of the Markov switching model, optimal weights conditional on state probabilities will substantially improve forecasts. When the size of the switches is moderate or regime estimates precise, weights that ignore the covariances of the states are more efficient as the additional estimation uncertainty introduced by estimating the covariances of the states dominates the forecasts. When switches are large yet state remain uncertain using the full second moment matrix of the Markov chain leads to more precise forecasts. However, the difference between the two optimal weights is small compared to the overall gains in forecast accuracy. Second, when one expects to forecast predominantly observations from the middle regime in a three state model, using a linear model will lead

to forecasts that are effectively the optimal forecasts from the Markov switching model but with the benefit of substantially reduced estimation uncertainty.

2.A Mathematical details

2.A.1 Derivations conditional on states

Weights for two-state Markov switching model

In order to derive weights (2.10)–(2.13), define $\lambda = \frac{\beta_2-\beta_1}{\sigma_2}$ and $q = \frac{\sigma_1}{\sigma_2}$, $\pi_1 = \frac{1}{T}\sum_{t=1}^{T} s_{1t}$, and $\pi_2 = \frac{1}{T}\sum_{t=1}^{T} s_{2t}$. Then we have

$$\begin{aligned}\mathbf{M} &= \mathbf{Q} + \tilde{\mathbf{S}}'\boldsymbol{\lambda}\boldsymbol{\lambda}'\tilde{\mathbf{S}} \\ &= q^2\mathbf{S}_1 + \mathbf{S}_2 + \lambda^2\mathbf{s}_2\mathbf{s}_2'\end{aligned}$$

where $\mathbf{S}_i$ is a $T \times T$ diagonal matrix with typical t,t-element $s_{i,t}$. The inverse of $\mathbf{M}$ is

$$\begin{aligned}\mathbf{M}^{-1} &= (q^2\mathbf{S}_1 + \mathbf{S}_2)^{-1} - \frac{\lambda^2(q^2\mathbf{S}_1 + \mathbf{S}_2)^{-1}\mathbf{s}_2\mathbf{s}_2'(q^2\mathbf{S}_1 + \mathbf{S}_2)^{-1}}{1 + \lambda^2\mathbf{s}_2'(q^2\mathbf{S}_1 + \mathbf{S}_2)^{-1}\mathbf{s}_2} \\ &= \frac{1}{q^2}\mathbf{S}_1 + \mathbf{S}_2 - \frac{\lambda^2(\frac{1}{q^2}\mathbf{S}_1 + \mathbf{S}_2)\mathbf{s}_2\mathbf{s}_2'(\frac{1}{q^2}\mathbf{S}_1 + \mathbf{S}_2)}{1 + \lambda^2\mathbf{s}_2'(\frac{1}{q^2}\mathbf{S}_1 + \mathbf{S}_2)\mathbf{s}_2} \\ &= \frac{1}{q^2}\mathbf{S}_1 + \mathbf{S}_2 - \frac{\lambda^2\mathbf{s}_2\mathbf{s}_2'}{1 + \lambda^2 T\pi_2}\end{aligned}$$

The weights are given by

$$\mathbf{w} = \lambda^2\mathbf{M}^{-1}\mathbf{s}_2 s_{2,T+1} + \frac{\mathbf{M}^{-1}\boldsymbol{\iota}}{\boldsymbol{\iota}'\mathbf{M}^{-1}\boldsymbol{\iota}}\left(1 - \lambda^2\boldsymbol{\iota}'\mathbf{M}^{-1}\mathbf{s}_2 s_{2,T+1}\right)$$

The various components needed to calculate the weights are given by

$$\begin{aligned}\mathbf{M}^{-1}\mathbf{s}_2 &= \mathbf{s}_2 - \frac{\lambda^2 T\pi_2}{1 + \lambda^2 T\pi_2}\mathbf{s}_2 \\ &= \frac{1}{1 + \lambda^2 T\pi_2}\mathbf{s}_2 \\ \mathbf{M}^{-1}\boldsymbol{\iota} &= \frac{1}{q^2}\mathbf{s}_1 + \mathbf{s}_2 - \frac{\lambda^2 T\pi_2}{1 + \lambda^2 T\pi_2}\mathbf{s}_2 \\ &= \frac{\mathbf{s}_1(1 + \lambda^2 T\pi_2) + q^2\mathbf{s}_2}{q^2(1 + \lambda^2 T\pi_2)}\end{aligned}$$

and

$$\boldsymbol{\iota}'\mathbf{M}^{-1}\mathbf{s}_2 = \frac{T\pi_2}{1 + \lambda^2 T\pi_2}, \quad \boldsymbol{\iota}'\mathbf{M}^{-1}\boldsymbol{\iota} = T\frac{\pi_1 + \lambda^2 T\pi_1\pi_2 + q^2\pi_2}{q^2(1 + \lambda^2 T\pi_2)}$$

This yields the weights

$$\begin{aligned}\mathbf{w} &= \lambda^2 \frac{1}{1+\lambda^2 T\pi_2}\mathbf{s}_2 s_{2,T+1} + \frac{1}{T}\frac{\mathbf{s}_1(1+\lambda^2 T\pi_2)+q^2\mathbf{s}_2}{\pi_2 q^2+\pi_1(1+T\pi_2\lambda^2)}\left(1-\lambda^2\frac{T\pi_2 s_{2,T+1}}{1+\lambda^2 T\pi_2}\right)\\ &= \frac{1}{1+\lambda^2 T\pi_2}\left\{\mathbf{s}_2 s_{2,T+1} + \frac{1}{T}\frac{\mathbf{s}_1(1+\lambda^2 T\pi_2)+q^2\mathbf{s}_2}{\pi_2 q^2+\pi_1(1+T\pi_2\lambda^2)}\left[1+\lambda^2 T\pi_2(1-s_{2,T+1})\right]\right\}\end{aligned}$$

Suppose $s_{2,T+1} = s_{2,t} = 1$, then

$$\begin{aligned}w_{(2,2)} &= \frac{1}{1+\lambda^2 T\pi_2}\left(\lambda^2 + \frac{1}{T}\frac{q^2}{\pi_2 q^2+\pi_1(1+T\pi_2\lambda^2)}\right)\\ &= \frac{1}{1+\lambda^2 T\pi_2}\frac{1}{T}\frac{1}{\pi_2 q^2+\pi_1(1+T\pi_2\lambda^2)}\left[q^2(1+\lambda^2 T\pi_2)+\lambda^2 T\pi_1(1+\lambda^2 T\pi_2)\right]\\ &= \frac{1}{T}\frac{q^2+\lambda^2 T\pi_1}{\pi_2 q^2+\pi_1(1+T\pi_2\lambda^2)}\end{aligned}$$

when $s_{2,T+1} = 1$, $s_{2,t} = 0$, then

$$\begin{aligned}w_{(2,1)} &= \frac{1}{1+\lambda^2 T\pi_2}\left(\frac{1}{T}\frac{1+\lambda^2 T\pi_2}{\pi_2 q^2+\pi_1(1+T\pi_2\lambda^2)}\right)\\ &= \frac{1}{T}\frac{1}{\pi_2 q^2+\pi_1(1+T\pi_2\lambda^2)}\end{aligned}$$

when $s_{2,T+1} = 0, s_{2,t} = 1$, then

$$\begin{aligned}w_{(1,2)} &= \frac{1}{T}\frac{1}{1+\lambda^2 T\pi_2}\frac{q^2(1+\lambda^2 T\pi_2)}{\pi_2 q^2+\pi_1(1+T\pi_2\lambda^2)}\\ &= \frac{1}{T}\frac{q^2}{\pi_2 q^2+\pi_1(1+T\pi_2\lambda^2)}\end{aligned}$$

finally, when $s_{2,T+1} = 0, s_{2,t} = 0$, then

$$\begin{aligned}w_{(1,1)} &= \frac{1}{T}\frac{1}{1+\lambda^2 T\pi_2}\frac{(1+\lambda^2 T\pi_2)^2}{\pi_2 q^2+\pi_1(1+T\pi_2\lambda^2)}\\ &= \frac{1}{T}\frac{1+\lambda^2 T\pi_2}{\pi_2 q^2+\pi_1(1+T\pi_2\lambda^2)}\end{aligned}$$

In order to show the symmetry of the weights, consider the definition of λ and q conditional on the regime $s_{i,T+1}$. If $s_{2,T+1} = 1$, define $\lambda = \frac{\beta_2-\beta_1}{\sigma_2}$ and $q = \frac{\sigma_1}{\sigma_2}$, but if $s_{1,T+1} = 1$,

define $\lambda_* = \frac{\beta_1 - \beta_2}{\sigma_1}$ and $q_* = \frac{\sigma_2}{\sigma_1}$. Then, $\lambda^2 = \lambda_*^2/q_*^2$ and we have for $w_{(1,2)}$ and $w_{(1,1)}$

$$\begin{aligned} w_{(1,2)} &= \frac{1}{T}\frac{q^2}{\pi_2 q^2 + \pi_1(1 + T\pi_2\lambda^2)} \\ &= \frac{1}{T}\frac{1/q_*^2}{\pi_2/q_*^2 + \pi_1(1 + 1/q_*^2 T\pi_2\lambda_*^2)} \\ &= \frac{1}{T}\frac{1}{\pi_1 q_*^2 + \pi_2(1 + T\pi_1\lambda_*^2)} \end{aligned}$$

$$\begin{aligned} w_{(1,1)} &= \frac{1}{T}\frac{1 + \lambda^2 T\pi_2}{\pi_2 q^2 + \pi_1(1 + T\pi_2\lambda^2)} \\ &= \frac{1}{T}\frac{1 + 1/q_*^2\lambda_*^2 T\pi_2}{\pi_2/q_*^2 + \pi_1(1 + 1/q_*^2 T\pi_2\lambda_*^2)} \\ &= \frac{1}{T}\frac{q_*^2 + \lambda_*^2 T\pi_2}{\pi_1 q_*^2 + \pi_2(1 + T\pi_1\lambda_*^2)} \end{aligned}$$

The symmetry of the weights is a natural consequence of the fact that the Markov Switching model is invariant under a relabeling of the states.

Weights and MSFE for m-state Markov switching model

To derive weights for an m-state Markov switching model, we will concentrate on $s_{k,T+1} = 1$ as we have shown above that the weights are symmetric. In this case, define $\lambda_i = (\beta_i - \beta_k)/\sigma_k$ and $q_i = \sigma_i/\sigma_k$. The model is given by

$$\begin{aligned} y_t &= \sum_{i=1}^{m} \beta_i s_{it} + \sum_{i=1}^{m} \sigma_i s_{it}\varepsilon_t \\ &= \beta_k + \sum_{i=1}^{m}(\beta_i - \beta_k)s_{it} + \sum_{i=1}^{m} \sigma_i s_{it}\varepsilon_t \\ &= \sigma_k \left(\frac{\beta_k}{\sigma_k} + \sum_{i=1}^{m} \lambda_i s_{it} + \sum_{i=1}^{m} q_i s_{it}\varepsilon_t \right) \end{aligned}$$

For the observation at $T + 1$ we have

$$\frac{1}{\sigma_k} y_{T+1} = \frac{\beta_k}{\sigma_k} + \varepsilon_{T+1}$$

The forecast error is

$$\frac{1}{\sigma_k}(y_{T+1} - \mathbf{w}'\mathbf{y}) = \varepsilon_{T+1} - \sum_{i=1}^{m} \lambda_i \mathbf{w}'\mathbf{s}_i - \sum_{i=1}^{m} q_i \mathbf{w}'\mathbf{S}_i\boldsymbol{\varepsilon}$$

Squaring and taking expectations gives

$$\mathrm{E}\left[\sigma_k^{-2}(y_{T+1} - \mathbf{w}'\mathbf{y})^2\right] = 1 + \sum_{i=1}^{m}\sum_{j=1}^{m}\lambda_i\lambda_j\mathbf{w}'\mathbf{s}_i\mathbf{s}_j'\mathbf{w} + \sum_{i=1}^{m} q_i^2\mathbf{w}\mathbf{S}_i\mathbf{w}$$

Implementing the constraint $\sum_{t=1}^{T} w_t = 1$ by a Lagrange multiplier and taking the derivative gives

$$\begin{aligned}\mathbf{w} &= \left(\sum_{i=1}^{m}\sum_{j=1}^{m}\lambda_i\lambda_j\mathbf{w}'\mathbf{s}_i\mathbf{s}_j' + \sum_{i=1}^{m} q_i^2\mathbf{w}\mathbf{S}_i\right)^{-1}(-\theta\boldsymbol{\iota}) \\ &= -\theta\mathbf{M}^{-1}\boldsymbol{\iota}\end{aligned} \tag{2.36}$$

The inverse can be expressed analytically through the Sherman Morrison formula as

$$\begin{aligned}\mathbf{M}^{-1} &= \sum_{i=1}^{m}\frac{1}{q_i^2}\mathbf{S}_i - \frac{\left(\sum_{i=1}^{m}\frac{1}{q_i^2}\mathbf{s}_i\right)\left(\sum_{i=1}^{m}\sum_{j=1}^{m}\lambda_i\lambda_j\mathbf{s}_i\mathbf{s}_j'\right)\left(\sum_{i=1}^{m}\frac{1}{q_i^2}\mathbf{s}_i\right)}{1 + \left(\sum_{j=1}^{m}\lambda_j\mathbf{s}_j'\right)\left(\sum_{i=1}^{m}\frac{1}{q_i^2}\mathbf{s}_i\right)\left(\sum_{i=1}^{m}\lambda_j\mathbf{s}_i\right)} \\ &= \sum_{i=1}^{m}\frac{1}{q_i^2}\mathbf{S}_i - \frac{\sum_{i=1}^{m}\sum_{j=1}^{m}\frac{\lambda_i}{q_i^2}\frac{\lambda_j}{q_j^2}\mathbf{s}_i\mathbf{s}_j'}{1 + T\sum_{i=1}^{m}\frac{\lambda_i^2}{q_i^2}\pi_i}\end{aligned}$$

Multiplying with $\boldsymbol{\iota}$ as in equation (2.36) gives

$$\mathbf{M}^{-1}\boldsymbol{\iota} = \sum_{i=1}^{m}\frac{1}{q_i^2}\mathbf{s}_i - \frac{T\sum_{i=1}^{m}\sum_{j=1}^{m}\frac{\lambda_i}{q_i^2}\frac{\lambda_j}{q_j^2}\pi_j\mathbf{s}_i}{1 + T\sum_{i=1}^{m}\frac{\lambda_i^2}{q_i^2}\pi_i}$$

Since the weights should sum up to one, we have

$$\begin{aligned}\boldsymbol{\iota}'\mathbf{w} &= \left(T\sum_{i=1}^{m}\frac{1}{q_i^2}\pi_i - \frac{T^2\sum_{i=1}^{m}\sum_{j=1}^{m}\frac{\lambda_i}{q_i^2}\frac{\lambda_j}{q_j^2}\pi_j\pi_i}{1 + T\sum_{i=1}^{m}\frac{\lambda_i^2}{q_i^2}\pi_i}\right)(-\theta) \\ &= 1\end{aligned}$$

which gives

$$\begin{aligned}\theta &= \frac{1 + T\sum_{j=1}^{m}\frac{\lambda_i^2}{q_i^2}\pi_i}{T}\left[\sum_{i=1}^{m}\frac{1}{q_i^2}\pi_i + T\sum_{i=1}^{m}\sum_{j=1}^{m}\left(\frac{1}{q_i^2}\frac{\lambda_j}{q_j^2}\pi_i\pi_j - \frac{\lambda_i}{q_i^2}\frac{\lambda_j}{q_j^2}\pi_j\pi_i\right)\right]^{-1} \\ &= \frac{1 + T\sum_{j=1}^{m}\frac{\lambda_i^2}{q_i^2}\pi_i}{T}\left[\sum_{i=1}^{m}\frac{1}{q_i^2}\pi_i + T\sum_{i=1}^{m}\sum_{j=1}^{m}\frac{1}{q_i^2}\frac{1}{q_j^2}\pi_i\pi_j\lambda_j(\lambda_j - \lambda_i)\right]^{-1}\end{aligned}$$

The weights are then given by

$$\mathbf{w} = \frac{1}{T}\frac{\sum_{i=1}^{m}\frac{1}{q_i^2}\mathbf{s}_i + T\sum_{i=1}^{m}\sum_{j=1}^{m}\frac{1}{q_i^2}\frac{1}{q_j^2}\pi_j\lambda_j(\lambda_j-\lambda_i)\mathbf{s}_i}{\sum_{i=1}^{m}\frac{1}{q_i^2}\pi_i + T\sum_{i=1}^{m}\sum_{j=1}^{m}\frac{1}{q_i^2}\frac{1}{q_j^2}\pi_i\pi_j\lambda_j(\lambda_j-\lambda_i)}$$

So that if $s_{lt}=1$ the weight at time t is

$$w_t = \frac{1}{T}\frac{\frac{1}{q_l^2} + T\sum_{j=1}^{m}\frac{1}{q_l^2}\frac{1}{q_j^2}\pi_j\lambda_j(\lambda_j-\lambda_l)}{\sum_{i=1}^{m}\frac{1}{q_i^2}\pi_i + T\sum_{i=1}^{m}\sum_{j=1}^{m}\frac{1}{q_i^2}\frac{1}{q_j^2}\pi_i\pi_j\lambda_j(\lambda_j-\lambda_i)}$$

The MSFE is easy to derive by noting that we can substitute the first order condition for the weights

$$\begin{aligned}\mathrm{E}\left[\sigma_k^{-2}(y_{T+1}-\mathbf{w}'\mathbf{y})^2\right] &= 1 + \sum_{i=1}^{m}\sum_{j=1}^{m}\lambda_i\lambda_j\mathbf{w}'\mathbf{s}_i\mathbf{s}_j'\mathbf{w} + \sum_{i=1}^{m}q_i^2\mathbf{w}\mathbf{S}_i\mathbf{w}\\ &= 1-\theta\\ &= 1+w_{(k,k)}\end{aligned}$$

where $w_{(k,k)}$ is the weight when $s_{k,T+1}=s_{kt}=1$.

2.A.2 Derivations conditional on state probabilities

Large T approximation for optimal weights

Rewrite (2.21) as

$$w_t = \frac{1}{T}\frac{d_t[\frac{1}{T}+\lambda^2\frac{1}{T}\sum_{t'=1}^{T}d_{t'}(\xi_{2,T+1}-\xi_{2t'})(\xi_{2t}-\xi_{2t'})]}{\frac{1}{T}\left(\frac{1}{T}\sum_{t'=1}^{T}d_{t'}\right)+\lambda^2\left[\frac{1}{T}\sum_{t'=1}^{T}d_{t'}\xi_{2t'}^2\frac{1}{T}\sum_{t'=1}^{T}d_{t'}-\left(\frac{1}{T}\sum_{t'=1}^{T}d_{t'}\xi_{2t'}\right)^2\right]} \tag{2.37}$$

where

$$d_t = \left[\lambda^2\xi_{2t}(1-\xi_{2t})+q^2+(1-q^2)\xi_{2t}\right]^{-1}$$

To perform the large sample approximation we need to establish that $\frac{1}{T}\sum_{t=1}^{T}d_t<\infty$, $\frac{1}{T}\sum_{t=1}^{T}\xi_{2t}d_t<\infty$ and $\frac{1}{T}\sum_{t=1}^{T}\xi_{2t}^2d_t<\infty$. Proving the first of these relations implies the other two, since $0\leqslant\xi_{2t}\leqslant 1$. Define $a_t=\frac{1}{d_t}$. We then need to prove that $a_t>0$. The only scenario where $a_t=0$ is when $\xi_{2t}=0$ and $q^2=0$, so the only restriction that we must impose to obtain $a_t>0$ is that $q^2>0$. Then

$$\frac{1}{T}\sum_{t=1}^{T}d_t = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{a_t}\leqslant\frac{1}{T}T\frac{1}{a_{\min}} = \frac{1}{a_{\min}}<\infty$$

where $a_{\min}$ is the minimum value of a_t over $t = 1, 2, \ldots, T$.

Denote $\bar{d} = \frac{1}{T}\sum_{t=1}^{T} d_t$, $\bar{d\xi} = \frac{1}{T}\sum_{t=1}^{T} d_t\xi_{2t}$, and $\overline{d\xi^2} = \frac{1}{T}\sum_{t=1}^{T} d_t\xi_{2t}^2$, then (2.37) can be written as

$$\begin{aligned} w_t &= \frac{1}{T}d_t\left[\frac{\frac{1}{T}}{\frac{1}{T}\bar{d}+\lambda^2\left(\overline{d\xi^2}\bar{d}-\bar{d\xi}^2\right)}+\frac{\lambda^2(\xi_{2t}\xi_{2,T+1}\bar{d}-\xi_{2t}\bar{d\xi}-\xi_{2,T+1}\bar{d\xi}+\overline{d\xi^2})}{\frac{1}{T}\bar{d}+\lambda^2\left(\overline{d\xi^2}\bar{d}-\bar{d\xi}^2\right)}\right] \\ &= \frac{1}{T}d_t\left[\frac{1}{T}\frac{1}{\lambda^2\left(\overline{d\xi^2}\bar{d}-\bar{d\xi}^2\right)}\frac{1}{1+\frac{\theta}{T}}+\frac{\lambda^2(\xi_{2t}\xi_{2,T+1}\bar{d}-\xi_{2t}\bar{d\xi}-\xi_{2,T+1}\bar{d\xi}+\overline{d\xi^2})}{\lambda^2\left(\overline{d\xi^2}\bar{d}-\bar{d\xi}^2\right)}\frac{1}{1+\frac{\theta}{T}}\right] \\ &= \frac{1}{T}d_t\frac{\lambda^2(\xi_{2t}\xi_{2,T+1}\bar{d}-\xi_{2t}\bar{d\xi}-\xi_{2,T+1}\bar{d\xi}+\overline{d\xi^2})}{\lambda^2\left(\overline{d\xi^2}\bar{d}-\bar{d\xi}^2\right)}+\mathcal{O}\left(T^{-2}\right) \end{aligned}$$

where $\theta = \frac{\bar{d}}{\lambda^2\left(\overline{d\xi^2}\bar{d}-\bar{d\xi}^2\right)} = \frac{1}{\lambda^2\sum_{t=1}^{T}\tilde{d}_t(\xi_{2t}-\frac{1}{T}\sum_{t'=1}^{T}\tilde{d}_{t'}\xi_{2t'})^2}$ where $\tilde{d}_t = d_t/\sum_{t'} d_{t'}$. The numerator is nonzero unless for the trivial case when ξ_{2t} is constant for all t. Using this and the result that $\bar{d}$, $\bar{d\xi}$ and $\overline{d\xi^2}$ are finite for any T proves that we can apply the expansion in terms of θ/T. Dividing w_t by $\sum_{t=1}^{T} d_t$ yields (2.23).

Weights and MSFE for standard Markov switching model

The Markov switching weights can be written as

$$\begin{aligned} \mathbf{w}_{\text{MS}} &= \frac{\xi_{1,T+1}\boldsymbol{\xi}_1}{\sum_{t=1}^{T}\xi_{1t}}+\frac{\xi_{2,T+1}\boldsymbol{\xi}_2}{\sum_{t=1}^{T}\xi_{2t}} \\ &= \frac{1}{T}\frac{\xi_{2,T+1}\boldsymbol{\xi}_2}{\bar{\xi}_2}+\frac{1}{T}\frac{(1-\xi_{2,T+1})(\boldsymbol{\iota}-\boldsymbol{\xi}_2)}{(1-\bar{\xi}_2)} \\ &= \frac{1}{T}\frac{1}{\bar{\xi}_2(1-\bar{\xi}_2)}(\xi_{2,T+1}\boldsymbol{\xi}_2(1-\bar{\xi}_2+\bar{\xi}_2)+\bar{\xi}_2\boldsymbol{\iota}-\bar{\xi}_2\xi_{2,T+1}\boldsymbol{\iota}-\bar{\xi}_2\boldsymbol{\xi}_2) \\ &= \frac{1}{T}\frac{1}{\bar{\xi}_2(1-\bar{\xi}_2)}(\xi_{2,T+1}-\bar{\xi}_2)(\boldsymbol{\xi}_2-\bar{\xi}_2\boldsymbol{\iota})+\bar{\xi}_2(1-\bar{\xi}_2) \\ &= \frac{1}{T}+\frac{1}{T}\frac{(\xi_{2,T+1}-\bar{\xi}_2)(\boldsymbol{\xi}_2-\bar{\xi}_2\boldsymbol{\iota})}{\bar{\xi}_2(1-\bar{\xi}_2)} \end{aligned} \tag{2.38}$$

For a general vector of weights $\mathbf{w}$, subject to $\sum_{t=1}^{T} w_t = 1$, and assuming a constant error variance, we have the following MSFE

$$\begin{aligned} \text{E}[\sigma^{-2}e_{T+1}^2] &= 1+\lambda^2\xi_{2,T+1}+\mathbf{w}'\mathbf{M}\mathbf{w}-2\lambda^2\mathbf{w}'\boldsymbol{\xi}\xi_{2,T+1} \\ &= 1+\lambda^2\xi_{2,T+1}+\lambda^2(\mathbf{w}'\boldsymbol{\xi})^2+\mathbf{w}'\mathbf{D}\mathbf{w}-2\lambda^2\mathbf{w}'\boldsymbol{\xi}\xi_{2,T+1} \end{aligned} \tag{2.39}$$

where $\mathbf{D} = (1+\lambda^2\sigma_\xi^2)\mathbf{I}$.

Using (2.38) we have that

$$\begin{aligned}\mathbf{w}'_{\text{MS}}\boldsymbol{\xi} &= \bar{\xi}_2 + \frac{\xi_{2,T+1}-\bar{\xi}_2}{(1-\bar{\xi}_2)\bar{\xi}_2}\left(\frac{1}{T}\sum_{t=1}^{T}\xi_t^2 - T\bar{\xi}_2^2\right)\\ &= \bar{\xi}_2 + \frac{\xi_{2,T+1}-\bar{\xi}_2}{(1-\bar{\xi}_2)\bar{\xi}_2}\left[\bar{\xi}_2(1-\bar{\xi}_2)-\sigma_\xi^2\right]\\ &= \xi_{2,T+1} - \frac{\xi_{2,T+1}-\bar{\xi}_2}{\bar{\xi}_2(1-\bar{\xi}_2)}\sigma_\xi^2\end{aligned}$$

where we have used (2.24), and

$$\mathbf{w}'_{\text{MS}}\mathbf{D}\mathbf{w}_{\text{MS}} = (1+\lambda^2\sigma_\xi^2)\left\{\frac{1}{T} + \frac{(\xi_{2,T+1}-\bar{\xi}_2)^2}{T\bar{\xi}_2^2(1-\bar{\xi}_2)^2}\left[\bar{\xi}_2(1-\bar{\xi}_2)-\sigma_\xi^2\right]\right\}$$

So that the MSFE is

$$\begin{aligned}\mathrm{E}[\sigma^{-2}e_{T+1}^2]_{\text{MS}} &= 1+\lambda^2\xi_{2,T+1} + \lambda^2\left[\xi_{2,T+1}^2 - 2\frac{\xi_{2,T+1}(\xi_{2,T+1}-\bar{\xi}_2)\sigma_\xi^2}{\bar{\xi}_2(1-\bar{\xi}_2)} + \frac{(\xi_{2,T+1}-\bar{\xi}_2)^2\sigma_\xi^4}{\bar{\xi}_2^2(1-\bar{\xi}_2)^2}\right]\\ &\quad - \lambda^2\left[2\xi_{2,T+1}^2 - 2\frac{\xi_{2,T+1}(\xi_{2,T+1}-\bar{\xi}_2)\sigma_\xi^2}{\bar{\xi}_2(1-\bar{\xi}_2)}\right]\\ &\quad + (1+\lambda^2\sigma_\xi^2)\frac{1}{T}\left\{1 + \frac{(\xi_{2,T+1}-\bar{\xi}_2)^2}{\bar{\xi}_2^2(1-\bar{\xi}_2)^2}\left[\bar{\xi}_2(1-\bar{\xi}_2)-\sigma_\xi^2\right]\right\}\\ &= 1+\lambda^2\xi_{2,T+1}(1-\xi_{2,T+1}) + \lambda^2\frac{(\xi_{2,T+1}-\bar{\xi}_2)^2\sigma_\xi^4}{\bar{\xi}_2^2(1-\bar{\xi}_2)^2}\\ &\quad + (1+\lambda^2\sigma_\xi^2)\frac{1}{T}\left\{1 + \frac{(\xi_{2,T+1}-\bar{\xi}_2)^2}{\bar{\xi}_2^2(1-\bar{\xi}_2)^2}\left[\bar{\xi}_2(1-\bar{\xi}_2)-\sigma_\xi^2\right]\right\}\\ &= 1+\lambda^2\xi_{2,T+1}(1-\xi_{2,T+1}) + (1+\lambda^2\sigma_\xi^2)\frac{1}{T}\\ &\quad + \frac{(\xi_{2,T+1}-\bar{\xi}_2)^2}{\bar{\xi}_2^2(1-\bar{\xi}_2)^2}\left\{\lambda^2\sigma_\xi^4 + (1+\lambda^2\sigma_\xi^2)\frac{1}{T}\left[\bar{\xi}_2(1-\bar{\xi}_2)-\sigma_\xi^2\right]\right\}\end{aligned}$$

MSFE for Markov switching model using optimal weights

Equation (2.22) for an arbitrary number of states is derived as follows

$$\begin{aligned}
\mathrm{E}[\sigma^{-2}e_{T+1}^2] &= (\boldsymbol{\iota}'\mathbf{M}^{-1}\boldsymbol{\iota})^{-1}(1-\boldsymbol{\iota}'\mathbf{M}^{-1}\tilde{\boldsymbol{\xi}}\tilde{\xi}_{T+1})^2+ \\
&\quad + \sum_{j=2}^{m}\lambda_j^2\xi_{j,T+1} - \tilde{\xi}_{T+1}^2\tilde{\boldsymbol{\xi}}'\mathbf{M}^{-1}\tilde{\boldsymbol{\xi}} + \sum_{j=1}^{m} q_j^2\xi_{j,T+1} \\
&= \frac{1+\tilde{\boldsymbol{\xi}}'\mathbf{D}^{-1}\tilde{\boldsymbol{\xi}}}{\boldsymbol{\iota}'\mathbf{D}^{-1}\boldsymbol{\iota}(1+\tilde{\boldsymbol{\xi}}'\mathbf{D}^{-1}\tilde{\boldsymbol{\xi}}) - (\boldsymbol{\iota}\mathbf{D}^{-1}\tilde{\boldsymbol{\xi}})^2}\left[1+\frac{\tilde{\xi}_{T+1}^2(\boldsymbol{\iota}'\mathbf{D}^{-1}\tilde{\boldsymbol{\xi}})^2}{(1+\tilde{\boldsymbol{\xi}}'\mathbf{D}^{-1}\tilde{\boldsymbol{\xi}})^2}+\right. \\
&\quad \left.-2\frac{\tilde{\xi}_{T+1}\boldsymbol{\iota}'\mathbf{D}^{-1}\tilde{\boldsymbol{\xi}}}{1+\tilde{\boldsymbol{\xi}}'\mathbf{D}^{-1}\tilde{\boldsymbol{\xi}}}\right] + \tilde{\xi}_{T+1}^2 - \frac{\tilde{\xi}_{T+1}^2\tilde{\boldsymbol{\xi}}'\mathbf{D}^{-1}\tilde{\boldsymbol{\xi}}}{1+\tilde{\boldsymbol{\xi}}'\mathbf{D}^{-1}\tilde{\boldsymbol{\xi}}} + \frac{1}{d_{T+1}} \\
&= \frac{1+\sum_{t=1}^{T}\tilde{\xi}_t^2 - 2\tilde{\xi}_{T+1}\sum_{t=1}^{T}d_t\tilde{\xi}_t + \tilde{\xi}_{T+1}^2\sum_{t=1}^{T}d_t}{\sum_{t=1}^{T}d_t(1+\sum_{t'=1}^{T}d_{t'}\tilde{\xi}_{t'}^2) - (\sum_{t=1}^{T}d_t\tilde{\xi}_t)^2} + \frac{1}{d_{T+1}} \\
&= \frac{w_{T+1}}{d_{T+1}} + \frac{1}{d_{T+1}} \\
&= \frac{1}{d_{T+1}}(1+w_{T+1})
\end{aligned}$$

Derivation of (2.32)

To save on notation, in the following we use $p(s_{jt}|s_{i,t+m},\Omega_T)$ to write $p(s_{jt}=1|s_{i,t+m}=1,\Omega_T)$. To derive (2.32), take for example a three state model and calculate

$$\begin{aligned}
p(s_{jt}|s_{i,t+3},\Omega_T) &= \sum_{k=0}^{2} p(s_{jt}|s_{k,t+1},s_{i,t+3},\Omega_T)p(s_{k,t+1}|s_{i,t+3},\Omega_T) \\
&= \sum_{k=0}^{2} p(s_{jt}|s_{k,t+1},\Omega_t)\sum_{l=0}^{2} p(s_{k,t+1}|s_{l,t+2},\Omega_{t+1})p(s_{l,t+2}|s_{i,t+3},\Omega_{t+2}) \\
&= \sum_{k=0}^{2}\frac{p_{jk}p(s_{jt}|\Omega_t)}{p(s_{k,t+1}|\Omega_t)}\sum_{l=0}^{2}\frac{p_{kl}p(s_{k,t+1}|\Omega_{t+1})}{p(s_{l,t+2}|\Omega_{t+1})}\frac{p_{li}p(s_{l,t+2}|\Omega_{t+2})}{p(s_{i,t+3}|\Omega_{t+2})} \\
&= \frac{p(s_{jt}|\Omega_t)}{p(s_{i,t+3}|\Omega_{t+2})}\sum_{k=0}^{2}\sum_{l=0}^{2} p_{jk}a_{t+1}^k p_{kl}a_{t+2}^l p_{li} \\
&= \frac{p(s_{jt}|\Omega_t)}{p(s_{i,t+3}|\Omega_{t+2})}\left(\mathbf{P}'\mathbf{A}_{t+1}\mathbf{P}'\mathbf{A}_{t+2}\mathbf{P}'\right)_{j,i}
\end{aligned}$$

where $a_{t+1}^k = \frac{p(s_{k,t+1}=1|\Omega_{t+1})}{p(s_{k,t+1}=1|\Omega_t)}$. On the second line we use that the regime s_t depends on future observations only through s_{t+1}.

2.A.3 The MSFE with exogenous regressors

The expected MSFE is given by

$$
\begin{aligned}
\mathrm{E}(\sigma_m^{-2} e_{T+1}^2) &= \sum_{i=1}^{m} \mathrm{E}(s_{i,T+1}) \mathbf{x}_{T+1}' \mathbf{\Lambda}_{ij} \mathbf{x}_{T+1} + \sum_{i=1}^{m} \mathrm{E}(s_{i,T+1}) q_i^2 \varepsilon_{T+1}^2 \\
&+ \mathbf{x}_{T+1}' (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1} \sum_{i=1}^{m} \sum_{j=1}^{m} \mathrm{E}[(\mathbf{X}'\mathbf{W}\mathbf{S}_i\mathbf{X}) \mathbf{\Lambda}_{ij} (\mathbf{X}'\mathbf{S}_j\mathbf{W}\mathbf{X})] (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1} \mathbf{x}_{T+1} \\
&+ \mathbf{x}_{T+1}' (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1} \sum_{i=1}^{m} q_i^2 \mathbf{X}'\mathbf{W}\mathrm{E}(\mathbf{S}_i)\mathbf{W}\mathbf{X} (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1} \mathbf{x}_{T+1} \\
&- 2\mathbf{x}_{T+1}' (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1} \sum_{i=1}^{m} \sum_{j=1}^{m} \mathrm{E}\left(\mathbf{X}'\mathbf{W}\mathbf{S}_i\mathbf{X}\mathbf{\Lambda}_{ij} s_{j,T+1}\right) \mathbf{x}_{T+1}
\end{aligned}
$$

2.B Additional Monte Carlo results

2.B.1 Monte Carlo results for $T = 50$ and $T = 100$

Tables 2.8 and 2.9 report the results for the mean only model for $T = 50$ and 100 and complement the results in Tables 2.3 and 2.4 in this chapter.

2.B.2 Exogenous regressors

In this set of experiments, we use the set up of the experiments of the mean only, two state model and add an exogenous regressor to the model, such that $\mathbf{x}_t = [1, z_t]'$, $z_t \sim \mathrm{N}(0, \sigma_z^2)$ and $\sigma_z = 1/2$ is chosen such that the centered R^2 is of a similar magnitude to the model with a constant only. The latter requirement is due to the fact that an important determinant of the quality of the forecasts is how well identified the states are and increasing the R^2 would improve the identification.

Table 2.10 displays the results for models that include an exogenous regressor. The optimal forecast are obtained by using an asymptotic approximation to the covariance matrix in (2.34). As the ratio of parameters to estimate versus the number of observations increases, the performance of the optimal weights $w_{\hat{s}}$ is less pronounced but the differences are generally small and the conclusions from experiments with mean only models carry over to the case of exogenous regressors.

2.B.3 Monte Carlo results for MSI and MSM models

Table 2.11 presents Monte Carlo results for the models that are frequently used in empirical applications. These models are the m-state Markov switching in intercept (MSI) and Markov switching in mean (MSM) models which include p lags of the dependent variable. We analyze the performance of the optimal weights for an MSI(2)-AR(2) and MSM(2)-AR(2) model. For both models, Table 2.11 shows that the improvements by using optimal weights are consistent with the results for the Markov switching model with no lagged dependent variables. However, the additional parameter estimates imply noise that leads to slightly less pronounced differences in MSFE compared to the intercept only model.

Table 2.8: Monte Carlo results: two states, mean only models

		$T = 50$			$T = 100$		
λ	$\tilde{\sigma}^2_{\hat{\xi}_{\|T}}$	$w_{\hat{s}}$	$w_{\hat{\xi}}$	$w_{\hat{\mathbf{M}}}$	$w_{\hat{s}}$	$w_{\hat{\xi}}$	$w_{\hat{\mathbf{M}}}$
Switches in mean							
1	0.0-0.1	0.982	1.008	1.008	0.993	1.005	1.005
	0.1-0.2	0.991	1.026	1.030	0.997	1.013	1.022
	0.2-0.3	0.996	1.034	1.039	0.999	1.019	1.032
	0.3-0.4	0.999	1.036	1.042	1.000	1.024	1.037
2	0.0-0.1	0.996	1.009	1.017	0.999	1.005	1.023
	0.1-0.2	1.001	1.009	1.025	1.002	0.994	1.034
	0.2-0.3	1.004	0.983	1.002	1.003	0.977	1.004
	0.3-0.4	1.005	0.961	0.977	1.004	0.960	0.973
3	0.0-0.1	1.000	0.997	1.009	1.000	0.997	1.022
	0.1-0.2	1.004	0.969	0.999	1.005	0.961	0.993
	0.2-0.3	1.007	0.926	0.950	1.007	0.920	0.944
	0.3-0.4	1.009	0.890	0.907	1.007	0.892	0.912
Switches in mean and variance ($q^2 = 2$)							
1	0.0-0.1	0.984	1.001	1.002	0.992	1.001	1.001
	0.1-0.2	0.990	1.016	1.018	0.996	1.009	1.013
	0.2-0.3	0.996	1.029	1.032	0.999	1.014	1.021
	0.3-0.4	1.000	1.028	1.034	1.001	1.018	1.026
2	0.0-0.1	0.993	1.009	1.011	0.998	1.005	1.019
	0.1-0.2	0.999	1.015	1.028	1.002	0.999	1.030
	0.2-0.3	1.002	1.003	1.018	1.003	0.992	1.021
	0.3-0.4	1.006	0.983	0.998	1.003	0.987	1.003
3	0.0-0.1	0.998	1.003	1.016	1.000	0.999	1.027
	0.1-0.2	1.004	0.985	1.011	1.003	0.980	1.025
	0.2-0.3	1.007	0.953	0.971	1.007	0.946	0.962
	0.3-0.4	1.009	0.929	0.942	1.007	0.920	0.939

Note: The table reports the ratio of the MSFE of the optimal weights to that of the Markov switching weights. $y_t = \beta_1 s_{1t} + \beta_2 s_{2t} + (\sigma_1 s_{1t} + \sigma_2 s_{2t})\varepsilon_t$ where $\varepsilon_t \sim \text{N}(0,1)$, $\sigma_2^2 = 0.25$, $q^2 = \sigma_1^2/\sigma_2^2$. Column labels: $\lambda = (\beta_2 - \beta_1)/\sigma_2$, $\tilde{\sigma}^2_{\hat{\xi}_{|T}}$ is the normalized variance in of the smoothed probability vector (2.35). $w_{\hat{s}}$: forecasts from weights based on estimated parameters and state probabilities. $w_{\hat{\xi}}$: forecasts from weights conditional on state probabilities. $w_{\hat{\mathbf{M}}}$ are the weights based on numerically inverting $\hat{\mathbf{M}}$.

Table 2.9: Monte Carlo results: three states, intercept only models

		$T = 50$			$T = 100$		
$\{\lambda_{31}, \lambda_{21}\}$	$\tilde{\sigma}^2_{\hat{\xi}_{\vert T}}$	$w_{\hat{s}}$	$w_{\hat{\xi}}$	$w_{\hat{\mathbf{M}}}$	$w_{\hat{s}}$	$w_{\hat{\xi}}$	$w_{\hat{\mathbf{M}}}$
{2,1}	0.0-0.1	0.996	1.035	1.033	0.998	1.025	1.027
	0.1-0.2	0.998	1.033	1.037	0.999	1.027	1.046
	0.2-0.3	0.999	1.027	1.032	1.000	1.012	1.027
	0.3-0.4	1.001	1.017	1.025	1.001	1.007	1.018
{3,1}	0.0-0.1	0.998	1.020	1.016	0.999	1.011	1.026
	0.1-0.2	1.000	1.011	1.013	1.001	0.998	1.013
	0.2-0.3	1.002	0.991	0.993	1.002	0.971	0.986
	0.3-0.4	1.004	0.962	0.967	1.003	0.939	0.953
{3.5,2}	0.0-0.1	0.999	1.014	1.013	1.000	1.009	1.013
	0.1-0.2	1.000	1.004	1.003	1.001	0.994	1.008
	0.2-0.3	1.002	0.983	0.988	1.002	0.964	0.979
	0.3-0.4	1.004	0.946	0.947	1.003	0.933	0.946

Note: The table reports the ratio of the MSFE of the optimal weights to that of the Markov switching weights. For details see Table 2.3.

Table 2.10: Monte Carlo results: two states, models with exogenous regressors

		$T = 50$			$T = 100$		
λ	$\tilde{\sigma}^2_{\hat{\xi}_{\vert T}}$	$w_{\hat{s}}$	$w_{\hat{\xi}}$	$w_{\hat{\mathbf{M}}}$	$w_{\hat{s}}$	$w_{\hat{\xi}}$	$w_{\hat{\mathbf{M}}}$
1	0.0-0.1	0.962	0.988	0.986	0.986	1.002	1.002
	0.1-0.2	0.973	1.021	1.001	0.993	1.014	1.018
	0.2-0.3	0.991	1.025	1.021	0.999	1.023	1.028
	0.3-0.4	0.995	1.030	1.028	1.000	1.026	1.032
2	0.0-0.1	0.990	1.000	1.002	0.999	1.003	1.013
	0.1-0.2	1.004	1.008	1.016	1.006	0.997	1.031
	0.2-0.3	1.011	0.999	1.013	1.011	0.978	1.009
	0.3-0.4	1.012	0.986	0.999	1.019	0.956	0.991
3	0.0-0.1	1.005	1.004	1.013	1.005	1.001	1.027
	0.1-0.2	1.018	0.998	1.026	1.020	0.979	1.033
	0.2-0.3	1.031	0.983	1.010	1.043	0.935	1.008
	0.3-0.4	1.020	0.969	0.991	1.051	0.919	0.958

Note: The table reports the ratio of the MSFE of the optimal asymptotic weights to that of the Markov switching weights. DGP: $y_t = x_t'\beta_1 + \sigma\left(x_t'\lambda s_{2t} + \varepsilon_t\right)$ where $\varepsilon_t \sim \text{NID}(0,1)$. Also $\sigma^2 = 0.25$, $\beta_1 = 1$ and $x_t = [1, z_t]$ where $z_t \sim \text{N}(0, 0.25)$. For the column labels see the footnote of Table 2.3.

Table 2.11: Monte Carlo results: MSI and MSM models

		$T = 50$			$T = 100$		
λ	$\tilde{\sigma}^2_{\hat{\xi}_{\|T}}$	$w_{\hat{s}}$	$w_{\hat{\xi}}$	$w_{\hat{\mathbf{M}}}$	$w_{\hat{s}}$	$w_{\hat{\xi}}$	$w_{\hat{\mathbf{M}}}$
MSI							
1	0.0-0.1	0.988	1.008	1.002	0.994	1.006	1.006
	0.1-0.2	0.994	1.019	1.016	0.997	1.016	1.020
	0.2-0.3	0.997	1.018	1.018	0.999	1.017	1.026
2	0.0-0.1	0.997	1.005	1.006	0.999	1.003	1.020
	0.1-0.2	1.000	1.005	1.017	1.002	0.994	1.030
	0.2-0.3	1.003	0.993	1.007	1.003	0.985	1.018
3	0.0-0.1	1.000	0.999	1.004	1.000	0.999	1.012
	0.1-0.2	1.004	0.983	1.026	1.004	0.972	1.020
	0.2-0.3	1.005	0.970	0.986	1.005	0.944	0.981
MSM							
1	0.0-0.1	0.991	1.010	1.008	0.994	1.019	1.020
	0.1-0.2	0.994	1.023	1.017	0.996	1.033	1.042
	0.2-0.3	0.995	1.029	1.037	0.998	1.033	1.043
2	0.0-0.1	0.996	1.011	1.009	0.999	1.012	1.028
	0.1-0.2	0.998	1.015	1.019	1.000	1.010	1.034
	0.2-0.3	0.999	1.015	1.022	1.001	1.007	1.024
3	0.0-0.1	0.999	1.004	1.004	1.000	1.002	1.015
	0.1-0.2	1.000	1.002	1.013	1.002	0.991	1.012
	0.2-0.3	1.000	1.006	1.007	1.003	0.974	0.983

Note: The table reports the ratio of the MSFE of the optimal weights to that of the Markov switching weights. DGP MSI: $y_t = \beta_1 s_{1t} + \beta_2 s_{2t} + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \sigma\varepsilon_t$ where $\varepsilon_t \sim \text{N}(0,1)$. DGP MSM: $y_t = \beta_1 s_{1t} + \beta_2 s_{2t} + \phi_1(y_{t-1} - \beta_{s_{t-1}}) + \phi_2(y_{t-2} - \beta_{s_{t-2}}) + \sigma\varepsilon_t$, $\sigma^2 = 0.25$, $\phi_1 = 0.4$, $\phi_2 = -0.3$. Column labels as in Table 2.3.

Chapter 3

A near optimal test for structural breaks when forecasting under squared error loss

3.1 Introduction

Structural breaks present a major challenge to forecasters as they require information about the time of the break and parameter estimates for the post-break sample. However, often these can be estimated only imprecisely (Elliott and Müller, 2007, 2014). Furthermore, forecasts are typically evaluated using mean square forecast error loss, which implies a bias-variance trade-off and suggests that ignoring small breaks will lead to more accurate forecasts than incorporating them into the model (Pesaran and Timmermann, 2005). If sufficiently small breaks can be ignored, the question is: what constitutes sufficiently small?

We develop a test for equal forecast accuracy that compares the expected mean square forecast error (MSFE) of a one-step-ahead forecast from the post-break sample to that of a forecast that uses the full sample. Under a known break date, the break size for which post-break sample and full sample forecasts achieve equal predictive accuracy is one standard deviation of the distribution of the parameter estimates. Under a local break of unknown timing, the uncertainty around the break date increases the variance of the post-break sample forecast and the break size of equal forecast accuracy is much larger, up to three standard deviations in terms of the distribution of the parameter estimates.

Building on the work of Andrews (1993) and Piterbarg (1996), we derive a test for the critical break size, which is optimal as the size tends to zero. Simulations of asymptotic power show that our test is near optimal for conventional choices of the nominal size, which is largely due to the size of the breaks that are allowed under the null. In the process, we show that the near optimality of the test follows from an optimality argument of the estimated break date by maximizing a Wald test statistic. This optimality does not depend on whether

the Wald-statistic is used in its homoskedastic form or whether a heteroskedastic version is used, as long as the estimator of the variance is consistent. We also show that post-test inference following a rejection remains standard if the size of the test is small.

While our test uses much of the asymptotic framework of Andrews (1993), it is substantially different from extant break point tests, such as those of Ploberger et al. (1989), Andrews (1993), Andrews and Ploberger (1994), Elliott and Müller (2007, 2014), and Elliott et al. (2015b). While those tests focus on the difference between (sub-)sets of parameters of a model before and after a break date, our measure is the forecast accuracy of the entire model. Our test, therefore, allows for a break in the parameters under the null of equal forecast accuracy.

In line with much of the forecasting literature, our loss function is the mean squared forecast error. Like the work of Trenkler and Toutenburg (1992) and Clark and McCracken (2012), our test is inspired by the in-sample MSE test of Toro-Vizcarrondo and Wallace (1968) and Wallace (1972). However, compared to Trenkler and Toutenburg (1992) and Clark and McCracken (2012), our test is much simpler in that, under a known break date, our test statistic has a known distribution that is free of unknown parameters.

Our test is different from forecast accuracy tests of the kind suggested by Diebold and Mariano (1995) and extended by, among others, Clark and McCracken (2001); a recent review is by Clark and McCracken (2013). These tests assess forecast accuracy *ex post*. In contrast, the test we propose in this chapter is an *ex ante* test of the accuracy of forecasts of models that do or do not account for breaks.

Giacomini and Rossi (2009) assess forecast breakdowns in the sense that the forecast performance of a model is not in line with the in-sample fit of the model. They consider forecast breakdowns in historically made forecasts as well as prediction of forecast breakdowns. Our approach is more targeted asking whether a structural beak, which is one of the possible sources of forecast breakdown, needs to be addressed from a forecast perspective.

The competing forecasts in our test are those using the full sample and using the post-break sample. Recently, Pesaran et al. (2013) showed that forecasts based on post-break samples can be improved by using all observations and weighting them such that the MSFE is minimized. We show that this estimator can be written as a shrinkage estimator in the tradition of Thompson (1968), where the shrinkage estimator averages between the full sample estimator and post-break sample estimator with a weight that is equivalent to the test statistic introduced in this chapter.

Under a known break date, the performance of shrinkage estimators is well known, see for example Magnus (2002). However, their properties depend critically on the fact that the break date is known, which implies that the estimator from the post-break sample is unbiased. Under local breaks, this may not be the case and the forecasting performance of the shrinkage estimator compared to the full sample forecast is not immediately clear.

Since the shrinkage estimator does not take break date uncertainty into account, it will likely put too much weight on the post-break sample forecast. We find that for small break sizes, where the break date is not accurately identified, the shrinkage forecast is less accurate than the full sample forecast. However, compared to the post-break sample forecast, we find that the shrinkage estimator is almost uniformly more accurate. We propose a second version of our test that compares the forecast accuracy of the shrinkage estimator and the full sample forecast.

Substantial evidence of structural breaks has been found in macroeconomic and financial time series by Stock and Watson (1996), Rapach and Wohar (2006), Rossi (2006), Paye and Timmermann (2006), and others. Hence, we apply our test to the macroeconomic and financial time series and use the FRED-MD data set by McCracken and Ng (2015) We find that breaks that are important for forecasting under MSFE loss are between a factor two to three less frequent than the usual sup-Wald test by Andrews (1993) would indicate. Incorporating only the breaks suggested by our test substantially reduces the average MSFE in this data set compared to the forecasts that take all breaks suggested by Andrew's sup-Wald test into account.

This chapter is structured as follows. We start with the motivating example of the linear regression model with one break of known timing in Section 3.2. The model is generalized in Section 3.3 using the methodology of Andrews (1993). In Section 3.4, we derive the test and show its weak optimality. We extend the test to cover the optimal weights or shrinkage forecast in Section 3.4.4 . Simulation results in Section 3.5 shows that the weak optimality of the test is in fact quite strong, with power very close to the optimal, but infeasible test that knows the true break date. Finally, the application of our tests to the large set of time series in the FRED-MD data set is presented in Section 3.6.

3.2 Motivating example: structural break of known timing in a linear model

In order to gain intuition, initially consider a linear regression model with a structural break at time T_b

$$y_t = \boldsymbol{x}_t'\boldsymbol{\beta}_t + \varepsilon_t, \qquad \varepsilon_t \sim \text{i.i.d.}(0, \sigma^2) \tag{3.1}$$

where

$$\boldsymbol{\beta}_t = \begin{cases} \boldsymbol{\beta}_1 & \text{if } t \leqslant T_b \\ \boldsymbol{\beta}_2 & \text{if } t > T_b \end{cases} \tag{3.2}$$

$\boldsymbol{x}_t$ is a $k \times 1$ vector of exogenous regressors, $\boldsymbol{\beta}_i$ a $k \times 1$vector of parameters, and the break date, T_b, is initially assumed to be known. The parameter vectors $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ can be estimated

by OLS on the two subsamples. If the break is ignored, a single set of parameter estimates, $\hat{\boldsymbol{\beta}}_F$, can be obtained using OLS on the full sample.

Denote $\boldsymbol{V}_i = (T_i - T_{i-1})\text{Var}(\hat{\boldsymbol{\beta}}_i)$, for $i = 1, 2$, $T_0 = 0$, $T_1 = T_b$, $T_2 = T$ and $\boldsymbol{V}_F = T\text{Var}(\hat{\boldsymbol{\beta}}_F)$ as the covariance matrices of the vectors of coefficient estimates. Initially, assume these matrices to be known; later they will be replaced by their probability limits.

In this chapter, we would like to test whether the expected mean squared forecast error (MSFE) from the forecast using the full sample, $\hat{y}^F_{T+1} = \mathbf{x}'_{T+1}\boldsymbol{\beta}_F$, is smaller than that of the post-break sample, $\hat{y}^P_{T+1} = \mathbf{x}'_{T+1}\boldsymbol{\beta}_2$.

The MSFE for the forecast from the post-break sample parameter estimate, $\boldsymbol{\beta}_2$, is

$$\begin{aligned} R(\boldsymbol{x}'_{T+1}\hat{\boldsymbol{\beta}}_2) &= \text{E}\left[\left(\boldsymbol{x}'_{T+1}\hat{\boldsymbol{\beta}}_2 - \boldsymbol{x}'_{T+1}\boldsymbol{\beta}_2 - \varepsilon_{T+1}\right)^2\right] \\ &= \frac{1}{T - T_b}\boldsymbol{x}_{T+1}\boldsymbol{V}_2\boldsymbol{x}_{T+1} + \sigma^2 \end{aligned} \tag{3.3}$$

and that using the full sample estimate, $\boldsymbol{\beta}_F$, is

$$\begin{aligned} R(\boldsymbol{x}'_{T+1}\hat{\boldsymbol{\beta}}_F) &= \text{E}\left[\left(\boldsymbol{x}'_{T+1}\hat{\boldsymbol{\beta}}_F - \boldsymbol{x}'_{T+1}\boldsymbol{\beta}_2 - \varepsilon_{T+1}\right)^2\right] \\ &= \text{E}\left[\left(\boldsymbol{x}'_{T+1}\hat{\boldsymbol{\beta}}_F - \boldsymbol{x}'_{T+1}\boldsymbol{\beta}_2\right)\right]^2 + \frac{1}{T}\boldsymbol{x}'_{T+1}\boldsymbol{V}_F\boldsymbol{x}_{T+1} + \sigma^2 \\ &= \left[\frac{T_b}{T}\boldsymbol{x}'_{T+1}\boldsymbol{V}_F\boldsymbol{V}_1^{-1}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)\right]^2 + \frac{1}{T}\boldsymbol{x}'_{T+1}\boldsymbol{V}_F\boldsymbol{x}_{T+1} + \sigma^2 \end{aligned} \tag{3.4}$$

Comparing (3.3) and (3.4), we see that the full sample forecast is at least as accurate as the post-break sample forecast if

$$\begin{aligned} \zeta &= T\tau_b^2 \frac{\left[\boldsymbol{x}'_{T+1}\boldsymbol{V}_F\boldsymbol{V}_1^{-1}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)\right]^2}{\boldsymbol{x}'_{T+1}\left(\frac{\boldsymbol{V}_2}{1-\tau_b} - \boldsymbol{V}_F\right)\boldsymbol{x}_{T+1}} \\ &\xrightarrow{p} T\tau_b(1-\tau_b)\frac{\left[\boldsymbol{x}'_{T+1}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)\right]^2}{\boldsymbol{x}'_{T+1}\boldsymbol{V}\boldsymbol{x}_{T+1}} \\ &\leqslant 1 \end{aligned} \tag{3.5}$$

where $\tau_b = T_b/T$ and the third line assumes that the covariance matrices asymptotically satisfy $\text{plim}_{T\to\infty}\boldsymbol{V}_i \to \boldsymbol{V}$ for $i = 1, 2, F$.

From (3.5) it can be observed that, under $H_0 : \zeta = 1$, the size of the break $\boldsymbol{x}'_{T+1}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)$ is symmetric in ζ. Additionally, (3.3) suggests that breaks that occur at the end of the sample will lead to a larger mean squared forecast error than breaks that occur at the beginning.

To test $H_0 : \zeta = 1$ note that

$$\begin{aligned}\hat{\zeta}(\tau) &= T\tau^2 \frac{\left[\boldsymbol{x}'_{T+1}\hat{\boldsymbol{V}}_F\hat{\boldsymbol{V}}_1^{-1}(\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_2)\right]^2}{\boldsymbol{x}'_{T+1}\left(\frac{\hat{\boldsymbol{V}}_2}{1-\tau} - \hat{\boldsymbol{V}}_F\right)\boldsymbol{x}_{T+1}} \\ &= \frac{\left[\boldsymbol{x}'_{T+1}(\hat{\boldsymbol{\beta}}_F - \hat{\boldsymbol{\beta}}_2)\right]^2}{\boldsymbol{x}'_{T+1}\widehat{\text{Var}}(\hat{\boldsymbol{\beta}}_F - \hat{\boldsymbol{\beta}}_2)\boldsymbol{x}_{T+1}} \xrightarrow{d} \chi^2(1,\zeta)\end{aligned} \tag{3.6}$$

with $\xrightarrow{d}$ denoting convergence in distribution and a consistent estimator of the covariance matrix is used in the denominator.

A more conventional and asymptotically equivalent form of the test statistic is

$$\hat{\zeta}(\tau) = T\frac{\left[\boldsymbol{x}'_{T+1}(\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_2)\right]^2}{\boldsymbol{x}'_{T+1}\left(\frac{\hat{\boldsymbol{V}}_1}{\tau} + \frac{\hat{\boldsymbol{V}}_2}{1-\tau}\right)\boldsymbol{x}_{T+1}} \xrightarrow{d} \chi^2(1,\zeta) \tag{3.7}$$

This is a standard Wald test using the regressors at $t = T + 1$ as weights.

The results of the test will, in general, differ from the outcomes of the classical Wald test on the difference between the parameter vectors $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ for two reasons. The first is that the multiplication by $\boldsymbol{x}_{T+1}$ can render large breaks irrelevant for forecasting, or small breaks relevant. The first scenario is more likely due to the fact that breaks in the coefficients of $\boldsymbol{\beta}$ potentially cancel in the inner product $\boldsymbol{x}'_{T+1}\boldsymbol{\beta}$. The second reason is that under $H_0 : \zeta = 1$, we compare the test statistic against the critical values of the non-central χ^2-distribution, instead of the central χ^2-distribution. The critical values of these distributions differ substantially: the $\alpha = 0.05$ critical value of the $\chi^2(1)$ is 3.84 and that of the $\chi^2(1,1)$ is 7.00.

As is clear from (3.5), if the difference in the parameters, $\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2$, converges to zero at a rate $T^{-1/2+\epsilon}$ for some $\epsilon > 0$, then the test statistic diverges to infinity as $T \to \infty$, which is unlikely to reflect the uncertainty surrounding the break date in empirical applications. We will therefore consider breaks that are local in nature, i.e. $\boldsymbol{\beta_2} = \boldsymbol{\beta}_1 + \frac{1}{\sqrt{T}}\boldsymbol{\eta}$, rendering a finite test statistic in the asymptotic limit. Local breaks have been intensively studied in the recent literature, see for example Elliott and Müller (2007, 2014) and Elliott et al. (2015b). An implication of local breaks is that no consistent estimator for the break date is available, which mimics practical situations. A consequence is that post-break parameters cannot be consistently estimated. This will deteriorate post-break window forecasts compared the full sample forecast, which, in turn, increases the break size that yields equal forecasting performance between full and post-break sample estimation windows.

3.3 General model set-up and estimation

We consider a general, possibly non-linear, parametric model, where parameters are estimated using GMM. The general estimation framework is that used by Andrews (1993). The observed data are given by a triangular array of random variables $\{\boldsymbol{W}_t = (\boldsymbol{Y}_t, \boldsymbol{X}_t) : 1 \leqslant t \leqslant T\}$, $\boldsymbol{Y}_t = (y_1, \dots y_t)$ and $\boldsymbol{X}_t = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_t)'$. Assumptions can be made with regard to the dependency of $\boldsymbol{W}_t$ such that the results below apply to a wide range of time series models. We make the following additional assumption on the noise and the relation between y_t, lagged values of y_t and exogenous regressors $\boldsymbol{x}_t$

Assumption 1 *The model for the dependent variable y_t consists of a signal and additive noise*

$$y_t = f(\boldsymbol{\beta}_t, \boldsymbol{\delta}; \boldsymbol{X}_t, \boldsymbol{Y}_{t-1}) + \varepsilon_t \tag{3.8}$$

where the function f is fixed and differentiable with respect to the parameter vector $\boldsymbol{\theta}_t = (\boldsymbol{\beta}_t', \boldsymbol{\delta}')'$.

In the model (3.8), the parameter vector $\boldsymbol{\delta}$ is known to be constant. The parameter vector $\boldsymbol{\beta}_t$ could be subject to a structural break. When ignoring the break, parameters are estimated by minimizing the sample analogue of the population moment conditions

$$\frac{1}{T}\sum_{t=1}^{T} \mathrm{E}[m(\boldsymbol{W}_t, \boldsymbol{\beta}, \boldsymbol{\delta})] = 0$$

which requires solving

$$\begin{aligned} &\frac{1}{T}\sum_{t=1}^{T} m(\boldsymbol{W}_t, \hat{\boldsymbol{\beta}}_F, \hat{\boldsymbol{\delta}})'\hat{\boldsymbol{\gamma}}\frac{1}{T}\sum_{t=1}^{T} m(\boldsymbol{W}_t, \hat{\boldsymbol{\beta}}_F, \hat{\boldsymbol{\delta}}) = \\ &\inf_{\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\delta}}} \frac{1}{T}\sum_{t=1}^{T} m(\boldsymbol{W}_t, \tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\delta}})'\hat{\boldsymbol{\gamma}}\frac{1}{T}\sum_{t=1}^{T} m(\boldsymbol{W}_t, \tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\delta}}) \end{aligned} \tag{3.9}$$

where $\hat{\boldsymbol{\beta}}_F$ is estimator based on the full estimation window. We assume throughout the weighting matrix $\boldsymbol{\gamma} = \boldsymbol{S}^{-1}$ and

$$\boldsymbol{S} = \lim_{T\to\infty} \mathrm{Var}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} m(\boldsymbol{W}_t, \boldsymbol{\beta}, \boldsymbol{\delta})\right)$$

for which a consistent estimator is assumed to be available.

As discussed above, we consider a null hypothesis that allows local breaks, defined by

$$\boldsymbol{\beta}_t = \boldsymbol{\beta}_1 + \frac{1}{\sqrt{T}}\boldsymbol{\eta}(\tau)$$

where $\boldsymbol{\eta}(\tau) = \boldsymbol{b}I[\tau < \tau_b]$, $\boldsymbol{b}$ is a vector of constants, and $\tau = t/T$. The pre-break parameter vector, $\boldsymbol{\beta}_1$, and the post-break parameter vector, $\boldsymbol{\beta}_2$, satisfy the partial sample moment conditions

$$\frac{1}{\tau T}\sum_{t=1}^{T} m(\boldsymbol{W}_t, \boldsymbol{\beta}_1, \boldsymbol{\delta}) = \mathbf{0}, \quad \text{and} \quad \frac{1}{T}\sum_{t=T\tau+1}^{T} m(\boldsymbol{W}_t, \boldsymbol{\beta}_2, \boldsymbol{\delta}) = \mathbf{0}$$

Define

$$\bar{m}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\delta}, \tau) = \frac{1}{T}\sum_{t=1}^{T\tau}\begin{pmatrix} m(\boldsymbol{W}_t, \boldsymbol{\beta}_1, \boldsymbol{\delta}) \\ \mathbf{0} \end{pmatrix} + \frac{1}{T}\sum_{t=T\tau+1}^{T}\begin{pmatrix} \mathbf{0} \\ m(\boldsymbol{W}_t, \boldsymbol{\beta}_2, \boldsymbol{\delta}) \end{pmatrix}$$

then, partial sum GMM estimators can be obtained by solving (3.9) with $m(\cdot)$ replaced by $\bar{m}(\cdot)$ and $\hat{\boldsymbol{\gamma}}$ replaced by

$$\hat{\gamma}(\tau) = \begin{pmatrix} \frac{1}{\tau}\hat{\boldsymbol{S}}^{-1} & \mathbf{0} \\ \mathbf{0} & \frac{1}{1-\tau}\hat{\boldsymbol{S}}^{-1} \end{pmatrix}$$

The aim is to determine whether the full sample estimates lead to a more precise forecast in the mean square forecast error sense than the post-break sample estimates. The forecasts are constructed as

$$\begin{aligned} \hat{y}^F_{T+1} &= f(\hat{\boldsymbol{\beta}}_F, \hat{\boldsymbol{\delta}}; \boldsymbol{X}_t, \boldsymbol{Y}_{t-1}) \\ \hat{y}^P_{T+1} &= f(\hat{\boldsymbol{\beta}}_2, \hat{\boldsymbol{\delta}}; \boldsymbol{X}_t, \boldsymbol{Y}_{t-1}) \end{aligned} \tag{3.10}$$

Throughout, we condition on the both the exogenous and lagged dependent variables that are needed to construct the forecast. The comparison between $\hat{y}^F_{T+1}$ and $\hat{y}^P_{T+1}$ is non-standard as, under a local break, even the parameters of the model that incorporates the break are inconsistent.

In order to compare the forecasts in (3.10), we start by providing the asymptotic properties of the estimators in a model that incorporates the break and in a model that ignores the break. Proofs for weak convergence of the estimators towards Gaussian processes indexed by the break date τ are given by Andrews (1993). The asymptotic distributions depend on the following matrices, for which consistent estimators are assumed to be available,

$$\begin{aligned} \boldsymbol{M} &= \lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\mathrm{E}\left[\frac{\partial m(\boldsymbol{W}_t, \boldsymbol{\beta}, \boldsymbol{\delta})}{\partial \boldsymbol{\beta}}\right] \\ \boldsymbol{M}_\delta &= \lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\mathrm{E}\left[\frac{\partial m(\boldsymbol{W}_t, \boldsymbol{\beta}, \boldsymbol{\delta})}{\partial \boldsymbol{\delta}}\right] \end{aligned}$$

To simplify the notation, define

$$\begin{aligned}\bar{\boldsymbol{X}}' &= \boldsymbol{M}'\boldsymbol{S}^{-1/2}\\ \bar{\boldsymbol{Z}}' &= \boldsymbol{M}_\delta'\boldsymbol{S}^{-1/2}\end{aligned}$$

Partial sample estimator: The partial sample estimators converge to the following Gaussian process indexed by τ

$$\begin{aligned}\sqrt{T}\begin{pmatrix}\hat{\boldsymbol{\beta}}_1(\tau)-\boldsymbol{\beta}_2\\ \hat{\boldsymbol{\beta}}_2(\tau)-\boldsymbol{\beta}_2\\ \hat{\boldsymbol{\delta}}-\boldsymbol{\delta}\end{pmatrix} \to &\begin{bmatrix}\tau\bar{\boldsymbol{X}}'\bar{\boldsymbol{X}} & \boldsymbol{0} & \tau\bar{\boldsymbol{X}}'\bar{\boldsymbol{Z}}\\ \boldsymbol{0} & (1-\tau)\bar{\boldsymbol{X}}'\bar{\boldsymbol{X}} & (1-\tau)\bar{\boldsymbol{X}}'\bar{\boldsymbol{Z}}\\ \tau\bar{\boldsymbol{Z}}'\bar{\boldsymbol{X}} & (1-\tau)\bar{\boldsymbol{Z}}'\bar{\boldsymbol{X}} & \bar{\boldsymbol{Z}}'\bar{\boldsymbol{Z}}\end{bmatrix}^{-1}\\ &\times\begin{bmatrix}\bar{\boldsymbol{X}}'\boldsymbol{B}(\tau)+\bar{\boldsymbol{X}}'\bar{\boldsymbol{X}}\int_0^\tau\boldsymbol{\eta}(s)ds\\ \bar{\boldsymbol{X}}'[\boldsymbol{B}(1)-\boldsymbol{B}(\tau)]+\bar{\boldsymbol{X}}'\bar{\boldsymbol{X}}\int_\tau^1\boldsymbol{\eta}(s)ds\\ \bar{\boldsymbol{Z}}'\boldsymbol{B}(1)+\bar{\boldsymbol{Z}}'\bar{\boldsymbol{X}}\int_0^1\boldsymbol{\eta}(s)ds\end{bmatrix}\end{aligned} \tag{3.11}$$

where $\boldsymbol{B}(\tau)$ is a Brownian motion defined on the interval $[0,1]$. In line with Andrews (1993) we subtract $\boldsymbol{\beta}_2$ from both estimators $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$. This lines up with our interest in forecasting future observations, which are functions of $\boldsymbol{\beta}_2$ only, and the remainder that arises if $\tau \neq \tau_b$, is absorbed in the integral on the right hand side.

Define the projection matrix that projects onto the columns of $\bar{\boldsymbol{X}}$ as $\boldsymbol{P}_{\bar{X}} = \bar{\boldsymbol{X}}(\bar{\boldsymbol{X}}'\bar{\boldsymbol{X}})^{-1}\bar{\boldsymbol{X}}'$, its orthogonal complement as $\boldsymbol{M}_{\bar{X}} = \boldsymbol{I} - \boldsymbol{P}_{\bar{X}}$ and, additionally,

$$\begin{aligned}\boldsymbol{V} &= (\bar{\boldsymbol{X}}'\bar{\boldsymbol{X}})^{-1}\\ \boldsymbol{Q} &= \bar{\boldsymbol{Z}}'\boldsymbol{M}_{\bar{X}}\bar{\boldsymbol{Z}}\\ \boldsymbol{L} &= (\bar{\boldsymbol{X}}'\bar{\boldsymbol{X}})^{-1}\bar{\boldsymbol{X}}'\bar{\boldsymbol{Z}}(\bar{\boldsymbol{Z}}'\boldsymbol{M}_{\bar{X}}\bar{\boldsymbol{Z}})^{-1}\\ \tilde{\boldsymbol{Q}} &= (\bar{\boldsymbol{X}}'\bar{\boldsymbol{X}})^{-1}\bar{\boldsymbol{X}}'\bar{\boldsymbol{Z}}(\bar{\boldsymbol{Z}}'\boldsymbol{M}_{\bar{X}}\bar{\boldsymbol{Z}})^{-1}\bar{\boldsymbol{Z}}'\bar{\boldsymbol{X}}(\bar{\boldsymbol{X}}'\bar{\boldsymbol{X}})^{-1}\end{aligned} \tag{3.12}$$

The inverse in (3.11) can be calculated using blockwise inversion. The result is the asymptotic variance covariance matrix of $\left(\hat{\boldsymbol{\beta}}_1(\tau)', \hat{\boldsymbol{\beta}}_2(\tau)', \hat{\boldsymbol{\delta}}'\right)'$

$$\boldsymbol{\Sigma}_P = \begin{pmatrix}\frac{1}{\tau}\boldsymbol{V}+\tilde{\boldsymbol{Q}} & \tilde{\boldsymbol{Q}} & -\boldsymbol{L}\\ \tilde{\boldsymbol{Q}} & \frac{1}{1-\tau}\boldsymbol{V}+\tilde{\boldsymbol{Q}} & -\boldsymbol{L}\\ -\boldsymbol{L}' & -\boldsymbol{L}' & \boldsymbol{Q}^{-1}\end{pmatrix}$$

Hence,

$$\sqrt{T}\begin{pmatrix} \hat{\boldsymbol{\beta}}_1(\tau) - \boldsymbol{\beta}_2 \\ \\ \hat{\boldsymbol{\beta}}_2(\tau) - \boldsymbol{\beta}_2 \\ \\ \hat{\boldsymbol{\delta}} - \boldsymbol{\delta} \end{pmatrix} \rightarrow \begin{pmatrix} \frac{1}{\tau}\left[(\bar{\boldsymbol{X}}'\bar{\boldsymbol{X}})^{-1}\bar{\boldsymbol{X}}'\boldsymbol{B}(\tau) + \int_0^\tau \boldsymbol{\eta}(s)ds\right] \\ \quad -(\bar{\boldsymbol{X}}'\bar{\boldsymbol{X}})^{-1}\bar{\boldsymbol{X}}'\bar{\boldsymbol{Z}}(\bar{\boldsymbol{Z}}'\boldsymbol{M}_{\bar{X}}\bar{\boldsymbol{Z}})^{-1}\bar{\boldsymbol{Z}}'\boldsymbol{M}_{\bar{X}}\boldsymbol{B}(1) \\ \frac{1}{1-\tau}\left[(\bar{\boldsymbol{X}}'\bar{\boldsymbol{X}})^{-1}\bar{\boldsymbol{X}}'(\boldsymbol{B}(1) - \boldsymbol{B}(\tau)) + \int_\tau^1 \boldsymbol{\eta}(s)ds\right] \\ \quad -(\bar{\boldsymbol{X}}'\bar{\boldsymbol{X}})^{-1}\bar{\boldsymbol{X}}'\bar{\boldsymbol{Z}}(\bar{\boldsymbol{Z}}'\boldsymbol{M}_{\bar{X}}\bar{\boldsymbol{Z}})^{-1}\bar{\boldsymbol{Z}}'\boldsymbol{M}_{\bar{X}}\boldsymbol{B}(1) \\ (\bar{\boldsymbol{Z}}'\boldsymbol{M}_{\bar{X}}\bar{\boldsymbol{Z}})^{-1}\bar{\boldsymbol{Z}}'\boldsymbol{M}_{\bar{X}}\boldsymbol{B}(1) \end{pmatrix} \tag{3.13}$$

Several terms can be recognized to be analogous to what would be obtained in a multivariate regression problem using the Frisch-Waugh theorem.

Full sample estimator: Estimators that ignore the break converge to

$$\sqrt{T}\begin{pmatrix} \hat{\boldsymbol{\beta}}_F - \boldsymbol{\beta}_2 \\ \hat{\boldsymbol{\delta}} - \boldsymbol{\delta} \end{pmatrix} \rightarrow \begin{bmatrix} \bar{\boldsymbol{X}}'\bar{\boldsymbol{X}} & \bar{\boldsymbol{X}}'\bar{\boldsymbol{Z}} \\ \bar{\boldsymbol{Z}}'\bar{\boldsymbol{X}} & \bar{\boldsymbol{Z}}'\bar{\boldsymbol{X}} \end{bmatrix}^{-1} \begin{bmatrix} \bar{\boldsymbol{X}}'\boldsymbol{B}(1) + \bar{\boldsymbol{X}}'\bar{\boldsymbol{X}}\int_0^1 \boldsymbol{\eta}(s)ds \\ \bar{\boldsymbol{Z}}'\boldsymbol{B}(1) + \bar{\boldsymbol{Z}}'\bar{\boldsymbol{X}}\int_0^1 \boldsymbol{\eta}(s)ds \end{bmatrix} \tag{3.14}$$

Using the notation defined in (3.12), the inverse in (3.14) can be written as

$$\boldsymbol{\Sigma}_F = \begin{pmatrix} \boldsymbol{V} + \tilde{\boldsymbol{Q}} & -\boldsymbol{L} \\ -\boldsymbol{L}' & \boldsymbol{Q}^{-1} \end{pmatrix}$$

and, therefore,

$$\sqrt{T}\begin{pmatrix} \hat{\boldsymbol{\beta}}_F - \boldsymbol{\beta}_2 \\ \\ \hat{\boldsymbol{\delta}} - \boldsymbol{\delta} \end{pmatrix} \rightarrow \begin{pmatrix} (\bar{\boldsymbol{X}}'\bar{\boldsymbol{X}})^{-1}\bar{\boldsymbol{X}}'\boldsymbol{B}(1) + \int_0^1 \boldsymbol{\eta}(s)ds \\ \quad -(\bar{\boldsymbol{X}}'\bar{\boldsymbol{X}})^{-1}\bar{\boldsymbol{X}}'\bar{\boldsymbol{Z}}(\bar{\boldsymbol{Z}}'\boldsymbol{M}_{\bar{X}}\bar{\boldsymbol{Z}})^{-1}\bar{\boldsymbol{Z}}'\boldsymbol{M}_{\bar{X}}\boldsymbol{B}(1) \\ (\bar{\boldsymbol{Z}}'\boldsymbol{M}_{\bar{X}}\bar{\boldsymbol{Z}})^{-1}\bar{\boldsymbol{Z}}'\boldsymbol{M}_{\bar{X}}\boldsymbol{B}(1) \end{pmatrix} \tag{3.15}$$

Note that for the parameters $\hat{\boldsymbol{\delta}}$, the expression is identical to partial sample estimator.

Later results require the asymptotic covariance between the estimators from the full sample and the break model, which is

$$\begin{aligned} \text{Cov}(\hat{\boldsymbol{\beta}}_2(\tau), \hat{\boldsymbol{\beta}}_F) &\xrightarrow{p} (\bar{\boldsymbol{X}}'\bar{\boldsymbol{X}})^{-1} + (\bar{\boldsymbol{X}}'\bar{\boldsymbol{X}})^{-1}\bar{\boldsymbol{X}}'\bar{\boldsymbol{Z}}(\bar{\boldsymbol{Z}}'\boldsymbol{M}_{\bar{X}}\bar{\boldsymbol{Z}})^{-1}\bar{\boldsymbol{Z}}'\bar{\boldsymbol{X}}(\bar{\boldsymbol{X}}'\bar{\boldsymbol{X}})^{-1} \\ &= \operatorname*{plim}_{T\to\infty} \text{Var}(\hat{\boldsymbol{\beta}}_F) \end{aligned}$$

which corresponds to the results by Hausman (1978) that under the null of no misspecification, a consistent and asymptotically efficient estimator should have zero covariance with its difference from an consistent but asymptotically inefficient estimator, i.e.

$$\operatorname*{plim}_{T\to\infty} \text{Cov}(\hat{\boldsymbol{\beta}}_F, \hat{\boldsymbol{\beta}}_F - \hat{\boldsymbol{\beta}}_2(\tau)) = \boldsymbol{0}$$

A difference to the case considered here is that, under a local structural break, $\hat{\boldsymbol{\beta}}_F$ and $\hat{\boldsymbol{\beta}}_2(\tau)$ are both inconsistent estimators.

3.4 Testing for a structural break

3.4.1 A break of known timing

Initially, we will assume that the break date is known in order to illustrate our approach. In a second step, we will extend the test to an unknown break date. Following Assumption 1, forecasts are obtained by applying a fixed, differentiable function to the $p+q$ parameters of the model conditional on a set of regressors of dimension $k = p+q$ by $(\boldsymbol{x}_{T+1}, \boldsymbol{z}_{T+1})$

$$\hat{y}_{T+1} = f(\hat{\boldsymbol{\beta}}_2, \hat{\boldsymbol{\delta}})$$

where we omit the dependence on the regressors for notational convenience.

For a known break date, the results of the previous section imply the following asymptotic distribution of the parameters

$$\sqrt{T}\begin{pmatrix} \hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1 \\ \hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2 \\ \hat{\boldsymbol{\delta}} - \boldsymbol{\delta} \end{pmatrix} \xrightarrow{d} N\left[\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \frac{1}{\tau}\boldsymbol{V} + \tilde{\boldsymbol{Q}} & \tilde{\boldsymbol{Q}} & -\boldsymbol{L} \\ \tilde{\boldsymbol{Q}} & \frac{1}{1-\tau}\boldsymbol{V} + \tilde{\boldsymbol{Q}} & -\boldsymbol{L} \\ -\boldsymbol{L}' & -\boldsymbol{L}' & \boldsymbol{Q}^{-1} \end{pmatrix}\right]$$

The full sample estimator satisfies

$$\lim_{T\to\infty} \hat{\boldsymbol{\beta}}_F = \lim_{T\to\infty} \left[\hat{\boldsymbol{\beta}}_2 + \tau_b(\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_2)\right]$$

and

$$\sqrt{T}\begin{pmatrix} \hat{\boldsymbol{\beta}}_F - \boldsymbol{\beta}_2 \\ \hat{\boldsymbol{\delta}} - \boldsymbol{\delta} \end{pmatrix} \xrightarrow{d} N\left[\begin{pmatrix} \tau_b(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2) \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{V} + \tilde{\boldsymbol{Q}} & -\boldsymbol{L} \\ -\boldsymbol{L}' & \boldsymbol{Q}^{-1} \end{pmatrix}\right]$$

Define $f_{\beta_2} = \frac{\partial f(\boldsymbol{\beta}_2, \boldsymbol{\delta})}{\partial \boldsymbol{\beta}_2}$ and $f_\delta = \frac{\partial f(\boldsymbol{\beta}_2, \boldsymbol{\delta})}{\partial \boldsymbol{\delta}}$. Using a Taylor expansion and the fact that the breaks are local in nature, we have that

$$\begin{aligned}
\sqrt{T}\left(f(\hat{\boldsymbol{\beta}}_2, \hat{\boldsymbol{\delta}}) - f(\boldsymbol{\beta}_2, \boldsymbol{\delta})\right) &= \sqrt{T}\left[f'_{\beta_2}(\hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2) + f'_\delta(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}) + O(T^{-1})\right] \\
&\xrightarrow{d} N\left(0, f'_{\beta_2}\text{Var}(\hat{\boldsymbol{\beta}}_2)f_{\beta_2} + q\right) \\
\sqrt{T}\left(f(\hat{\boldsymbol{\beta}}_F, \hat{\boldsymbol{\delta}}) - f(\boldsymbol{\beta}_2), \boldsymbol{\delta})\right) &= \sqrt{T}\left[f'_{\beta_2}(\hat{\boldsymbol{\beta}}_F - \boldsymbol{\beta}_2) + f'_\delta(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}) + O(T^{-1})\right] \\
&\xrightarrow{d} N\left(\tau_b f'_{\beta_2}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2), f'_{\beta_2}\text{Var}(\hat{\boldsymbol{\beta}}_F)f_{\beta_2} + q\right)
\end{aligned}$$

where $q = \text{plim}_{T\to\infty} T \cdot \left[f_\delta' \text{Var}(\hat{\boldsymbol{\delta}}) f_\delta + 2 f_{\beta_2}' \text{Cov}(\hat{\boldsymbol{\beta}}_F, \hat{\boldsymbol{\delta}}) f_\delta \right]$ and we use that $\text{plim}_{T\to\infty} T \cdot \text{Cov}(\hat{\boldsymbol{\beta}}_F, \hat{\boldsymbol{\delta}}) = \text{plim}_{T\to\infty} T \cdot \text{Cov}(\hat{\boldsymbol{\beta}}_2, \hat{\boldsymbol{\delta}})$. Using previous results on the covariance matrix of the estimators, and the notation in (3.12), we have

$$
\begin{aligned}
T \cdot f_{\beta_2}' \text{Var}(\hat{\boldsymbol{\beta}}_2) f_{\beta_2} &\xrightarrow{p} \frac{1}{1-\tau_b} f_{\beta_2}' \boldsymbol{V} f_{\beta_2} + f_{\beta_2}' \tilde{\boldsymbol{Q}} f_{\beta_2} \\
T \cdot f_{\beta_2}' \text{Var}(\hat{\boldsymbol{\beta}}_F) f_{\beta_2} &\xrightarrow{p} f_{\beta_2}' \boldsymbol{V} f_{\beta_2} + f_{\beta_2}' \tilde{\boldsymbol{Q}} f_{\beta_2}
\end{aligned}
$$

For the expected MSFEs using $\boldsymbol{\beta}_2$ and $\boldsymbol{\beta}_F$, we have

$$
\begin{aligned}
T\text{E}\left[\left(f(\hat{\boldsymbol{\beta}}_2, \hat{\boldsymbol{\delta}}) - f(\boldsymbol{\beta}_2, \boldsymbol{\delta}) \right)^2 \right] &\xrightarrow{p} \frac{1}{1-\tau_b} f_{\beta_2}' \boldsymbol{V} f_{\beta_2} + f_{\beta_2}' \tilde{\boldsymbol{Q}} f_{\beta_2} + q \\
T\text{E}\left[\left(f(\hat{\boldsymbol{\beta}}_F, \hat{\boldsymbol{\delta}}) - f(\boldsymbol{\beta}_2, \boldsymbol{\delta}) \right)^2 \right] &\xrightarrow{p} \left[\tau_b f_{\beta_2}' (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2) \right]^2 + f_{\beta_2}' \boldsymbol{V} f_{\beta_2} + f_{\beta_2}' \tilde{\boldsymbol{Q}} f_{\beta_2} + q
\end{aligned}
$$

Hence, the full sample based forecast improves over the post-break sample based forecast if

$$
\zeta = T(1-\tau_b)\tau_b \frac{\left[f_{\beta_2}' (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2) \right]^2}{f_{\beta_2}' \boldsymbol{V} f_{\beta_2}} \leqslant 1 \tag{3.16}
$$

Similar to Section 3.2, a test for $H_0 : \zeta = 1$ can be derived by by noting that $T \cdot \text{Var}(\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_2) = T \cdot \left[\text{Var}(\hat{\boldsymbol{\beta}}_1) + \text{Var}(\hat{\boldsymbol{\beta}}_2) - 2\text{Cov}(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2) \right] \xrightarrow{p} \frac{1}{\tau_b(1-\tau_b)} \boldsymbol{V}$ and, therefore,

$$
\hat{\zeta} = T(1-\tau_b)\tau_b \frac{\left[f_{\beta_2}' (\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_2) \right]^2}{\hat{\omega}} \xrightarrow{d} \chi^2(1, \zeta) \tag{3.17}
$$

where $\text{plim}_{T\to\infty} \hat{\omega} = f_{\beta_2}' \boldsymbol{V} f_{\beta_2}$. The test statistic, $\hat{\zeta}$, can be compared against the critical values of the $\chi^2(1,1)$ distribution to test for equal forecast performance.

The above can be immediately applied to the simple structural break model (3.1) where $f(\hat{\boldsymbol{\beta}}_2; \boldsymbol{x}_{T+1}) = \boldsymbol{x}_{T+1}' \hat{\boldsymbol{\beta}}_2$, and $f_{\beta_2} = \boldsymbol{x}_{T+1}$. The full sample forecast is more accurate if

$$
\zeta = T\tau_b(1-\tau_b) \frac{\left[\boldsymbol{x}_{T+1}' (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2) \right]^2}{\boldsymbol{x}_{T+1}' \boldsymbol{V} \boldsymbol{x}_{T+1}} \leqslant 1 \tag{3.18}
$$

identical to the result in (3.5).

3.4.2 A local break of unknown timing

If the timing of the break is unknown and $\tau < \tau_b$, then the estimator of $\boldsymbol{\beta}_2$ is biased as can be seen from the last term in (3.13). The difference between the expected asymptotic MSFE of the partial sample forecast and that of the full sample forecast, standardized by the variance

of the partial sample forecast, is

$$\begin{aligned}\Delta &= \left\{R(\hat{\boldsymbol{\beta}}_2(\hat{\tau}),\hat{\boldsymbol{\delta}}) - R(\hat{\boldsymbol{\beta}}_F,\hat{\boldsymbol{\delta}})\right\}/f'_{\beta_2}\boldsymbol{V}f_{\beta_2}\\ &= T\left\{\mathrm{E}\left[(f(\hat{\boldsymbol{\beta}}_2(\hat{\tau}),\hat{\boldsymbol{\delta}}) - f(\boldsymbol{\beta}_2,\boldsymbol{\delta}))^2\right] - \mathrm{E}\left[(f(\hat{\boldsymbol{\beta}}_F,\hat{\boldsymbol{\delta}}) - f(\boldsymbol{\beta}_2,\boldsymbol{\delta}))^2\right]\right\}/f'_{\beta_2}\boldsymbol{V}f_{\beta_2}\\ &= T\left\{\mathrm{E}\left[\left(f'_{\beta_2}(\hat{\boldsymbol{\beta}}_2(\hat{\tau}) - \boldsymbol{\beta}_2)\right)^2\right] - \mathrm{E}\left[f'_{\beta_2}(\hat{\boldsymbol{\beta}}_F - \boldsymbol{\beta}_2)\right]^2 - f'_{\beta_2}\mathrm{Var}(\hat{\boldsymbol{\beta}}_F)f_{\beta_2}\right\}/f'_{\beta_2}\boldsymbol{V}f_{\beta_2}\end{aligned}$$

where $R(\hat{\boldsymbol{\theta}})$ is the asymptotic MSFE under parameter estimates $\hat{\boldsymbol{\theta}}$. The derivations are provided in Appendix 3.A.1. Using (3.13) and (3.15) we obtain

$$\begin{aligned}\Delta &= \frac{R(\hat{\boldsymbol{\beta}}_2(\hat{\tau}),\hat{\boldsymbol{\delta}}) - R(\hat{\boldsymbol{\beta}}_F,\hat{\boldsymbol{\delta}})}{f'_{\beta_2}\boldsymbol{V}f_{\beta_2}}\\ &= \mathrm{E}\left\{\left[\frac{1}{1-\hat{\tau}}\frac{f'_{\beta_2}\boldsymbol{V}\bar{\boldsymbol{X}}'(\boldsymbol{B}(1)-\boldsymbol{B}(\hat{\tau}))}{\sqrt{f'_{\beta_2}\boldsymbol{V}f_{\beta_2}}} + \frac{1}{1-\hat{\tau}}\int_{\hat{\tau}}^1 \frac{f'_{\beta_2}\boldsymbol{\eta}(s)}{\sqrt{f'_{\beta_2}\boldsymbol{V}f_{\beta_2}}}ds\right]^2\right\}\\ &\quad - \left(\int_0^1 \frac{f'_{\beta_2}\boldsymbol{\eta}(s)}{\sqrt{f'_{\beta_2}\boldsymbol{V}f_{\beta_2}}}ds\right)^2 - 1 \qquad (3.19)\end{aligned}$$

Note that (3.19) makes no assumption about the form of the instability, which is governed by $\boldsymbol{\eta}(\tau)$. Define $J(\tau) = \int_\tau^1 \frac{f'_{\beta_2}\boldsymbol{\eta}(s)}{\sqrt{f'_{\beta_2}\boldsymbol{V}f_{\beta_2}}}ds$ and note that for fixed f'_{β_2} the continuous mapping theorem yields $\frac{f'_{\beta_2}\boldsymbol{V}\bar{\boldsymbol{X}}'[\boldsymbol{B}(1)-\boldsymbol{B}(\tau)]}{\sqrt{f'_{\beta_2}\boldsymbol{V}f_{\beta_2}}} = B(1)-B(\tau)$, where $B(\cdot)$ is a one-dimensional Brownian motion. Then

$$\Delta = \mathrm{E}_{f(\tau)}\left\{\left[\frac{1}{1-\tau}(B(1)-B(\tau)) + \frac{1}{1-\tau}J(\tau)\right]^2\right\} - J(1)^2 - 1$$

which could be used to test whether the use of a moving window will outperform an expanding window under various forms of parameter instability. The expectation simplifies if the size of the moving window is exogenously set to some fraction of the total number of observations.

Under a structural break, $\boldsymbol{\eta}(\tau) = \boldsymbol{b}I[\tau < \tau_b]$

$$\Delta = \mathrm{E}_{f(\hat{\tau})}\left\{\left[\frac{1}{1-\hat{\tau}}\left(B(1)-B(\hat{\tau})\right) + \theta_{\tau_b}\frac{\tau_b-\hat{\tau}}{1-\hat{\tau}}I[\hat{\tau}<\tau_b]\right]^2\right\} - \theta_{\tau_b}^2\tau_b^2 - 1 \qquad (3.20)$$

where $\theta_{\tau_b} = \frac{f'_{\beta_2}\boldsymbol{b}}{\sqrt{f'_{\beta_2}\boldsymbol{V}f_{\beta_2}}}$. The subscript τ_b is added for notational purposes in the following section. This is related to the standardized break size ζ from the previous section by

$$\theta_{\tau_b} = \sqrt{\frac{\zeta}{\tau_b(1-\tau_b)}} \tag{3.21}$$

If $\hat{\tau} = \tau_b$, the critical break size of the previous section is obtained. However, under an unknown break date, in general, $\hat{\tau} \neq \tau_b$ and (3.20) cannot immediately be used for testing purposes.

It is, however, interesting to observe that since Δ is symmetric around $\theta_{\tau_b} = 0$, $\Delta > 0$ for $\theta_{\tau_b} = 0$, and (3.20) quadratically decreases away from $\theta_{\tau_b} = 0$, there is a break size $|\theta_{\tau_b}|$ for each τ_b for which $\Delta = 0$. Numerical results depicted in Figure 3.11 clearly show that equal predictive accuracy is attained for a unique break size. This makes it an excellent candidate test statistic. Analogous to the case where the break date is known, we simply use the Wald test statistic

$$W(\tau) = T\frac{\left[f'_{\beta_2}(\hat{\boldsymbol{\beta}}_2(\tau) - \hat{\boldsymbol{\beta}}_1(\tau))\right]^2}{f'_{\beta_2}\left(\frac{\hat{\boldsymbol{V}}_1}{\tau} - \frac{\hat{\boldsymbol{V}}_2}{1-\tau}\right)f_{\beta_2}} \tag{3.22}$$

We will show below that for sufficiently small size the test statistic in (3.22) identifies the true break date up to a constant that vanishes with decreasing size, which establishes a weak form of optimality of the sup-Wald test even when the break size for which $\Delta = 0$ is non-constant over τ_b.

Since the function f'_{β_2} is fixed, the results in Andrews (1993) and the continuous mapping theorem show that under local alternatives $W(\tau)$ in (3.22) converges to

$$\begin{aligned} Q^*(\tau) &= \left(\frac{B(\tau) - \tau B(1)}{\sqrt{\tau(1-\tau)}} + \sqrt{\frac{1-\tau}{\tau}}\int_0^\tau \eta(s)ds - \sqrt{\frac{\tau}{1-\tau}}\int_\tau^1 \eta(s)ds\right)^2 \\ &= \left(\frac{B(\tau) - \tau B(1)}{\sqrt{\tau(1-\tau)}} + \mu(\tau;\theta_{\tau_b})\right)^2 \end{aligned} \tag{3.23}$$

The first term of (3.23) is a self-normalized Brownian bridge with expectation zero and variance equal to one. For a fixed break date, $Q^*(\tau)$ follows a non-central χ^2 distribution with one degree of freedom and non-centrality parameter $\mu(\tau;\theta_{\tau_b})^2$. For the structural break model, we have

$$\mu(\tau;\theta_{\tau_b}) = \theta_{\tau_b}\left[\sqrt{\frac{1-\tau}{\tau}}\tau_b I[\tau_b < \tau] + \sqrt{\frac{\tau}{1-\tau}}(1-\tau_b)I[\tau_b \geqslant \tau]\right] \tag{3.24}$$

For the optimality results below the following assumption is made with regard to the non-centrality parameter

Assumption 2 *The function $\mu(\tau;\theta_{\tau_b})$ is maximized (or minimized) if and only if $\tau = \tau_b$*

Under this assumption, for a sufficiently small nominal size, rejections are found only for break locations that are close to τ_b. For the structural break model it is easy to verify that Assumption 2 holds. The extremum value is given by $\mu(\tau_b;\theta_{\tau_b}) = \theta_{\tau_b}\sqrt{\tau_b(1-\tau_b)} = \zeta^{1/2}$.

3.4.3 Weak optimality

In this section, we will show that the Wald test is weakly optimal when the null hypothesis (and consequently the critical values) depend on the unknown break date. Using arguments of Piterbarg (1996), we show that under a general form of instability, only points in a small neighborhood around the maximum instability point τ_b contribute to the probability of exceeding a constant boundary, u, in the limit where the size of the test tends to zero. In a second step we extend the analysis by considering a null hypothesis that depends on an unknown and weakly identified parameter τ_b. In this case, critical values will also depend τ_b. If the critical values vary sufficiently slow with τ_b, then using an estimate $\hat{\tau}$ leads to a weakly optimal test in the sense that it has larger or equal power compared to the test that knows τ_b in the limit where the size of the test goes to zero. In Section 3.5 we provide evidence that the power of the test under estimated τ_b is close to that of the infeasible test under known τ_b for values of u corresponding to standard size values.

Location concentration

To prove that only points in a small neighborhood of the true break date contribute to the probability of exceeding a distant boundary, we require the following preliminaries.

Lemma 1 *Suppose $Z(\tau)$ is a symmetric Gaussian process, i.e. $P(Z(\tau) > u) = P(-Z(\tau) > u)$, then*

$$P\left(\sup_{\tau\times c}[Z(\tau) + \mu(\tau;\theta_{\tau_b})]c > u\right) = P\left(\sup_{\tau\in I} Z(\tau) > u - |\mu(\tau;\theta_{\tau_b})|\right)[1 + o(1)]$$

where $c = \pm 1$ and the supremum is taken jointly over $\tau \in I = [\tau_{\min}, \tau_{\max}]$ and c.

Proof: Consider first $\mu(\tau;\theta_{\tau_b}) > 0$ then

$$\begin{aligned} P(Z(\tau) + \mu(\tau;\theta_{\tau_b}) > u, \tau \in I) &= P(Z(\tau) > u - |\mu(\tau;\theta_{\tau_b})|, \tau \in I) \\ P(-Z(\tau) - \mu(\tau;\theta_{\tau_b}) > u, \tau \in I) &= P(Z(\tau) > u + |\mu(\tau;\theta_{\tau_b})|, \tau \in I) \end{aligned} \tag{3.25}$$

where $\tau \in I$ is shorthand notation for "*for some* $\tau \in I$". When $\mu(\tau;\theta_{\tau_b}) < 0$ we have

$$\begin{aligned} P(-Z(\tau) - \mu(\tau;\theta_{\tau_b}) > u, \tau \in I) &= P(Z(\tau) > u - |\mu(\tau;\theta_{\tau_b})|, \tau \in I) \\ P(Z(\tau) + \mu(\tau;\theta_{\tau_b}) > u, \tau \in I) &= P(Z(\tau) > u + |\mu(\tau;\theta_{\tau_b})|, \tau \in I) \end{aligned} \tag{3.26}$$

The bounds in the second lines of (3.25) and (3.26) are equal or larger then the bounds in the first lines. It follows from the results below that the crossing probabilities over the larger bounds are negligible compared to the crossing probabilities over the lower bounds. This implies that for any sign of $\mu(\tau;\theta_{\tau_b})$ as $u \to \infty$

$$P\left(\sup_{\tau\times c}[Z(\tau)+\mu(\tau;\theta_{\tau_b})]c > u\right) = P\left(\sup_{\tau\in I} Z(\tau) > u - |\mu(\tau;\theta_{\tau_b})|\right)[1+o(1)] \quad (3.27)$$

as required. ■

In the structural break model, $Z(\tau)$ is a locally stationary Gaussian process with correlation function $r(\tau,\tau+s)$, defined as follows

Definition 1 (Local stationarity) *A Gaussian process is locally stationary if there exists a continuous function $C(\tau)$ satisfying $0 < C(\tau) < \infty$*

$$\lim_{s\to 0}\frac{1-r(\tau,\tau+s)}{|s|^{\alpha}} = C(\tau) \textit{ uniformly in } \tau \geqslant 0$$

See Hüsler (1990). The correlation function can be written as

$$r(\tau,\tau+s) = 1 - C(\tau)|s|^{\alpha} \text{ as } s \to 0$$

The standardized Brownian bridge that we encounter in the structural break model is a locally stationary process with $\alpha = 1$ and local covariance function $C(\tau) = \frac{1}{2}\frac{1}{\tau(1-\tau)}$. Since $\tau \in [\tau_{\min},\tau_{\max}]$ with $0 < \tau_{\min} < \tau_{\max} < 1$, it holds that $0 < C(\tau) < \infty$.

Lemma 2 *Suppose $Z(\tau)$ is a locally stationary process with local covariance function $C(\tau)$ then if $\delta(u)u^2 \to \infty$ and $\delta(u) \to 0$ as $u \to \infty$*

$$P\left(\sup_{[\tau,\tau+\delta(u)]} Z(t) > u\right) = \frac{1}{\sqrt{2\pi}}\delta(u)u\exp\left(-\frac{1}{2}u^2\right)C(\tau) \quad (3.28)$$

Proof: see Hüsler (1990).

Given the above results, we can state the following

Theorem 1 (Location concentration) *Suppose $Q^*(\tau) = [Z(\tau)+\mu(\tau;\theta_{\tau_b})]^2$ where $Z(\tau)$ is a zero mean Gaussian process with variance equal to one and $|\mu(\tau;\theta_{\tau_b})|$ is a function that attains its unique maximum when $\tau = \tau_b$, then as $u \to \infty$*

$$P\left(\sup_{\tau\in I} Q^*(\tau) > u^2\right) = P\left(Z(\tau) > u - |\mu(\tau;\theta_{\tau_b})| \textit{ for some } \tau \in I_1\right)(1+o(1))$$

where $I = [\tau_{\min},\tau_{\max}]$, $I_1 = [\tau_b - \delta(u), \tau_b + \delta(u)]$ and $\delta(u) = u^{-1}\log^2 u$.

Proof: We start by noting that for $\tau \in I = [\tau_{\min}, \tau_{\max}]$

$$\begin{aligned} P\left(\sup_{\tau\in I} Q^*(\tau) > u^2\right) &= P\left(\sup_{\tau\in I} \sqrt{Q^*(\tau)} > u\right) \\ &= P\left(\sup_{\tau\in I} |Z(\tau) + \mu(\tau;\theta_{\tau_b})| > u\right) \\ &= P\left(\sup_{\tau\times c}[Z(\tau) + \mu(\tau;\theta_{\tau_b})]c > u\right) \qquad \text{with } c = \pm 1 \\ &= P\left(\sup_{\tau\in I} Z(\tau) > u - |\mu(\tau;\theta_{\tau_b})|\right)[1 + o(1)] \end{aligned}$$

where the supremum is taken jointly over $\tau \in I$ and c. The last equality follows from Lemma 1.

Now we proceed along the lines of Piterbarg (1996). As in Lemma 2, consider a region close to τ_b defined by $I_1 = [\tau_b - \delta(u), \tau_b + \delta(u)]$. In I_1, the minimum value of the boundary is given by

$$\underline{b} = \inf_{\tau\in I_1}[u - \mu(\tau;\theta_{\tau_b})] = u - |\mu(\tau_b;\theta_{\tau_b})| \tag{3.29}$$

so that

$$\begin{aligned} P_{I_1} &= P(Z(\tau) > u - |\mu(\tau;\theta_{\tau_b})| \text{ for some } \tau \in I_1) \\ &\leqslant P(Z(\tau) > \underline{b} \text{ for some } \tau \in I_1) \\ &= 2\delta(u)\underline{b}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\underline{b}^2\right)C(\tau_b) \\ &= \frac{2\delta(u)}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\underline{b}^2 + \log\underline{b}\right)C(\tau_b) \end{aligned}$$

where the third line follows from (3.28).

Define the region outside of I_1 as $I_A = I\backslash I_1$. Then in I_A, the minimum value of the boundary is given by

$$\underline{b}_A = u - |\mu(\tau_b + \delta(u);\theta_{\tau_b})| \tag{3.30}$$

Taking a Taylor expansion of $\mu(\tau_b + \delta(u);\theta_{\tau_b})$ around $\delta(u) = 0$ gives

$$\mu(\tau_b + \delta(u);\theta_{\tau_b}) = \mu(\tau_b;\theta_{\tau_b}) + \gamma\delta(u) + O\left[\delta(u)^2\right] \tag{3.31}$$

where $\gamma = \left.\frac{\partial\mu(\tau;\theta_{\tau_b})}{\partial\tau}\right|_{\tau=\tau_b}$. Then $\underline{b}_A = \underline{b} + \gamma\delta(u)$ and

$$\begin{aligned} P_{I_A} &= P(Z(\tau) > u - |\mu(\tau;\theta_{\tau_b})| \text{ for some } \tau \in I_A) \\ &\leqslant P(Z(\tau) > \underline{b}_A \text{ for some } \tau \in I_A) \\ &\leqslant \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\underline{b}^2 - \underline{b}\gamma\delta(u) - \frac{1}{2}\gamma^2\delta(u)^2 + \log(\underline{b} + \gamma\delta(u))\right)\overline{C} \end{aligned} \tag{3.32}$$

where the last line defines $\overline{C}$ by noting that

$$\sum_{I_k \in I_A} C(k\delta(u))\delta(u) \overset{\delta(u)\to 0}{=} \int_{I_A} C(\tau)d\tau \leqslant \int_I C(\tau)d\tau = \overline{C} < \infty \tag{3.33}$$

with I_k representing non-overlapping intervals of width $\delta(u)$ such that $\bigcup_{k=2}^{\infty} I_k = I_A$ and $k\delta(u) \in I_k$

Compare (3.32) to the probability of a test with a known break date to exceed the critical value

$$P_0 = P(Z(\tau_b) > u - |\mu(\tau_b; \theta_{\tau_b})|) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\underline{b}^2 - \log(\underline{b})\right) \tag{3.34}$$

where we use that

$$\frac{1}{\sqrt{2\pi}} \int_u^{\infty} \exp\left(-\frac{1}{2}x^2\right) dx = \frac{1}{\sqrt{2\pi}u} \exp\left(-\frac{1}{2}u^2\right) \text{ as } u \to \infty$$

Ignoring the lower order terms $-\frac{1}{2}\gamma^2\delta(u)^2 + \log(\underline{b} + \gamma\delta(u))$, equation (3.32) contains an extra term $\exp(-\underline{b}\gamma\delta(u))$ compared to (3.34). Using (3.29), this implies that $P_{I_A} = o(P_0)$ if

$$\frac{u\delta(u)}{\log u} \to \infty$$

Subsequently, if

$$\delta(u) = u^{-1}\log^2(u) \tag{3.35}$$

then all intervals outside of I_1 contribute $o(P_0)$ to the probability of crossing the boundary u. Under (3.35), we have that for P_{I_1} as $u \to \infty$

$$\begin{aligned} P_{I_1} \leqslant P_I &\leqslant P_{I_1} + P_{I_A} \\ &\leqslant P_{I_1} + o(P_0) \end{aligned}$$

We now only need to note that

$$\begin{aligned} P_{I_1} &= P(Z(\tau) > u - |\mu(\tau; \theta_{\tau_b})| \text{ for some } \tau \in I_1) \\ &\geqslant P(Z(\tau_b) > u - |\mu(\tau_b; \theta_{\tau_b})|) = P_0 \end{aligned}$$

To conclude that

$$P\left(\sup_{\tau \in I} Z(\tau) > u - |\mu(\tau; \theta_{\tau_b})|, \tau \in I\right) = P_{I_1}(1 + o(1)) \text{ as } u \to \infty$$

which completes the proof. ■

Note that, in (3.32), the term $\exp(\underline{b}\delta(u))^{-\gamma}$ ensures that $P_{I_A} = o(P_1)$. In the structural break model, we see that (3.31) is given by $\mu(\tau_b + \delta(u); \theta_{\tau_b}) = \theta_{\tau_b}\sqrt{\tau_b(1-\tau_b)} - \frac{1}{2}\theta_{\tau_b}\frac{1}{\sqrt{\tau_b(1-\tau_b)}}\delta(u) + O\left[\delta(u)^2\right]$. It is clear that γ scales linearly with the break size. Therefore, for a sufficiently large break, asymptotic optimality results are expected to extend to the practical case when u is finite. The simulations of asymptotic power presented in Section 3.5 confirm this.

Corollary 1 (Corollary 8.1 of Piterbarg (1996)) *As $u \to \infty$, the distribution of the break location denoted by D converges converges to a delta function located at $\tau = \tau_b$ for excesses over the boundary u^2, i.e.*

$$D\left(\hat{\tau} : Q^*(\hat{\tau}) = \sup_{\tau \in I} Q^*(\tau) \middle| \sup_{\tau \in I} Q^*(\tau) > u^2\right) \to \delta_{\tau_b} \text{ as } u \to \infty$$

This corollary implies that post-test parameter inference after a rejection of the null is in fact standard.

Weak optimality under location dependent boundaries

While the location concentration is essential to our proof, the problem we consider is further complicated by the fact that the size under the null hypothesis depends on the unknown break date. This translates into critical values that will also depend on the unknown break date. Theorem 1 indicates that simply plugging in the estimate of the break date that maximizes the Wald statistic could be a viable strategy to obtain a well behaved test. In fact, this strategy is weakly optimal in the sense of Andrews (1993) if the following assumption is satisfied.

Assumption 3 (Slowly varying critical values) *Suppose we test using a sequence of critical values that control size for every τ_b, i.e. $u = u(\tau_b)$ such that*

$$P\left(\sup_{\tau} Q^*(\tau) > u(\tau^*) \middle| \tau_b = \tau^*\right) = \alpha$$

then $u(\tau) - \mu(\tau; \theta_{\tau_b})$ should have a unique minimum on $I_1 = [\tau_b - \delta(u), \tau_b + \delta(u)]$ at $\tau = \tau_b$.

Suppose that $u(\tau_b)$ a differentiable function with respect to τ_b, then a sufficient condition is that the critical values are slowly varying with τ_b in comparison with the derivative of the function $\mu(\tau; \theta_{\tau_b})$ with respect to τ on the interval I_1, i.e.

$$\left|\frac{\partial u(\tau_b)}{\partial \tau_b}\right| < \left|\frac{\partial \mu(\tau; \theta_{\tau_b})}{\partial \tau}\right|$$

In the structural break model, the derivative on the right hand side occurs in (3.31) as $\gamma = \frac{\partial \mu(\tau; \theta_{\tau_b})}{\partial \tau} = \theta_{\tau_b}\frac{1}{\sqrt{\tau_b(1-\tau_b)}}$. The slowly varying condition relates the dependence of the critical

values on τ_b to the identification strength of the break date as the derivative of $\mu(\tau;\theta_{\tau_b})$ with respect to τ scales linearly with the break size. For the break size we know from Section 3.2 that $\theta_{\tau_b}\sqrt{\tau_b(1-\tau_b)} \geqslant 1$, where the equality holds if the break date is known with certainty. Then

$$\gamma = \frac{\theta_{\tau_b}}{\sqrt{\tau_b(1-\tau_b)}} \geqslant \frac{1}{\tau_b(1-\tau_b)}$$

A sufficient condition for the slowly varying assumption is therefore

$$\left|\frac{\partial u(\tau_b)}{\partial \tau_b}\right| \leqslant \frac{1}{\tau_b(1-\tau_b)} \tag{3.36}$$

which will show to hold once critical values have been obtained.

We now provide a result on the optimality of the test under Assumption 3. Let $v(\tau_b)$ denote the critical values of the optimal test conditional on the break location, i.e. $P(Q^*(\tau_b) > v(\tau_b)^2) = \alpha$, where $v(\tau_b)$ is the critical value of the test conditional on the true break date, τ_b.

Lemma 3 (Convergence of critical values) *Let $u(\tau)$ be the critical value that controls size for a given τ and let $v(\tau_b)$ be the critical value of the test on the true break date, then $u(\tau_b) - v(\tau_b) \to 0$.*

Proof: By definition of the critical values

$$P\left[\sup_\tau Q^*(\tau) > u(\tau_b)^2\right] = P\left[Z(\tau) > u(\tau_b) - |\mu(\tau;\theta_{\tau_b})| \text{ for some } \tau \in I_1\right] = \alpha$$
$$P\left[Q^*(\tau_b) > v(\tau_b)^2\right] = P\left[Z(\tau_b) > v(\tau_b) - |\mu(\tau_b;\theta_{\tau_b})|\right] = \alpha$$

Since τ in the first line is contained in I_1, we have by a Taylor series expansion of $\mu(\tau;\theta_{\tau_b})$ around τ_b that $\max|\mu(\tau;\theta_{\tau_b})| - |\mu(\tau_b;\theta_{\tau_b})| = O[\delta(u)]$ and consequently, $\max u(\tau_b) - v(\tau_b) = O(\delta(u))$. Since $\delta(u) \to 0$ as $u \to \infty$, the difference in the critical values $u(\tau_b) - v(\tau_b) \to 0$ as $u \to \infty$. ■

Theorem 2 (Weak optimality) *Under a slowly varying boundary*

$$\begin{aligned} &P_{H_a}\left[\sup_\tau Q^*(\tau) > u(\hat{\tau})^2\right] - P_{H_a}(Q^*\left[\tau_b) > v(\tau_b)^2\right] \\ &\geqslant P_{H_a}\left[Q^*(\tau_b) > u(\tau_b)\right] - P_{H_a}(Q^*\left[\tau_b) > v(\tau_b)^2\right] \\ &= 0 \end{aligned} \tag{3.37}$$

where $\hat{\tau} = \arg\sup_\tau Q^*(\tau)$ and P_{H_a} denotes the crossing probability under the alternative.

Proof: As before

$$P_{H_a}\left[\sup_\tau Q^*(\tau) > u(\hat{\tau})^2\right] = P_{H_a}\left[Z(\hat{\tau}) > u(\hat{\tau}) - \mu(\hat{\tau};\theta_{\tau_b})\right]$$

Under the slowly varying assumption, $u(\hat{\tau}) - \mu(\hat{\tau}; \theta_{\tau_b})$ has a unique minimum on I_1 at $\hat{\tau} = \tau_b$. Taking the supremum therefore necessarily leads to at least as many exceedances as considering $\tau = \tau_b$ alone, which proves the inequality in (3.37). The last line of (3.37) is follows from Lemma 3. ■

Testing with critical values independent of the break date

A test based on the Wald statistic (3.22) requires critical values that dependent on the estimated break date. It is, however, straightforward to derive a test statistic with critical values that are independent of the break date in the limit where $u \to \infty$.

Corollary 2 *A test statistic with critical values that are independent of τ_b for $u \to \infty$ is given by*

$$S(\hat{\tau}) = \sup_{\tau \in I} \sqrt{T} \frac{\left| f'_{\beta_2} \left(\hat{\boldsymbol{\beta}}_2(\tau) - \hat{\boldsymbol{\beta}}_1(\tau) \right) \right|}{\sqrt{f'_{\beta_2} \left(\frac{\hat{\boldsymbol{V}}_1}{\tau} + \frac{\hat{\boldsymbol{V}}_2}{1-\tau} \right) f_{\beta_2}}} - |\mu(\hat{\tau}; \theta_{\hat{\tau}})| \tag{3.38}$$

where $\hat{\tau}$ maximizes the first term of S or, equivalently, the Wald statistic (3.22)*.*

Proof: The test statistic converges to $S(\hat{\tau}) \to \sup_\tau |Z(\tau) + \mu(\tau; \theta_{\tau_b})| - |\mu(\hat{\tau}; \theta_{\hat{\tau}})|$ where $\hat{\tau}$ maximizes the first term. As shown before, exceedances of a high boundary are concentrated in the region $[\tau_b - \delta(u), \tau_b + \delta(u)]$ where $\delta(u) \to 0$ as $u \to \infty$. Then

$$\begin{aligned} P(S(\hat{\tau}) > u) &= P\left(\sup_{I_1} |Z(\tau) + \mu(\tau; \theta_{\tau_b})| - |\mu(\hat{\tau}; \theta_{\hat{\tau}})| > u \right) \\ &= P(Z(\hat{\tau}) > u - |\mu(\hat{\tau}; \theta_{\tau_b})| + |\mu(\hat{\tau}; \theta_{\hat{\tau}})|) \end{aligned}$$

Under the slowly varying assumption, the difference $-|\mu(\hat{\tau}; \theta_{\tau_b})| + |\mu(\hat{\tau}; \theta_{\hat{\tau}})| = O[\delta(u)]$. This implies that the critical values of $S(\hat{\tau})$ are independent of τ_b in the limit where $u \to \infty$. ■

3.4.4 Optimal weights or shrinkage forecasts

Pesaran et al. (2013) derive optimal weights for observations in the estimation sample such that, in the presence of a structural break, the MSFE of the one step ahead forecast is minimized. These optimal weights are derived under the assumption of a known break date and break size. Conditional on the break date, the optimal weights take one value for observations in the pre-break regime and one value for observations in the post-break regime. This implies that the forecast can be written as a weighted average of pre-break and post-break parameter estimates.

We can therefore write the optimally weighted forecast as a convex combination of the forecasts from pre-break observations and post-break observations

$$\hat{y}_{T+1}^{S}(\tau) = \omega \boldsymbol{x}_{T+1}' \hat{\boldsymbol{\beta}}_1 + (1-\omega)\boldsymbol{x}_{T+1}' \hat{\boldsymbol{\beta}}_2$$

where the optimal forecast is denoted with subscript S as we will see below that it is equal to a shrinkage forecast that shrinks the post-break sample based forecast in the direction of the full sample based forecast.

The expected mean squared error is

$$\begin{aligned} \mathrm{E}\left[T\left(\hat{y}_{T+1}^{S} - \boldsymbol{x}_{T+1}'\boldsymbol{\beta}_2\right)^2\right] &= \mathrm{E}\left[T\left(\omega \boldsymbol{x}_{T+1}' \hat{\boldsymbol{\beta}}_1 + (1-\omega)\boldsymbol{x}_{T+1}' \hat{\boldsymbol{\beta}}_2 - \boldsymbol{x}_{T+1}'\boldsymbol{\beta}_2\right)^2\right] \\ &= \omega^2 T\left[\boldsymbol{x}_{T+1}'\left(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\right)\right]^2 + \\ &\quad + \omega^2 \boldsymbol{x}_{T+1}'\left(\frac{1}{\tau_b} + \frac{1}{1-\tau_b}\right)\boldsymbol{V}\boldsymbol{x}_{T+1} \\ &\quad - 2\omega\frac{1}{1-\tau_b}\boldsymbol{x}_{T+1}'\boldsymbol{V}\boldsymbol{x}_{T+1} + \frac{1}{\tau_b}\boldsymbol{x}_{T+1}'\boldsymbol{V}\boldsymbol{x}_{T+1} \end{aligned} \tag{3.39}$$

see Appendix 3.A.4 for details.

Maximizing (3.39) with respect to ω yields

$$\omega^* = \tau_b\left[1 + T\frac{\left(\boldsymbol{x}_{T+1}'(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)\right)^2}{\boldsymbol{x}_{T+1}'\left(\frac{1}{\tau_b} + \frac{1}{1-\tau_b}\right)\boldsymbol{V}\boldsymbol{x}_{T+1}}\right]^{-1} \tag{3.40}$$

where the denominator contains the Wald statistic derived, W, in (3.5) and (3.18).

Alternatively, we can combine the full sample forecast and the post-break sample forecast. Since $\hat{\boldsymbol{\beta}}_F = \tau_b \hat{\boldsymbol{\beta}}_1 + (1-\tau_b)\hat{\boldsymbol{\beta}}_2$,

$$\begin{aligned} \hat{y}_{T+1}^{S} &= \omega \boldsymbol{x}_{T+1}' \hat{\boldsymbol{\beta}}_1 + (1-\omega)\boldsymbol{x}_{T+1}' \hat{\boldsymbol{\beta}}_2 \\ &= \frac{\omega}{\tau_b}\boldsymbol{x}_{T+1}' \hat{\boldsymbol{\beta}}_F + \left(1 - \frac{\omega}{\tau_b}\right)\hat{\boldsymbol{\beta}}_2 \end{aligned}$$

and the optimal weight on the full sample forecast is

$$\omega_F^* = \frac{\omega^*}{\tau_b} = \frac{1}{1 + W(\tau_b)} \tag{3.41}$$

The shrinkage estimator is therefore a convex combination of the full sample and post-break sample forecast with weights that are determined by the Wald test statistic.

The empirical results in Pesaran et al. (2013) suggest that uncertainty around the break date substantially deteriorates the accuracy of the optimal weights forecast in applications. As a consequence, Pesaran et al. (2013) derive robust optimal weight by integrating over the

break dates, which yield substantially more accurate forecasts. Given the impact that break date uncertainty has on choosing between the post-break and the full sample forecasts, it is not surprising that the same uncertainty should impact the weights. If this uncertainty is not taken into account, the weight on the post-break forecast will be too high. However, the lack of analytic expressions for the break date uncertainty complicates an analytic weighting scheme. Alternatively, this uncertainty can be taken into account by testing whether the break date uncertainty is small enough to justify using the shrinkage forecast.

As the Wald statistic in (3.41) requires the true break date, consider the shrinkage forecast for a general value of τ

$$\begin{aligned} \hat{y}^{S}_{T+1}(\tau) &= \frac{1}{1+W(\tau)}\boldsymbol{x}'_{T+1}\hat{\boldsymbol{\beta}}_F + \frac{W(\tau)}{1+W(\tau)}\boldsymbol{x}'_{T+1}\hat{\boldsymbol{\beta}}_2(\tau) \\ &\to \frac{1}{1+Q^*(\tau)}\boldsymbol{x}'_{T+1}\hat{\boldsymbol{\beta}}_F + \frac{Q^*(\tau)}{1+Q^*(\tau)}\boldsymbol{x}'_{T+1}\hat{\boldsymbol{\beta}}_2(\tau) \end{aligned} \tag{3.42}$$

where the last line holds by the continuous mapping theorem. The asymptotic expressions for $\hat{\boldsymbol{\beta}}_2$ and $\hat{\boldsymbol{\beta}}_F$ are provided in (3.13) and (3.15). The difference in MSFE between the shrinkage forecast and the full sample forecast depends on the distribution of the break date

$$\begin{aligned} \Delta_s \;=\; & \mathrm{E}_{\hat{\tau}}\left[\left(\frac{1}{1+Q^*(\hat{\tau})}\boldsymbol{x}'_{T+1}(\hat{\boldsymbol{\beta}}_F-\boldsymbol{\beta}_2) + \frac{Q^*(\hat{\tau})}{1+Q^*(\hat{\tau})}\boldsymbol{x}'_{T+1}(\hat{\boldsymbol{\beta}}_2(\hat{\tau})-\boldsymbol{\beta}_2)\right)^2\right] \\ & -\mathrm{E}\left[\left(\boldsymbol{x}'_{T+1}(\hat{\boldsymbol{\beta}}_F-\boldsymbol{\beta}_2)\right)^2\right] \end{aligned} \tag{3.43}$$

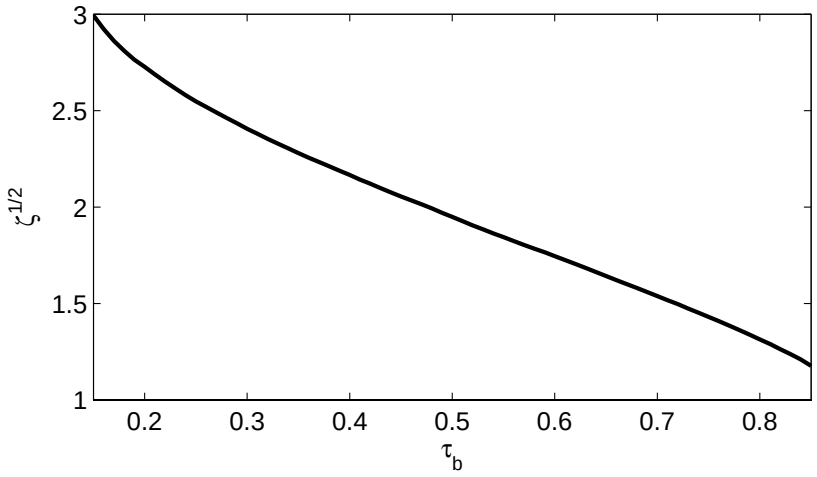
where we solve for $\Delta_s = 0$ numerically to obtain the break size that corresponds to equal predictive accuracy. Numerical results in Appendix 3.A.3 show that equal predictive accuracy is associated with a unique break size.

3.5 Simulations

3.5.1 Asymptotic analysis

The theoretical results of the previous section are derived under the assumption that the nominal size tends to zero. In this section, we investigate the properties of our test using simulations under conventional choices for nominal size, $\alpha = \{0.10, 0.05, 0.01\}$. We will study for which break size the difference between the MSFE from the post-break forecast equals that of the full sample forecast. Conditional on this break size, we use simulation to obtain critical values. Finally, we study the size and power properties of the resulting test.

Figure 3.1: Break size for equal predictive accuracy between post-break and full sample forecasts



Note: The graph shows the standardized break size, $\zeta^{1/2}$, in (3.44) for which the forecasts based on the post-break sample and the full sample achieve the same MSFE, that is, Δ in (3.19) equals zero.

Implementation

We simulate (3.23) with (3.24) for different combinations of the break date and break size $\{\tau_b, \theta_{\tau_b}\}$. Here, we focus on $\tau_b = \{\tau_{\min}, \tau_{\min} + \delta, \ldots, \tau_{\max}\}$ where $\tau_{\min} = 0.15$, $\tau_{\max} = 1 - \tau_{\min}$ and $\delta = 0.01$. Additional results for a wider grid with $\tau_{\min} = 0.05$ are presented in Appendix 3.B. For the break size parameter θ_{τ_b} we consider $\theta_{\tau_b} = \{0, 0.5, \ldots, 20\}$. The Brownian motion is approximated by dividing the $[0, 1]$ interval in $n = 1,000$ equally spaced parts, generating $\epsilon_i \sim N(0, 1)$ and $B(\tau) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n\tau} \epsilon_i$, see, for example, Bai and Perron (1998).

By maximizing (3.23) we obtain a distribution of the estimated break date $\hat{\tau}$ that can be used to evaluate (3.19). To approximate the expectation, we use 50,000 repetitions for each break date and break size. For each value of τ_b, a break size θ_{τ_b} is obtained for which the full sample forecast and the post-break forecast yield equal predictive accuracy using (3.19). This translates the null hypothesis of equal predictive accuracy into a null hypothesis regarding the break size conditional of the break date τ_b. By simulating under the null hypothesis for each τ_b, we obtain critical values that are dependent on τ_b. If the break date is estimated with sufficient accuracy, these critical values can be used for testing without correction. The size of the breaks that we find under the null hypothesis suggest that the estimated break date will, indeed, be quite accurate.

Post-break forecast versus full-sample forecast: break size for equal forecast accuracy

The break size for which the full sample and the post-break sample achieve equal predictive accuracy can be simulated using (3.19) as outlined above. Figure 3.1 shows the combinations

Table 3.1: Critical values and size of the W and S test statistics

		Critical values					Size				
Test	α	0.15	0.25	0.50	0.75	0.85	0.15	0.25	0.50	0.75	0.85
W	0.10	20.44	17.99	14.13	11.04	9.36	0.13	0.12	0.11	0.09	0.06
	0.05	23.71	20.99	16.74	13.30	11.37	0.07	0.06	0.06	0.04	0.03
	0.01	30.54	27.29	22.29	18.22	15.82	0.01	0.01	0.01	0.01	0.01
S	0.10	1.78	1.84	1.89	1.80	1.59	0.10	0.10	0.11	0.11	0.08
	0.05	2.12	2.18	2.23	2.14	1.94	0.05	0.05	0.06	0.05	0.04
	0.01	2.76	2.81	2.87	2.80	2.60	0.01	0.01	0.01	0.01	0.01

Note: Reported are critical values and size for, first, W, the Wald test statistic (3.17) and, second, S, the test statistic (3.38), which is independent of τ_b when the nominal size tends to zero.

of break size and break date for which equal predictive accuracy is obtained. The break size is given in units of the standardized break size,

$$\zeta^{1/2} = \sqrt{T(1-\tau_b)\tau_b}\frac{\boldsymbol{f}'_{\beta_2}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)}{\sqrt{\boldsymbol{f}'_{\beta_2}\boldsymbol{V}\boldsymbol{f}_{\beta_2}}} \tag{3.44}$$

so that it can be interpreted as a standard deviations from a standard normal.

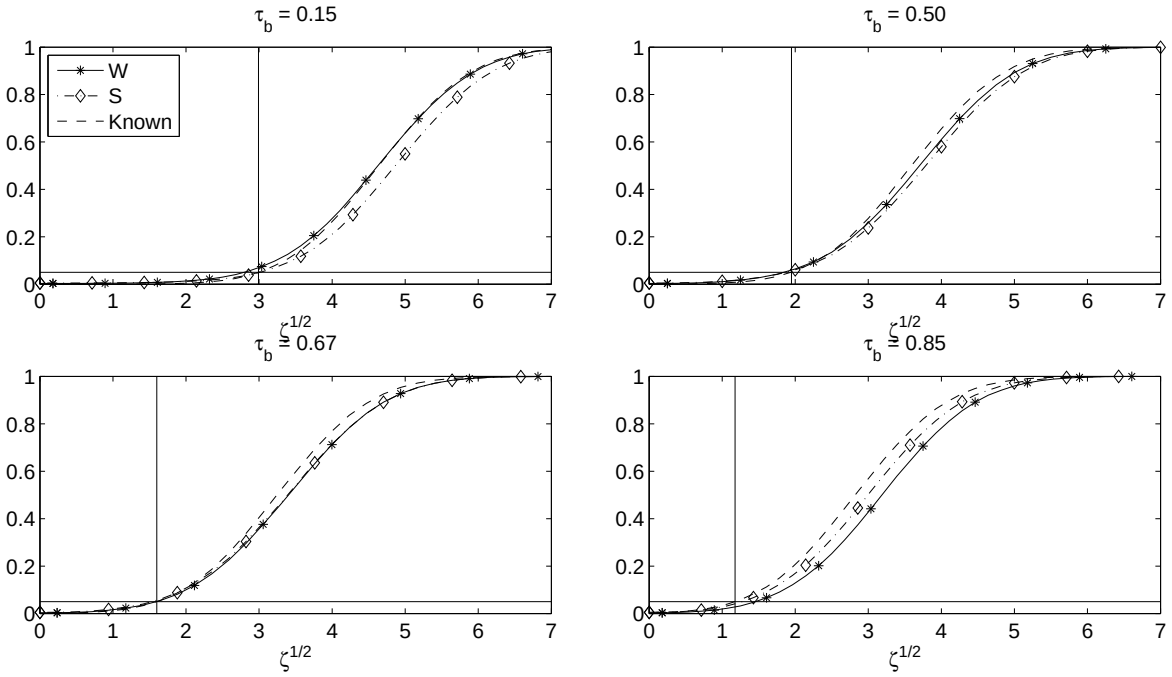
The figure shows that for each break date τ_b, the break size is substantially larger than the break size under a known break date, which yielded $\zeta^{1/2} = 1$. This illustrates the increase in the MSFE of the post-break sample forecast due to the fact that the break date needs to be estimated. The importance of this effect is clear. If a break occurs in the beginning of the sample, then we choose for the post-break forecast if the break size larger than three standard deviations. For breaks that occur closer to the end of the sample, this break size uniformly decreases. This provides further evidence for the intuition that breaks that occur at the end of the sample are the main reason for forecast failure.

Critical values, size and power

After finding the break size for which the post-break forecast and the full sample forecast yield equal predictive accuracy, we can compute critical values for both the Wald-type test statistic, W, in (3.22) and the α-asymptotic statistic, S, in (3.38) for a grid of break dates τ_b. Condition (3.36), which is required for the weak optimality result does hold for all τ_b—details are available in Appendix 3.A.2.

The third line of the right panel of Table 3.1 shows that the test has the correct size for $\alpha = 0.01$. For $\alpha = 0.05$ and 0.1 size is still very close to the nominal size, only at the beginning and the end of the sample some distortion occurs. However, using the corrected test statistic (3.38) largely remedies these size distortions.

Figure 3.2: Asymptotic power when testing between a post-break and full-sample forecast at $\alpha = 0.05$



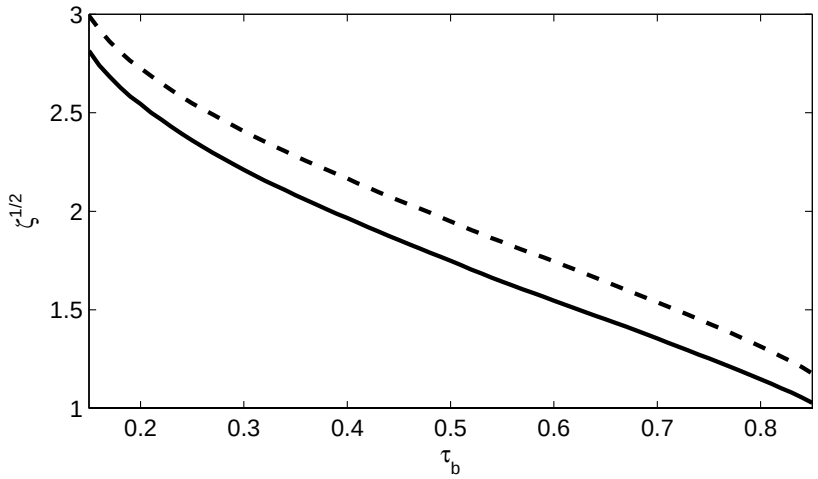
Note: The plots show the power for tests at a nominal size of $\alpha = 0.05$ with the null hypothesis given by the break size depicted in Figure 3.1. The panels show power for different values of the (unknown) break date. The power of infeasible test conditional on the true break date is given as the dashed line, that of the test statistic W as the solid line with stars, and that of the test statistic S as the dashed line with diamonds. The solid horizontal line indicates the nominal size, and the vertical solid line indicates the break size at which equal predictive accuracy is achieved corresponding to Figure 3.1.

The critical values are given in the left panel of Table 3.1. Critical values for a finer grid of the true break date can be found in Appendix 3.B. The large break size that yields equal forecast accuracy implies a major increase in critical values when using the Wald test statistic (3.22), compared to the standard values of Andrews (1993). For a nominal size of $[0.10, 0.05, 0.01]$ the critical values in Andrews are equal to $[7.17, 8.85, 12.35]$.

The critical values for the α-asymptotic test statistic, S, in (3.38) are independent of $\hat{\tau}$ in the limit where $\alpha \to 0$. Under a known break date, critical values would be from a one-sided normal distribution, that is, they would be $[1.64, 2.33, 2.58]$ for nominal size of $[0.10, 0.05, 0.01]$. The critical values for the corrected test, S, in (3.38) vary substantially less over $\hat{\tau}$ than those for the Wald statistic, W, in (3.22). The results in Section 3.4.3 suggest that the differences to the critical values that would be used if the break date is known diminish as $\alpha \to 0$ and this can indeed be observed in Table 3.1.

Given that the break sizes that lead to equal forecast performance are reasonably large, we expect the tests to have relatively good power properties. The power curves in Figure 3.2 show that the power of both tests is close to the power of the optimal test which uses the known break date to test whether the break size exceeds the boundary depicted in Figure 3.1.

Figure 3.3: Break size for equal predictive accuracy of shrinkage and full sample forecasts



Note: The solid line shows the standardized break size for which the shrinkage forecast (3.42) achieves the same MSFE as the full sample forecast, in which case (3.43) equals zero. For comparison, the dashed line shows the break size for which the post-break forecast and the full sample forecast achieve equal MSFE.

The good power properties are true for all break dates. This confirms that the theoretical results for vanishing nominal size extend to conventional choices of the nominal size.

Shrinkage forecast versus full-sample forecast

We now turn to the shrinkage forecast of Section 3.4.4. Figure 3.3 shows the combination of τ_b and break size for which the shrinkage forecast and full sample forecast that weights observations equally have the same MSFE, which is represented by the solid line in the graph. For comparison, the dashed line gives the combination of post-break forecast and full sample forecast that have the same MSFE. It can be seen that the break size for equal forecast performance for the shrinkage forecast is lower than for the post-break sample forecast. This implies that the shrinkage forecast is more precise than the post-break forecast for smaller break sizes for a given break date. However, the difference is relatively small and breaks need to be quite large before the shrinkage estimator is more precise than the full sample estimator.

In order to determine whether to use the shrinkage forecast, critical values can be obtained in a similar fashion as before and are presented in Table 3.2. Again, the size is close to the theoretical size with small size disturbances when using W, which are largely remedied when using S. Critical values on a finer grid of the true break date are presented in Appendix 3.B. Figure 3.4 displays the power curves of the tests that compare the shrinkage forecast and the full sample, equal weights forecast. Since, the break sizes for equal forecast performance are similar to the post-break sample forecast, it is not surprising that the prop-

Table 3.2: Critical values and size: shrinkage versus full sample forecasts

		Critical values					Size				
Test	α	0.15	0.25	0.50	0.75	0.85	0.15	0.25	0.50	0.75	0.85
W	0.10	19.01	16.63	12.95	10.19	8.82	0.14	0.13	0.11	0.08	0.06
	0.05	22.15	19.51	15.43	12.34	10.74	0.07	0.07	0.06	0.04	0.03
	0.01	28.74	25.57	20.74	17.03	15.02	0.02	0.01	0.01	0.01	0.01
S	0.10	1.85	1.90	1.93	1.82	1.63	0.10	0.10	0.11	0.11	0.08
	0.05	2.18	2.24	2.27	2.17	1.98	0.05	0.05	0.06	0.05	0.04
	0.01	2.82	2.87	2.91	2.82	2.63	0.01	0.01	0.01	0.01	0.01

Note: Reported are critical values and size when testing for equal MSFE of the shrinkage forecast (3.42) and the full sample forecast using, first, W, the Wald test statistic in (3.17) and, second, S, the test statistic (3.38) that is independent of τ_b when the nominal size tends to zero.

erties in terms of size and power of the tests for the shrinkage forecast are largely the same as those for the post-break forecast.

Shrinkage forecast versus the post-break forecast

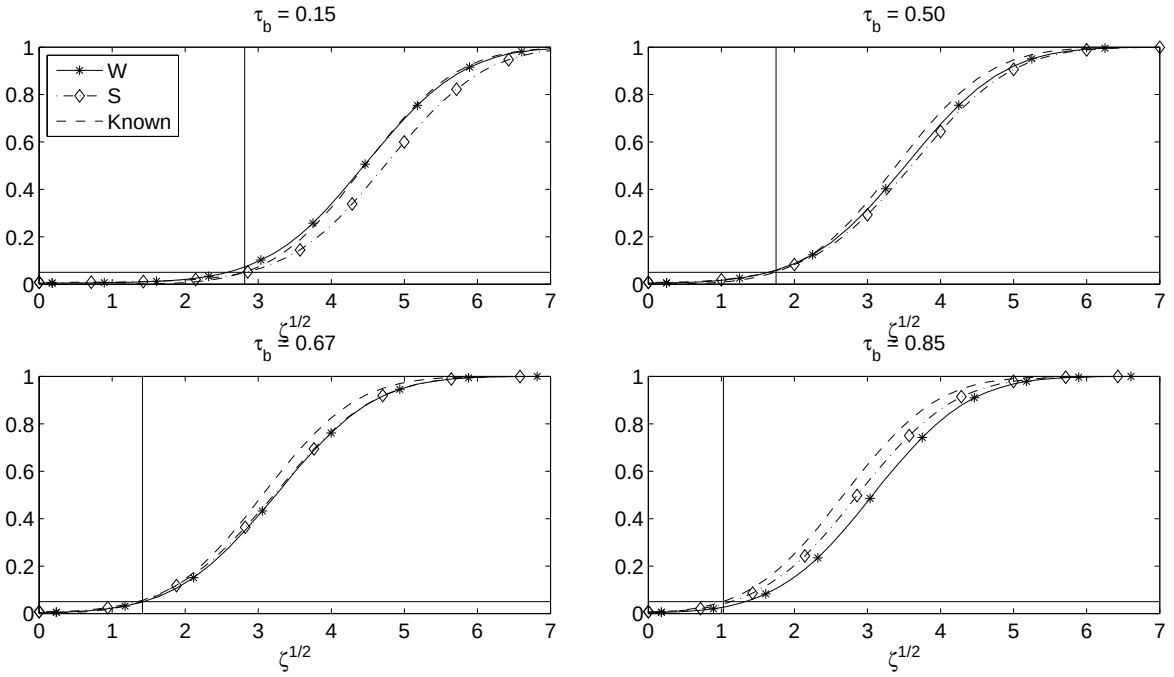
In addition to comparing post-break sample and shrinkage forecasts to the full sample forecast, we can investigate the break sizes that leads to equal forecast performance of the post-break forecast and the shrinkage forecast. Figure 3.5 plots the ratio of the MSFE of the shrinkage forecast over that of the post-break forecast. Nearly for all break sizes and dates, the shrinkage forecast outperforms the post-break forecast. Only when the break occurs at the end of the sample and is relatively large, the post-break forecast is slightly more accurate. This suggests that one can improve over the post-break estimator in a wide range of settings.

3.5.2 Finite sample analysis

Set up of the Monte Carlo experiments

We analyze the performance of the tests in finite sample for an AR(1) model with varying degree of persistence. We consider the two tests for equal predictive accuracy between the post-break forecast and the full-sample forecast based on the Wald statistic (3.17) and on the S-statistic (3.38). Next, we consider the same test statistics but now test for equal predictive accuracy between the shrinkage forecast (3.42) and the full-sample, equal weighted forecast. All tests are carried out at a nominal size $\alpha = 0.05$, using sample sizes of $T = \{120, 240, 480\}$ and break dates $\tau_b = [0.15, 0.25, 0.50, 0.75, 0.85]$. Parameter estimates are obtained by least squares, and the results are based on 10,000 repetitions.

Figure 3.4: Asymptotic power when testing at $\alpha = 0.05$ between the shrinkage and full-sample forecast



Note: The plots show asymptotic power curves when testing for equal predictive accuracy between the shrinkage forecast (3.42) and the full-sample forecast using the break size depicted in Figure 3.3 for different values of the break date τ_b. For more information, see the footnote of Figure 3.2.

The DGP is given by

$$y_t = \mu_t + \rho y_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim N(0, \sigma^2) \tag{3.45}$$

where $\sigma^2 = 1$ and

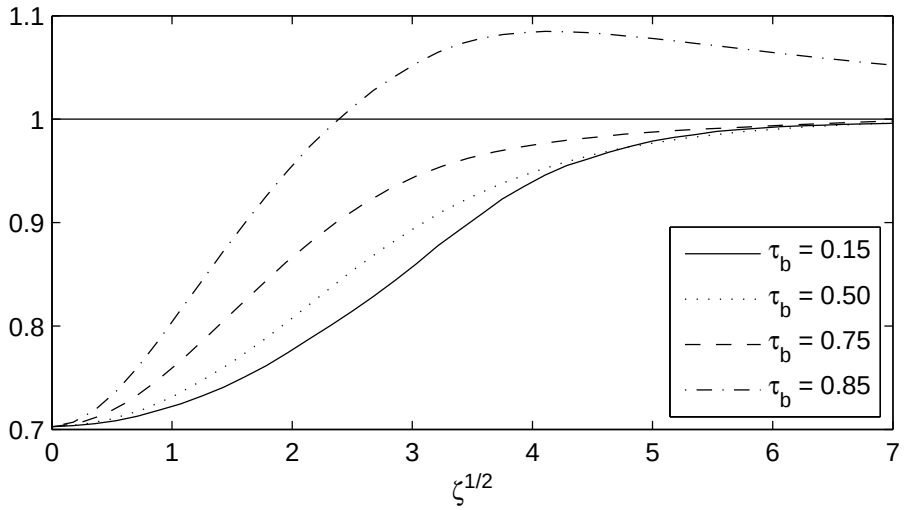
$$\mu_t = \begin{cases} \mu_1 & \text{if } t \leqslant \tau_b T \\ \mu_2 & \text{if } t > \tau_b T \end{cases}$$

We set $\mu_1 = -\mu_2$ and $\mu_1 = \frac{1}{2\sqrt{T}}\zeta^{1/2}(\tau_b) + \frac{1}{2}\frac{\lambda}{\sqrt{T\tau_b(1-\tau_b)}}$. To investigate the finite sample size of the tests, we choose $\lambda = 0$ which yields the asymptotic the break size from Figure 3.1. To investigate power, we choose $\lambda = \{1, 2\}$. The influence of the degree of persistence on the results is analyzed by varying $\rho = \{0.0, 0.3, 0.6, 0.9\}$.

Results

The results in Table 3.3 show that for models with low and moderate persistence, that is, $\rho = 0.0$ or 0.3, the size of the W and S tests are extremely close to the nominal size irrespective of the sample size and the break date. When persistency increases to $\rho = 0.9$, some size distortions become apparent for $T = 120$. This does, however, diminish as T increases.

Figure 3.5: Relative MSFE of shrinkage and post-break sample forecasts



Note: The graph shows the relative performance of the shrinkage forecast (3.42) and the post-break sample forecast as a function of the standardized break size $\zeta^{1/2}$ for different values of the break date τ_b. The horizontal solid line corresponds to equal predictive accuracy. Values below 1 indicate that the shrinkage forecast is more precise.

These size distortions are similar for W and S and are the result of the small effective sample size in this setting.

To analyze power, we increase the break size with $\lambda = \{1, 2\}$. For $T = 120$ it is slightly larger when the break is in the middle of the sample but this effect disappears with increasing T. Overall, differences between W and S are small.

The results for the tests that compare the shrinkage forecast against the full sample, equal weights forecast in Table 3.4 are very similar to the results for the test with the post-break sample forecast under the alternative. Size is very close to the nominal size for large effective sample sizes and power increases in λ and, mildly, in T.

Overall, the results suggest that the W and S tests have good size and power properties unless the persistence of the time series is very high and this is combined with a small effective T.

3.6 Application

We investigate the importance of structural breaks for 130 macroeconomic and financial time series from the St. Louis Federal Reserve database, which is a monthly updated database. The data are described by McCracken and Ng (2015), who suggest various transformations are applied to render the series stationary and to deal with discontinued series or changes in classification. In the vintage used here, the data start in 1959M01 and end in 2015M10. After the transformations, all 130 series are available from 1960M01 onwards.

Table 3.3: Finite sample analysis: size and power when testing between post-break and full-sample forecast

		$T = 120$					$T = 240$					$T = 480$				
ρ	λ	0.15	0.25	0.50	0.75	0.85	0.15	0.25	0.50	0.75	0.85	0.15	0.25	0.50	0.75	0.85
Wald-test (3.17)																
0.0	0	0.05	0.05	0.06	0.05	0.03	0.06	0.06	0.06	0.04	0.03	0.06	0.06	0.06	0.05	0.03
	1	0.17	0.20	0.22	0.21	0.17	0.21	0.22	0.23	0.21	0.16	0.24	0.24	0.23	0.21	0.16
	2	0.43	0.48	0.52	0.53	0.47	0.52	0.54	0.55	0.53	0.48	0.57	0.56	0.56	0.55	0.49
0.3	0	0.04	0.05	0.06	0.05	0.03	0.05	0.06	0.06	0.04	0.03	0.06	0.06	0.06	0.05	0.03
	1	0.13	0.17	0.21	0.21	0.17	0.18	0.20	0.22	0.20	0.16	0.22	0.23	0.22	0.21	0.16
	2	0.33	0.40	0.47	0.50	0.46	0.46	0.50	0.53	0.52	0.47	0.54	0.54	0.55	0.55	0.48
0.6	0	0.03	0.05	0.06	0.05	0.04	0.04	0.05	0.06	0.05	0.03	0.05	0.06	0.06	0.05	0.03
	1	0.08	0.12	0.19	0.20	0.16	0.13	0.17	0.20	0.20	0.15	0.18	0.20	0.22	0.21	0.15
	2	0.19	0.26	0.39	0.46	0.43	0.33	0.40	0.47	0.50	0.45	0.47	0.49	0.52	0.53	0.47
0.9	0	0.02	0.05	0.10	0.09	0.06	0.02	0.04	0.08	0.07	0.04	0.03	0.05	0.06	0.06	0.04
	1	0.04	0.07	0.17	0.24	0.20	0.04	0.08	0.16	0.21	0.16	0.07	0.11	0.17	0.20	0.15
	2	0.09	0.12	0.24	0.44	0.44	0.09	0.14	0.28	0.43	0.39	0.16	0.24	0.37	0.46	0.41
S-test (3.38)																
0.0	0	0.03	0.04	0.06	0.06	0.04	0.04	0.05	0.05	0.05	0.04	0.04	0.05	0.06	0.06	0.04
	1	0.13	0.16	0.21	0.23	0.22	0.16	0.18	0.21	0.23	0.21	0.17	0.19	0.21	0.23	0.21
	2	0.34	0.41	0.48	0.56	0.55	0.43	0.48	0.52	0.56	0.56	0.48	0.51	0.53	0.58	0.56
0.3	0	0.03	0.04	0.06	0.06	0.04	0.04	0.05	0.06	0.05	0.04	0.04	0.05	0.06	0.06	0.04
	1	0.09	0.14	0.19	0.23	0.22	0.13	0.16	0.20	0.23	0.20	0.16	0.18	0.21	0.23	0.21
	2	0.25	0.34	0.44	0.53	0.54	0.36	0.43	0.50	0.55	0.55	0.44	0.49	0.52	0.58	0.56
0.6	0	0.02	0.04	0.06	0.07	0.05	0.03	0.04	0.05	0.05	0.04	0.04	0.05	0.06	0.06	0.05
	1	0.05	0.09	0.17	0.23	0.21	0.09	0.13	0.19	0.22	0.21	0.13	0.16	0.20	0.23	0.21
	2	0.13	0.21	0.36	0.50	0.52	0.24	0.33	0.44	0.53	0.53	0.37	0.43	0.49	0.56	0.55
0.9	0	0.02	0.04	0.10	0.12	0.08	0.02	0.03	0.07	0.08	0.06	0.02	0.04	0.06	0.07	0.05
	1	0.03	0.05	0.16	0.28	0.26	0.02	0.06	0.14	0.24	0.22	0.04	0.08	0.16	0.23	0.21
	2	0.06	0.08	0.22	0.49	0.54	0.05	0.10	0.25	0.47	0.49	0.10	0.18	0.33	0.50	0.51

Note: The table presents finite sample size and power properties for the test comparing the post-break and full sample based forecasts. The DGP is $y_t = \mu_t + \rho y_{t-1} + \varepsilon_t$, $\varepsilon_t \sim N(0,1)$, $\mu_1 = -\mu_2$ and $\mu_1 = \frac{1}{2\sqrt{T}}\zeta^{1/2}(\tau_b) + \frac{1}{2}\frac{\lambda}{\sqrt{T\tau_b(1-\tau_b)}}$ where $\zeta^{1/2}(\tau_b)$ corresponds to Figure 3.1. The empirical size of the tests is obtained when $\lambda = 0$ and power when $\lambda = \{1, 2\}$. Tests are for a nominal size of 0.05.

The data are split into 8 groups: *output and income* (OI, 17 series), *labor market* (LM, 32 series), *consumption and orders* (CO, 10 series), *orders and inventories* (OrdInv, 11 series), *money and credit* (MC, 14 series), *interest rates and exchange rates* (IRER, 21 series), *prices* (P, 21 series), *stock market* (S, 4 series).

Following Stock and Watson (1996), we focus on linear autoregressive models of lag length $p = 1$ and $p = 6$ and test whether the intercept is subject to a break. We esti-

Table 3.4: Finite sample analysis: size and power when testing between shrinkage and full-sample forecast

		$T = 120$					$T = 240$					$T = 480$				
ρ	λ	0.15	0.25	0.50	0.75	0.85	0.15	0.25	0.50	0.75	0.85	0.15	0.25	0.50	0.75	0.85
Wald-test (3.17)																
0.0	0	0.05	0.06	0.06	0.04	0.03	0.06	0.06	0.06	0.04	0.03	0.07	0.07	0.06	0.05	0.03
	1	0.18	0.21	0.22	0.21	0.16	0.22	0.23	0.22	0.20	0.15	0.24	0.24	0.23	0.21	0.15
	2	0.45	0.49	0.52	0.52	0.46	0.53	0.55	0.55	0.53	0.47	0.57	0.57	0.56	0.54	0.48
0.3	0	0.05	0.06	0.06	0.05	0.03	0.06	0.06	0.06	0.04	0.03	0.06	0.07	0.06	0.05	0.03
	1	0.15	0.19	0.22	0.20	0.16	0.20	0.21	0.22	0.20	0.15	0.23	0.23	0.22	0.21	0.15
	2	0.36	0.42	0.48	0.51	0.45	0.48	0.51	0.53	0.52	0.46	0.55	0.55	0.55	0.54	0.47
0.6	0	0.04	0.06	0.07	0.05	0.04	0.05	0.06	0.06	0.04	0.03	0.06	0.06	0.06	0.05	0.03
	1	0.10	0.14	0.20	0.20	0.16	0.15	0.18	0.21	0.20	0.15	0.20	0.21	0.22	0.20	0.15
	2	0.22	0.30	0.42	0.47	0.43	0.36	0.42	0.49	0.50	0.44	0.48	0.51	0.53	0.52	0.46
0.9	0	0.03	0.07	0.12	0.10	0.07	0.04	0.05	0.09	0.07	0.05	0.04	0.06	0.07	0.06	0.04
	1	0.06	0.09	0.21	0.26	0.21	0.06	0.10	0.19	0.22	0.17	0.09	0.13	0.20	0.21	0.16
	2	0.11	0.15	0.30	0.48	0.45	0.12	0.18	0.34	0.46	0.41	0.20	0.28	0.41	0.47	0.42
S-test (3.38)																
0.0	0	0.04	0.05	0.06	0.06	0.04	0.04	0.05	0.06	0.05	0.04	0.04	0.05	0.06	0.06	0.04
	1	0.13	0.16	0.21	0.24	0.22	0.15	0.18	0.21	0.23	0.20	0.17	0.19	0.21	0.23	0.20
	2	0.34	0.42	0.49	0.56	0.55	0.42	0.47	0.52	0.56	0.55	0.46	0.50	0.53	0.58	0.56
0.3	0	0.03	0.05	0.06	0.06	0.04	0.04	0.05	0.06	0.05	0.04	0.04	0.05	0.06	0.06	0.04
	1	0.10	0.14	0.20	0.23	0.22	0.13	0.17	0.20	0.23	0.20	0.16	0.18	0.21	0.23	0.20
	2	0.25	0.35	0.45	0.54	0.54	0.36	0.43	0.50	0.56	0.54	0.43	0.48	0.52	0.58	0.55
0.6	0	0.03	0.05	0.07	0.07	0.05	0.03	0.05	0.06	0.06	0.04	0.04	0.05	0.06	0.06	0.05
	1	0.06	0.10	0.19	0.24	0.22	0.09	0.14	0.19	0.23	0.20	0.13	0.16	0.20	0.23	0.20
	2	0.14	0.23	0.39	0.52	0.52	0.25	0.35	0.45	0.54	0.53	0.37	0.43	0.49	0.56	0.55
0.9	0	0.02	0.05	0.12	0.12	0.09	0.02	0.04	0.09	0.09	0.06	0.03	0.04	0.07	0.08	0.06
	1	0.03	0.06	0.19	0.31	0.28	0.03	0.07	0.17	0.25	0.23	0.05	0.10	0.18	0.25	0.21
	2	0.06	0.11	0.27	0.53	0.56	0.07	0.12	0.30	0.50	0.51	0.12	0.21	0.37	0.52	0.52

Note: The table presents finite sample size and power properties of the tests comparing the shrinkage forecast (3.42) and the full-sample, equal weights forecast, using a nominal size of 0.05. For further details, see the footnote of Table 3.3.

mate parameters on a moving windows of 120 observations to decrease the likelihood of multiple breaks occurring in the estimation sample. Test results are based on heteroskedasticity robust Wald statistics, which use the following estimate of the covariance matrix $\hat{\boldsymbol{V}}_i = (\boldsymbol{X}_i'\boldsymbol{X}_i)^{-1}\boldsymbol{X}_i'\hat{\boldsymbol{\Omega}}_i\boldsymbol{X}_i(\boldsymbol{X}_i'\boldsymbol{X}_i)^{-1}$ with $[\hat{\boldsymbol{\Omega}}_i]_{kl} = \hat{\varepsilon}_k^2/(1-h_k)^2$ if $k = l$ and $[\hat{\boldsymbol{\Omega}}_i]_{kl} = 0$ otherwise, and h_k is the k-th diagonal element of $\boldsymbol{P}_{\boldsymbol{X}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$. See MacKinnon and White (1985) and Long and Ervin (2000) for discussions of different heteroskedasticity

Table 3.5: Fractions of estimation samples with a significant structural break

	supW	W	S	W^s	S^s
AR(1)	0.219	0.102	0.108	0.119	0.126
AR(6)	0.114	0.037	0.042	0.046	0.053

Note: supW refers to the Andrews' (1993) sup-Wald test, W and S refer to the tests developed in this chapter that compare post-break and full sample forecasts, and W^s and S^s refer to the tests that compare shrinkage and full sample forecasts. All tests are carried out at $\alpha = 0.05$.

robust covariance matrices. We have also obtained test results and forecasts using a larger window of 240 observations and using the homoskedastic Wald test and, qualitatively, our results do not depend on these choices.

To initialize the AR(p) model, we require p observations. Our first forecast is therefore for July 1970 and we recursively construct one-step ahead forecasts until the end of the sample.

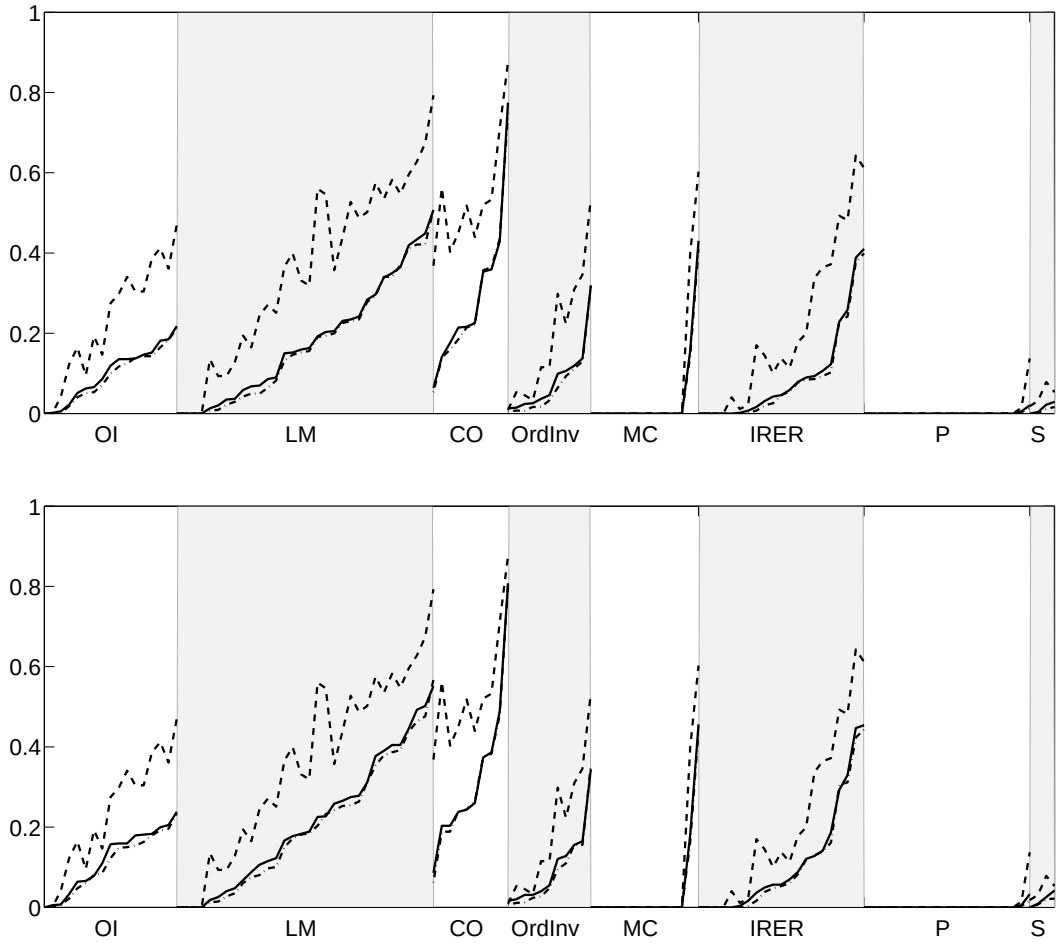
3.6.1 Structural break test results

In this forecast exercise, we will refer to the test of Andrews (1993) as *supW*, the Wald test statistic (3.17) as *W*, the test statistic (3.38) as *S*, and, when the alternative is the shrinkage forecast, these tests as W^s and S^s. In Table 3.5, we report the fraction of estimation samples where *supW* would indicate a break at a nominal size of $\alpha = 0.05$. This is contrasted with the fraction where our tests indicate a break also at a nominal size of $\alpha = 0.05$. It is clear that a large fraction of the breaks picked up by *supW* are judged as irrelevant for forecasting by *W* and *S*. The fraction of forecasts for which a break is indicated is lower by a factor of over two for the AR(1) and by factor of up to three for the AR(6).

Figure 3.6 displays the number of estimation samples per series for which the tests were significant when forecasting with the AR(1), where within each category we sort the series based on the fraction of breaks found by *W*. Across all categories the *supW* test is more often significant than the *W* and *S* test. Yet, we see substantial differences between categories. Whereas in the *labor market* and *consumption and orders* categories some of the series contain a significant breaks in up to 70% of the estimation samples when the *W* or *S* tests are used, the *prices* and *stock market* series hardly show any significant breaks from a forecasting perspective. This finding concurs with the general perception that, for these type of time series, simple linear models are very hard to beat in terms of MSFE.

Figure 3.7 displays the number of estimation samples with significant breaks for the AR(6) model. Compared to the results for the AR(1) in Figure 3.6, far fewer estimation samples contain a significant break, and this is true even in the consumption and orders

Figure 3.6: Fraction of significant structural break test statistics per series - AR(1)

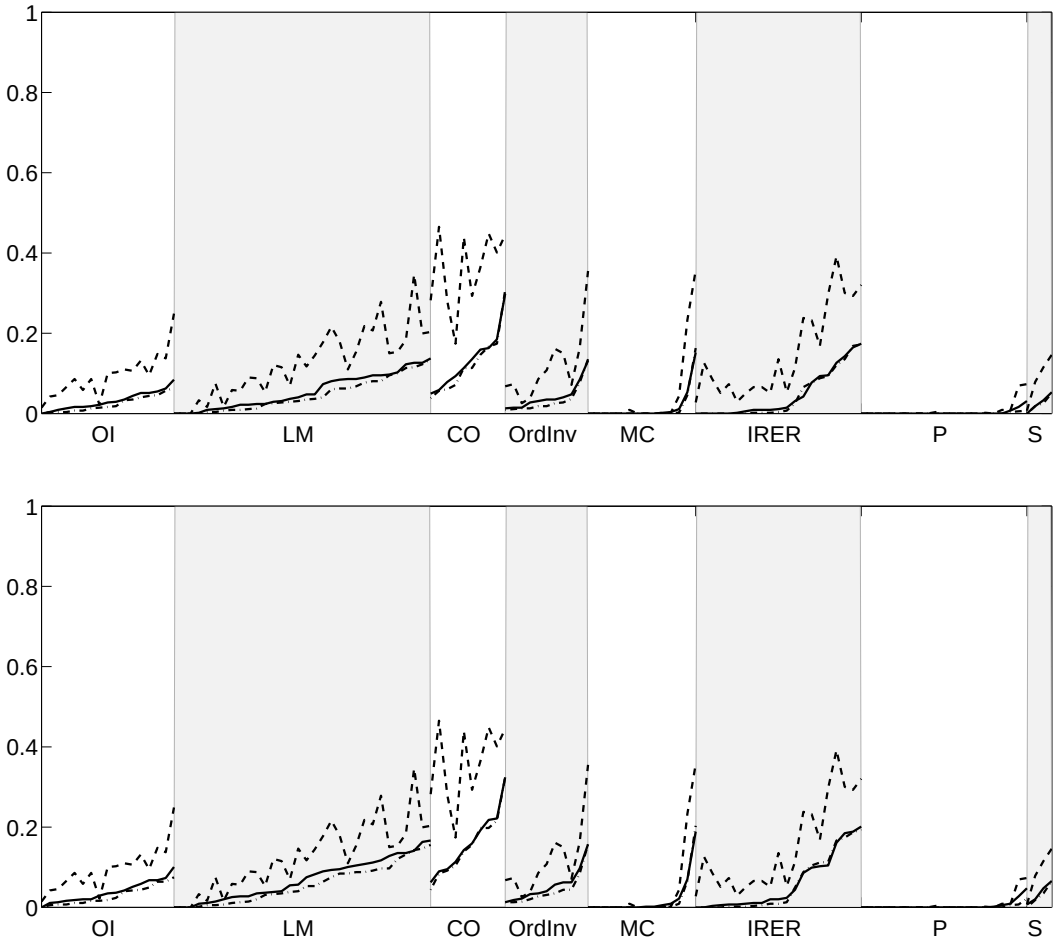


Note: The upper panel depicts the fraction of estimation samples with a significant break when testing under the alternative of the post-break forecast; the lower panel when testing under the alternative of the shrinkage forecast (3.42). Dashed lines indicate the fraction of estimation samples with significant sup-Wald test, dashed-dotted lines indicate the fraction of estimation samples where the break test W in (3.17) indicates a break, and solid lines indicate the fraction of estimation samples with significant S test in (3.38).

category, which contained series with many breaks when using the AR(1). Consistent with the results for the AR(1), however, the W and S tests find fewer estimation samples with breaks than the *supW* test for virtually all series.

Figure 3.8 shows the occurrence of significant breaks over the different estimation samples when using the AR(1) model, where the end date of the estimation sample is given on the horizontal axis. In the top panel are the results for the test comparing the post-break estimation window with the full estimation window. In the bottom panel are the tests comparing the shrinkage estimator and the full sample, equal weights estimator. It is clear that the *supW* test finds more breaks in for the vast majority of estimation samples, whereas the results from the W and S tests are extremely similar.

Figure 3.7: Fraction of significant structural break test statistics per series - AR(6)

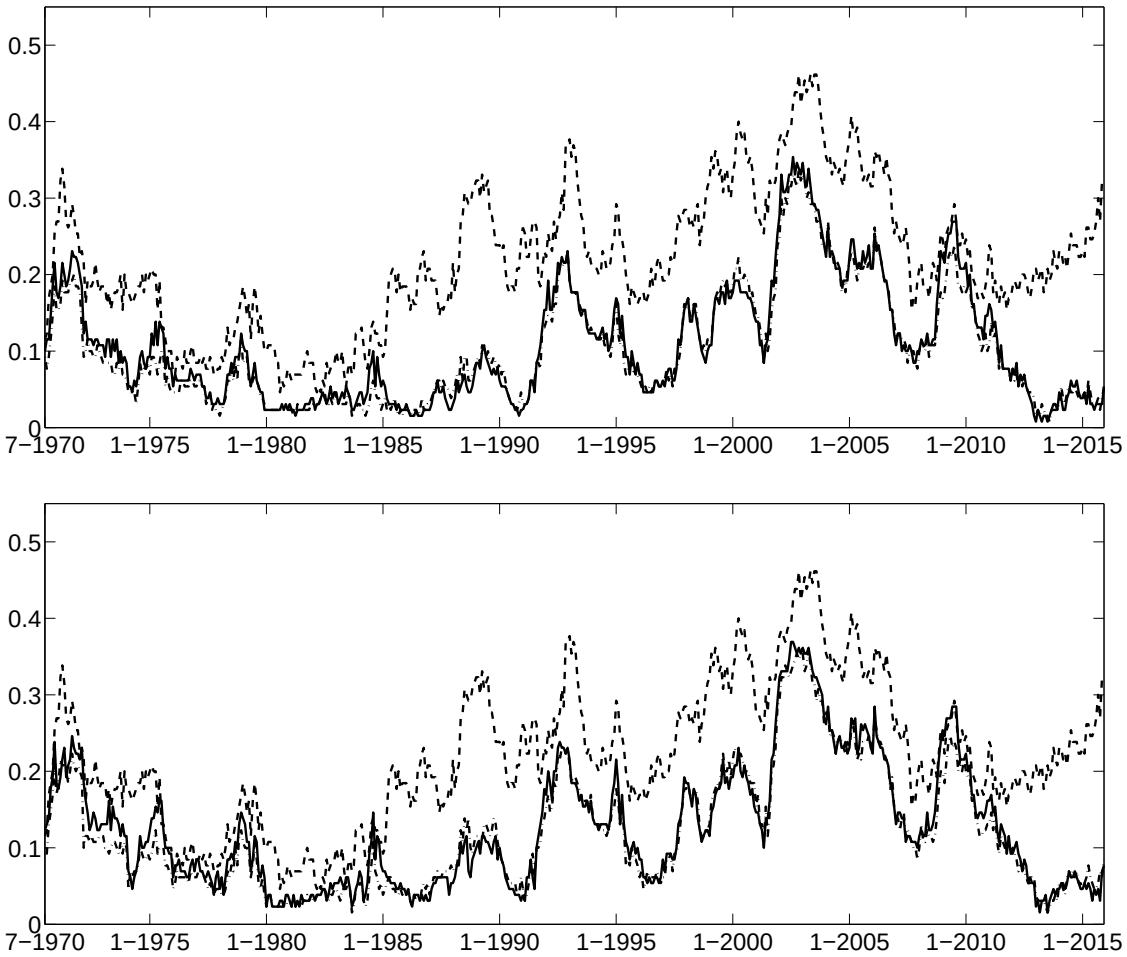


Note: See footnote of Table 3.6 for additional details.

A number of interesting episodes can be observed. While in the initial estimation samples the tests find a comparable number of samples with beaks, from 1985 the *supW* test finds many more series that contain breaks that are insignificant for the W and S test. This remains true until 2009 where the W and S tests find the same and, in the case of the shrinkage forecast, even more breaks that are relevant for forecasting than the *supW* test. From 2010 onwards, breaks that are relevant for forecasting decrease sharply, whereas the *supW* tests continues to find a large number of breaks.

Figure 3.9 shows the results but for the AR(6) model. All tests find fewer estimation samples with breaks compared to the AR(1) model. The evolution over the estimation samples is, however, similar to the AR(1) case. In the initial estimation samples up to 1985 all tests agree that a small number of series are subject to a structural break. From 1985 to 1990, however, the *supW* test finds breaks in up to a third of the estimation samples, which the W and S tests do not find important for forecasting. The same is true for breaks around 2000. In contrast, in the period following the dot com bubble and following the financial crisis of 2008/9 the W and the S tests find as many and, in the case of the shrinkage forecasts, more

Figure 3.8: Fraction of significant structural break test statistics over estimation samples – AR(1)



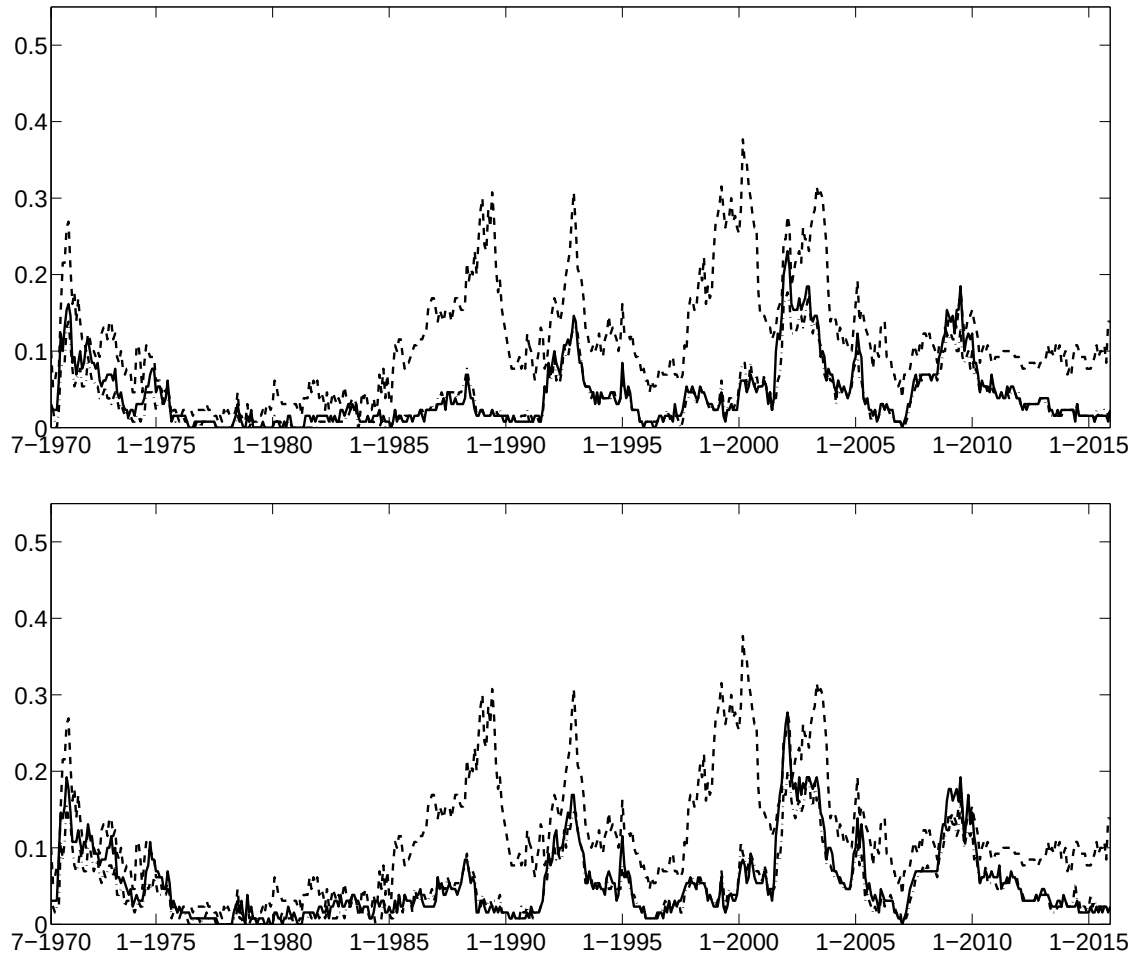
Note: The plots show the fractions of series with a significant break for each estimation sample when using an AR(1) model with a break in intercept. The top panel shows results when testing between the post-break sample based forecast and the full sample based forecast and the lower panel when testing between the shrinkage forecast and the full sample, equal weights forecast. The dashed line indicates the fraction of series when testing using the standard sup-Wald test at $\alpha = 0.05$, the solid line when testing using the S-test in (3.38), and the dashed-dotted line when testing using the W-test in (3.17). The dates displayed on the horizontal axis are the end dates of the estimation samples.

series, where taking a break into account will improve forecast accuracy than the *supW* test. Again, the number of series that should take breaks into account declines sharply towards the end of our sample when using the W and S tests but not when using the *supW* tests.

3.6.2 Forecast accuracy

In the next step, we are going to investigate whether forecasts conditional on the W and S tests are more accurate than forecasts based on the *supW* test. We use each test to determine whether to use the post-break or the full sample for forecasting or, alternatively, whether to use the shrinkage or the equal weights forecast. All tests are carried out at $\alpha = 0.05$.

Figure 3.9: Fraction of significant structural break test statistics over estimation samples – AR(6)



Note: The plots show the fractions of series with a significant break for each estimation sample when using an AR(6) model with a break in intercept. For additional details, see the footnote of Figure 3.8.

Table 3.6 reports the MSFE of the respective forecasting procedures relative to the MSFE of the forecast based on the supW test of Andrews with the results for the AR(1) in the top panel and those for the AR(6) in the bottom panel. For each model, we report the average relative MSFE over all series in the first line, followed by the average relative MSFE for the series in the different categories. We report only the results for the estimation windows where at least one test finds a break as the estimation samples where no test finds a break will to lead to identical full sample forecasts.

The results show that using the *W* test in place of the *supW* test leads to a 5.5% improvement in accuracy on average for the AR(1) and a 7.6% improvement in accuracy on average for the AR(6) model. This gain is similar for the S test with improvements of 4.9% and 6.5%. These improvements are found for series in all categories. The only exception is the use of the S test in the AR(1) model on the category 'prices'. This suggests that, while the improvements are modest, they are robust across the different series.

Table 3.6: Relative MSFE compared to the standard sup-Wald test

		Post-break		Shrinkage		
		W	S	W	S	supW
AR(1)	All series	0.948	0.953	0.948	0.949	0.983
	OI	0.972	0.981	0.970	0.972	0.986
	LM	0.950	0.951	0.948	0.948	0.979
	CO	0.978	0.973	0.975	0.969	0.992
	OrdInv	0.955	0.974	0.955	0.973	0.983
	MC	0.966	0.974	0.971	0.972	0.991
	IRER	0.878	0.891	0.889	0.892	0.974
	P	0.973	1.004	0.969	1.010	0.988
	S	0.924	0.961	0.926	0.928	0.979
AR(6)	All series	0.929	0.938	0.935	0.939	0.982
	OI	0.949	0.978	0.960	0.972	0.983
	LM	0.953	0.961	0.951	0.959	0.978
	CO	0.956	0.954	0.955	0.952	0.989
	OrdInv	0.926	0.953	0.935	0.948	0.983
	MC	0.948	0.957	0.960	0.974	0.990
	IRER	0.851	0.854	0.872	0.870	0.975
	P	0.921	0.940	0.939	0.914	0.985
	S	0.963	0.957	0.961	0.959	0.987

Note: The table reports the average of the ratio of the respective forecasts' MSFE over that of the forecasts resulting from the sup-Wald test of Andrews (1993) at $\alpha = 0.05$. Forecasts for which none of the tests indicate a break are excluded. Results are reported for the test statistic W in (3.17) and S in (3.38). 'Post-break' and 'Shrinkage' indicate that under the alternative the post-break forecast, respectively the shrinkage forecast (3.42), are used. The acronyms in the first column with corresponding series after excluding series without breaks (AR(1)|AR(6)): OI: output and income (16|17 series), LM: labor market (28|29), CO: consumption and orders (10|10), OrdInv: orders and inventories (11|11), MC: money and credit (2|8), IRER: interest rates and exchange rates (17|21), P: prices (2|6), S: stock market (4|4).

When the shrinkage forecast is used in conjunction with the W^s or S^s test, the accuracy of the forecasts is very similar as those of the post-break forecasts. This can be expected since we reject the test when the Wald statistic, that governs the amount of shrinkage, is relatively large. This implies that upon rejection of the test statistic, a forecast is used that is relatively close to the post-break forecast. The last column shows that using the shrinkage forecast in conjunction with the *supW* test leads to forecasts that, while more precise than post-break forecasts based on the same test, are clearly dominated by the W^s and S^s tests. In fact, for all categories and both models the W^s test leads to more accurate forecasts and the S^s tests for all categories and both models, with the exception of the AR(1) and prices.

3.7 Conclusion

In this chapter, we formalize the notion that small breaks might be better left ignored when forecasting. We quantify the break size that leads to equal forecast performance between a model based on the full sample and one based on a post-break sample. This break size is substantial, which points to a large penalty that is incurred by the uncertainty around the break date. A second finding is that the break size that leads to equal forecast performance depends on the unknown break date.

We derive a test for equal forecast performance. Under a local break no consistent estimator is available for the break date. Yet, we are able to prove weak optimality, in the sense that the power of an infeasible test conditional on the break date is achieved when we consider a small enough nominal size. This allows the critical values of the test to depend on the estimated break date. We show that under the break sizes we consider under our null hypothesis, this optimality is achieved relatively quickly, i.e. for finite nominal size. Simulations confirm this argument and show only a minor loss of power compared to the test is conditional on the true break date.

We apply the test on a large set of macroeconomic time series and find that breaks that are relevant for forecasting are rare. Pretesting using the test developed here improves over pretesting using the standard test of Andrews (1993) in terms of MSFE. Similar improvements can be made by considering an optimal weights or shrinkage estimator under the alternative.

3.A Additional mathematical details

3.A.1 Derivation of (3.19)

Define $\Delta = \Delta_1 - \Delta_2$ where

$$\begin{aligned}\Delta_1 &= T\mathrm{E}\left[\left(\partial_{\beta_2} f'(\hat{\boldsymbol{\beta}}_2(\hat{\tau}) - \boldsymbol{\beta}_2) + \partial_\delta f'(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta})\right)^2\right] \\ &= T\mathrm{E}\Bigg[\left(\partial_{\beta_2} f'(\hat{\boldsymbol{\beta}}_2(\hat{\tau}) - \boldsymbol{\beta}_2)\right)^2 + \left(\partial_\delta f'(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta})\right)^2 + \\ &\qquad +2\partial_{\beta_2} f'(\hat{\boldsymbol{\beta}}_2(\hat{\tau}) - \boldsymbol{\beta}_2)\partial_\delta f'(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta})\Bigg]\end{aligned} \tag{3.46}$$

and similarly for Δ_2

$$\begin{aligned}\Delta_2 &= T\mathrm{E}\left[\left(\partial_{\beta_2} f'(\hat{\boldsymbol{\beta}}_F(\hat{\tau}) - \boldsymbol{\beta}_2) + \partial_\delta f'(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta})\right)^2\right] \\ &= T\mathrm{E}\Bigg[\left(\partial_{\beta_2} f'(\hat{\boldsymbol{\beta}}_F - \boldsymbol{\beta}_2)\right)^2 + \left(\partial_\delta f'(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta})\right)^2 + \\ &\qquad +2\partial_{\beta_2} f'(\hat{\boldsymbol{\beta}}_F - \boldsymbol{\beta}_2)\partial_\delta f'(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta})\Bigg]\end{aligned} \tag{3.47}$$

In addition, we define

$$\begin{aligned}a &= \frac{1}{1-\hat{\tau}}\left[\partial_{\beta_2} f'(\bar{\boldsymbol{X}}'\bar{\boldsymbol{X}})^{-1}\bar{\boldsymbol{X}}'(\boldsymbol{B}(1) - \boldsymbol{B}(\hat{\tau})) + \int_{\hat{\tau}}^1 \partial_{\beta_2} f'\boldsymbol{\eta}(s)ds\right] \\ b &= \partial_{\beta_2} f'(\bar{\boldsymbol{X}}'\bar{\boldsymbol{X}})^{-1}\bar{\boldsymbol{X}}'\bar{\boldsymbol{Z}}(\bar{\boldsymbol{Z}}'\boldsymbol{M}_{\bar{X}}\bar{\boldsymbol{Z}})^{-1}\bar{\boldsymbol{Z}}'\boldsymbol{M}_{\bar{X}}\boldsymbol{B}(1) \\ c &= \partial_\delta f'(\bar{\boldsymbol{Z}}'\boldsymbol{M}_{\bar{X}}\bar{\boldsymbol{Z}})^{-1}\bar{\boldsymbol{Z}}'\boldsymbol{M}_{\bar{X}}\boldsymbol{B}(1) \\ d &= \partial_{\beta_2} f'(\bar{\boldsymbol{X}}'\bar{\boldsymbol{X}})^{-1}\bar{\boldsymbol{X}}'\boldsymbol{B}(1) + \int_0^1 \partial_{\beta_2} f'\boldsymbol{\eta}(s)ds\end{aligned} \tag{3.48}$$

We have from (3.13) and (3.15) that $\sqrt{T}\partial_{\beta_2} f'(\hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2) \to a - b$, $\sqrt{T}\partial_{\beta_2} f'(\hat{\boldsymbol{\beta}}_F - \boldsymbol{\beta}_2) \to d - b$ and $\sqrt{T}\partial_\delta f'(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}) \to c$ Then

$$\begin{aligned}\Delta_1 &\to \mathrm{E}[a^2 + b^2 - 2ab + c^2 + 2ca - 2cb] \\ \Delta_2 &\to \mathrm{E}[d^2 + b^2 - 2db + c^2 + 2cd - 2cb]\end{aligned} \tag{3.49}$$

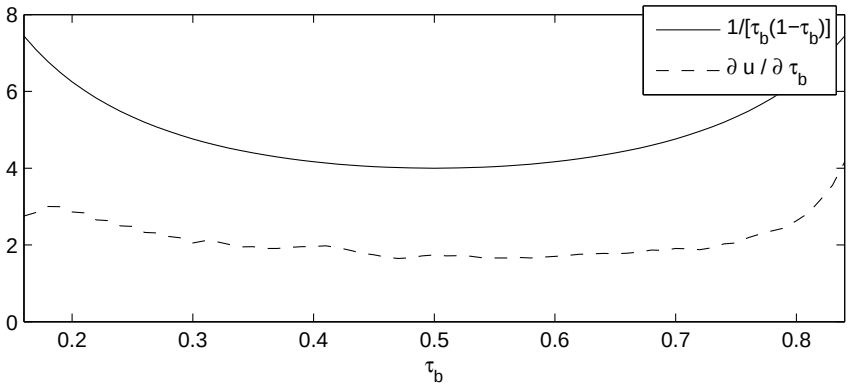
Now $\mathrm{E}[db] = \mathrm{E}[cd] = 0$ by the fact that $\mathrm{E}[\boldsymbol{B}(1)] = 0$, $\mathrm{E}[\boldsymbol{B}(1)\boldsymbol{B}(1)'] = \boldsymbol{I}$ and $\boldsymbol{M}_{\bar{\boldsymbol{X}}}\bar{\boldsymbol{X}} = \boldsymbol{O}$. Furthermore, as the distribution of the test statistic is independent of the estimation of $\boldsymbol{\delta}$ (Andrews, 1993), we can regard $\hat{\tau}$ as independent of $\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}$. This yields $\mathrm{E}[(a-d)c] = 0$. Concluding, we have

$$\Delta_1 - \Delta_2 = \mathrm{E}[a^2 - d^2] \tag{3.50}$$

3.A.2 Verifying condition (3.36)

In order to very that (3.36) holds, that is, that the condition for weak optimality, $\partial u(\tau_b)/\partial\tau_b < 1/[\tau_b(1-\tau_b)]$, holds. Observe that, in Figure 3.10, the dashed line, which depicts the derivative of the critical values for $\alpha = 0.05$ as a function of the break date τ_b and is obtained via simulation, is clearly below the solid line, which depicts the upper bound $[\tau_b(1-\tau_b)]^{-1}$.

Figure 3.10: Dependence of the critical values on the break date

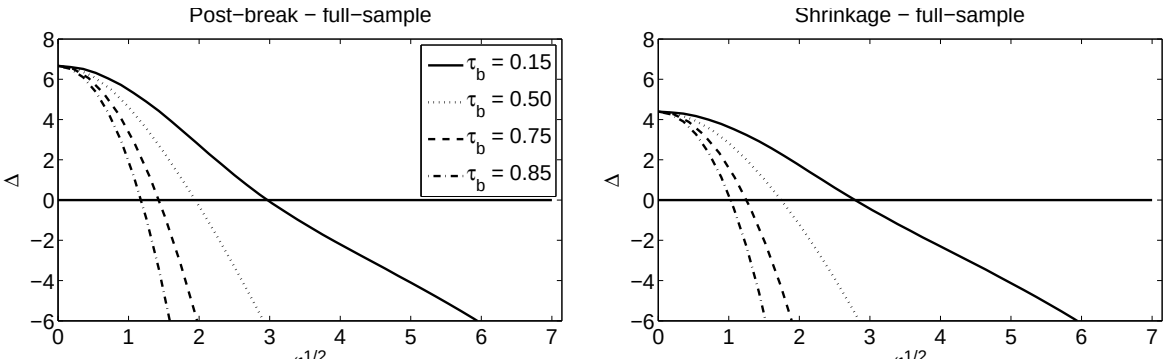


Note: The dashed line depicts the derivative of the critical values for $\alpha = 0.05$ as a function of the break date τ_b. The solid line depicting the upper bound $[\tau_b(1-\tau_b)]^{-1}$.

3.A.3 Uniqueness of the break size that yields equal forecast accuracy

In order to ensure the uniqueness of the break size that leads to equal forecast accuracy, we evaluate Δ in (3.19) and Δ_s in (3.43) numerically using the simulation set-up described in Section 3.5. The results in Figure 3.11 show that the break size that leads to equal forecast accuracy is, in fact, unique.

Figure 3.11: Difference in MSFE between the post-break forecast and full-sample forecast



Note: The left panel shows the difference in the asymptotic MSFE between the post-break forecast and the full-sample forecast as a function of the standardized break size $\zeta^{1/2}$ in (3.19) for $\tau_b = \{0.15, 0.50, 0.75, 0.85\}$. The right panel shows the difference in MSFE between the shrinkage forecast and the full-sample forecast in (3.43).

3.A.4 Derivation of equation (3.39)

We start by noting that (3.39)

$$\begin{aligned}\mathrm{E}\left[T\left(\hat{y}_{T+1}^{S}-\boldsymbol{x}_{T+1}'\boldsymbol{\beta}_2\right)^2\right] &= \mathrm{E}\left[T\left(\omega\boldsymbol{x}_{T+1}'\hat{\boldsymbol{\beta}}_1+(1-\omega)\boldsymbol{x}_{T+1}'\hat{\boldsymbol{\beta}}_2-\boldsymbol{x}_{T+1}'\boldsymbol{\beta}_2\right)^2\right] \\ &= \omega^2\mathrm{E}\left[T\left(\boldsymbol{x}_{T+1}'(\hat{\boldsymbol{\beta}}_1-\hat{\boldsymbol{\beta}}_2)\right)^2\right]+\frac{1}{\tau_b}\boldsymbol{x}_{T+1}'\boldsymbol{V}\boldsymbol{x}_{T+1} \\ &\quad +2\omega\boldsymbol{x}_{T+1}'\mathrm{E}\left[T\left(\hat{\boldsymbol{\beta}}_1-\hat{\boldsymbol{\beta}}_2\right)\left(\hat{\boldsymbol{\beta}}_2-\boldsymbol{\beta}_2\right)\right]\boldsymbol{x}_{T+1}\end{aligned} \tag{3.51}$$

We analyze the first and third term of the second equality separately. Using a bias-variance decomposition, the expectation in the first term can be calculated as

$$\begin{aligned}\mathrm{E}\left[T\left(\boldsymbol{x}_{T+1}'(\hat{\boldsymbol{\beta}}_1-\hat{\boldsymbol{\beta}}_2)\right)^2\right] &= \mathrm{E}\left[T\left(\boldsymbol{x}_{T+1}'(\hat{\boldsymbol{\beta}}_1-\hat{\boldsymbol{\beta}}_2)\right)\right]^2+T\mathrm{Var}\left[\boldsymbol{x}_{T+1}'(\hat{\boldsymbol{\beta}}_1-\hat{\boldsymbol{\beta}}_2)\right] \\ &= T\left(\boldsymbol{x}_{T+1}'\left(\boldsymbol{\beta}_1-\boldsymbol{\beta}_2\right)\right)^2+\boldsymbol{x}_{T+1}'\left(\frac{1}{\tau_b}+\frac{1}{1-\tau_b}\right)\boldsymbol{V}\boldsymbol{x}_{T+1}\end{aligned} \tag{3.52}$$

using that $\mathrm{Cov}(\hat{\boldsymbol{\beta}}_1,\hat{\boldsymbol{\beta}}_2)=0$. The term linear in ω is given by

$$\begin{aligned}\boldsymbol{x}_{T+1}'\mathrm{E}\left[T\left(\hat{\boldsymbol{\beta}}_1-\hat{\boldsymbol{\beta}}_2\right)\left(\hat{\boldsymbol{\beta}}_2-\boldsymbol{\beta}_2\right)\right]\boldsymbol{x}_{T+1} &= -\boldsymbol{x}_{T+1}'\mathrm{E}\left[T\left(\boldsymbol{\beta}_1-\boldsymbol{\beta}_2\right)\boldsymbol{\beta}_2'\right]\boldsymbol{x}_{T+1} \\ &\quad +\boldsymbol{x}_{T+1}'\mathrm{E}\left[T\hat{\boldsymbol{\beta}}_1\hat{\boldsymbol{\beta}}_2'-\hat{\boldsymbol{\beta}}_2\hat{\boldsymbol{\beta}}_2'\right]\boldsymbol{x}_{T+1} \\ &= -\frac{1}{1-\tau_b}\boldsymbol{x}_{T+1}'\boldsymbol{V}\boldsymbol{x}_{T+1}\end{aligned} \tag{3.53}$$

3.B Tables with critical values

Tables 3.7–3.8 contain critical values when the break is in the range $\tau_b = 0.15$ to 0.85, where Table 3.7 considers post-break sample and full sample based forecasts and Table 3.8 considers shrinkage forecast and full sample based forecasts. Tables 3.9–3.10 contain the critical values when the break can be in the range $\tau_b = 0.05$ to 0.95 for the same comparisons.

Table 3.7: Post-break versus full sample: critical values and size

τ_b	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85
$\zeta^{1/2}$	2.99	2.73	2.55	2.41	2.28	2.17	2.06	1.95	1.84	1.75	1.64	1.54	1.43	1.31	1.18
Wald test statistic (3.22)															
0.10	20.44	19.16	17.99	17.05	16.22	15.49	14.79	14.13	13.49	12.91	12.30	11.68	11.04	10.32	9.36
0.05	23.71	22.29	20.99	19.95	19.04	18.24	17.46	16.74	16.03	15.38	14.71	14.02	13.30	12.48	11.37
0.01	30.54	28.84	27.29	26.07	25.00	24.06	23.15	22.29	21.46	20.70	19.89	19.08	18.22	17.23	15.82
0.10	0.13	0.13	0.12	0.12	0.12	0.12	0.11	0.11	0.11	0.10	0.10	0.09	0.09	0.08	0.06
0.05	0.07	0.07	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.05	0.05	0.04	0.04	0.03
0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
S test statistic (3.38)															
0.10	1.78	1.82	1.84	1.86	1.87	1.88	1.89	1.89	1.88	1.87	1.86	1.83	1.80	1.73	1.59
0.05	2.12	2.16	2.18	2.20	2.21	2.22	2.23	2.23	2.22	2.22	2.20	2.18	2.14	2.08	1.94
0.01	2.76	2.79	2.81	2.83	2.85	2.86	2.86	2.87	2.86	2.86	2.85	2.83	2.80	2.74	2.60
0.10	0.10	0.10	0.10	0.10	0.10	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.10	0.08
0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.06	0.06	0.06	0.06	0.06	0.05	0.05	0.05	0.04
0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01

Note: The table reports critical values and size for the W and S test statistics that test the null hypothesis of equal MSFE of the post-break and full sample forecasts. For additional information, see the footnote of Table 3.1.

Table 3.8: Shrinkage versus full sample: critical values and size

τ_b	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85
$\zeta^{1/2}$	2.81	2.54	2.36	2.21	2.08	1.97	1.85	1.75	1.64	1.55	1.45	1.35	1.25	1.15	1.03
Wald test statistic (3.22)															
0.10	19.01	17.78	16.63	15.71	14.92	14.22	13.56	12.95	12.34	11.81	11.28	10.74	10.19	9.58	8.82
0.05	22.15	20.78	19.51	18.48	17.60	16.84	16.10	15.43	14.76	14.16	13.57	12.97	12.34	11.64	10.74
0.01	28.74	27.08	25.57	24.35	23.30	22.40	21.53	20.74	19.95	19.23	18.53	17.81	17.03	16.18	15.02
0.10	0.14	0.13	0.13	0.12	0.12	0.12	0.11	0.11	0.11	0.10	0.10	0.09	0.08	0.07	0.06
0.05	0.07	0.07	0.07	0.06	0.06	0.06	0.06	0.06	0.06	0.05	0.05	0.05	0.04	0.04	0.03
0.01	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
S test statistic (3.38)															
0.10	1.85	1.88	1.90	1.92	1.93	1.93	1.93	1.93	1.91	1.90	1.88	1.86	1.82	1.76	1.63
0.05	2.18	2.22	2.24	2.25	2.26	2.27	2.27	2.27	2.26	2.25	2.23	2.20	2.17	2.11	1.98
0.01	2.82	2.85	2.87	2.89	2.90	2.90	2.91	2.91	2.90	2.89	2.87	2.85	2.82	2.76	2.63
0.10	0.10	0.10	0.10	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.10	0.08
0.05	0.05	0.05	0.05	0.05	0.05	0.06	0.06	0.06	0.06	0.06	0.06	0.05	0.05	0.05	0.04
0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01

Note: The table reports critical values and size for the W and S test statistics that test the null hypothesis of equal MSFE of the shrinkage forecast (3.42) and the full sample forecast. For additional information, see the footnote of Table 3.2.

Table 3.9: Post-break versus full sample: critical values and size when searching $[0.05, 0.95]$

τ_b	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95
$\zeta^{1/2}$	4.17	3.69	3.44	3.25	3.10	2.97	2.85	2.73	2.63	2.53	2.43	2.32	2.21	2.10	1.97	1.84	1.68	1.48	1.19
Wald test statistic (3.22)																			
0.10	31.31	27.59	25.36	23.74	22.49	21.43	20.49	19.59	18.85	18.11	17.38	16.65	15.90	15.16	14.39	13.56	12.62	11.52	9.82
0.05	35.41	31.37	28.96	27.22	25.85	24.69	23.69	22.70	21.88	21.07	20.27	19.47	18.64	17.83	16.97	16.04	15.00	13.74	11.75
0.01	43.80	39.15	36.41	34.41	32.87	31.54	30.37	29.23	28.29	27.33	26.38	25.44	24.47	23.52	22.51	21.40	20.15	18.60	16.04
0.10	0.11	0.11	0.11	0.11	0.12	0.11	0.12	0.11	0.12	0.12	0.12	0.11	0.11	0.11	0.11	0.10	0.10	0.08	0.05
0.05	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.05	0.05	0.04	0.03
0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.00
S test statistic (3.38)																			
0.10	1.55	1.64	1.68	1.72	1.74	1.77	1.79	1.80	1.82	1.83	1.85	1.86	1.86	1.87	1.87	1.86	1.84	1.79	1.56
0.05	1.90	1.98	2.03	2.06	2.08	2.10	2.13	2.14	2.16	2.17	2.18	2.19	2.20	2.21	2.21	2.20	2.19	2.14	1.91
0.01	2.55	2.63	2.66	2.70	2.72	2.74	2.76	2.77	2.79	2.80	2.82	2.83	2.84	2.84	2.85	2.85	2.83	2.79	2.57
0.10	0.09	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.08
0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.06	0.06	0.06	0.06	0.04
0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01

Note: The table reports critical values and size for the W and S test statistics that test the null hypothesis of equal MSFE of the post-break and full sample forecasts. For additional information, see the footnote Table 3.1.

Table 3.10: Shrinkage versus full sample: critical values and size when searching $[0.05, 0.95]$

τ_b	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95
$\zeta^{1/2}$	4.04	3.56	3.30	3.10	2.95	2.81	2.68	2.56	2.46	2.35	2.24	2.13	2.02	1.90	1.78	1.65	1.49	1.31	1.05
Wald test statistic (3.22)																			
0.10	30.00	26.40	24.15	22.47	21.24	20.17	19.17	18.33	17.58	16.84	16.14	15.43	14.70	14.01	13.29	12.54	11.70	10.76	9.40
0.05	34.01	30.09	27.65	25.85	24.49	23.32	22.25	21.31	20.49	19.67	18.90	18.11	17.32	16.55	15.74	14.90	13.95	12.86	11.24
0.01	42.23	37.71	34.93	32.84	31.31	29.96	28.71	27.61	26.66	25.70	24.77	23.84	22.89	21.99	21.04	20.01	18.85	17.48	15.35
0.10	0.11	0.11	0.12	0.12	0.12	0.12	0.12	0.12	0.12	0.12	0.12	0.12	0.11	0.11	0.10	0.10	0.09	0.08	0.05
0.05	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.05	0.05	0.05	0.04	0.02
0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.00
S test statistic (3.38)																			
0.10	1.57	1.67	1.73	1.76	1.79	1.82	1.84	1.85	1.87	1.88	1.90	1.90	1.91	1.91	1.91	1.89	1.87	1.81	1.59
0.05	1.92	2.02	2.07	2.10	2.13	2.16	2.17	2.19	2.21	2.22	2.23	2.24	2.24	2.25	2.24	2.24	2.21	2.16	1.95
0.01	2.57	2.66	2.70	2.74	2.77	2.79	2.81	2.82	2.84	2.85	2.86	2.87	2.88	2.88	2.88	2.88	2.85	2.81	2.60
0.10	0.09	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.08
0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.06	0.06	0.06	0.06	0.06	0.06	0.04
0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01

Note: The table reports critical values and size for the W and S test statistics that test the null hypothesis of equal MSFE of the shrinkage forecast (3.42) and the full sample forecast. For additional information, see the footnote of Table 3.2.

Chapter 4

Controlled shrinkage and variable selection

4.1 Introduction and motivation

In the classical linear regression model, a response variable y_i for $i = 1, \dots n$ satisfies the following data generating process (DGP)

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i, \qquad \varepsilon_i \sim N(0, \sigma^2) \tag{4.1}$$

with $\boldsymbol{x}_i$ a $k \times 1$ vector of predictors and $\boldsymbol{\beta}$ a $k \times 1$ vector of coefficients. Biased shrinkage estimators for $\boldsymbol{\beta}$ are frequently considered as an alternative to the unbiased ordinary least squares (OLS) estimator

$$\hat{\boldsymbol{\beta}}^{OLS} = \left(\sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i'\right)\left(\sum_{i=1}^{n} \boldsymbol{x}_i y_i\right) \tag{4.2}$$

to increase prediction accuracy, and to achieve a more parsimonious model through variable selection. We distinguish two classes of shrinkage estimators. *Group shrinkage estimators* apply a common shrinkage factor to all coefficients, as for example the estimator by James and Stein (1961). Estimators in this class can dominate the OLS estimator in terms of estimation risk, but their interpretation is limited by the fact that no variable selection is performed. This contrasts with *predictor-specific shrinkage estimators* such as the lasso developed by Tibshirani (1996), the elastic net by Zou and Hastie (2005) and the adaptive lasso by Zou (2006). These estimators offer additional interpretation over the OLS estimator by allowing for variable selection. However, sharp risk bounds given by Donoho and Johnstone (1994) show that they do not dominate the risk of the OLS estimator. This paper introduces an estimator, Ctrl-shrink, that controls the fraction of repeated experiments in which the oracle shrinkage factor is exceeded, and contributes to both aforementioned classes.

In the following, we refer to the fraction of repeated experiments as the rate. For predictor-specific estimators, it is usually unknown at which rate variables are deleted that are part of the DGP. This can be harmful when variable selection is used as an exploratory tool to decide which variables need further investigation. Ctrl-shrink naturally controls the rate at which a variable is erroneously deleted as follows. For every coefficient in (4.1), an oracle shrinkage factor can be derived which depends only on the unknown (marginal) signal-to-noise ratio of the coefficients. By using the one-sided confidence level of the signal-to-noise ratio to calculate the shrinkage factor, the rate of exceeding the oracle factor is equal to a prespeficied level α. Ctrl-shrink then necessarily controls the rate of erroneously deleting a variable at the same level α. The estimator is shown to be an asymptotic oracle procedure as defined by Fan and Li (2001) and Zou (2006).

In addition, Ctrl-shrink can be used as a group shrinkage estimator when the shrinkage factor is based on the overall signal-to-noise ratio. This estimator dominates the risk of the OLS estimator when $k \geqslant 4$ and the rate α at which the oracle shrinkage factor is exceeded, is chosen to be smaller than or equal to 0.5. Refinements can be made to ensure that the dominance holds for $k \geqslant 3$.

A simulation exercise, using the lasso and the positive-part James-Stein estimator as benchmarks, shows that the risk of Ctrl-shrink is competitive. When used as a predictor-specific estimator, the quality of the variable selection is substantially improved. The rate at which variables are erroneously deleted is by definition bounded at a prespecified level α. For the lasso estimator this rate can be as high as 0.90 when the signal-to-noise ratio is small. We find that the rate at which the method correctly identifies the nonzero coefficients is consistently higher compared with the lasso when $\alpha = 0.10$. The performance of the estimator is robust to different realizations of the predictors $\boldsymbol{x}_i$. As a group shrinkage estimator, Ctrl-shrink offers a minor improvement over the James-Stein estimator.

We present empirical evidence for the usefulness of the developed estimator using prostate cancer data from Stamey et al. (1989), previously considered by Tibshirani (1996) to illustrate the lasso. We randomize over the split between the training set and the prediction set to avoid dependence on the choice of a particular prediction set. The Ctrl-shrink estimator improves over the OLS estimator, while the lasso performs worse by a wide margin. In concurrence with the simulation exercise, the risk is less sensitive to the location of the split point compared to the lasso.

The outline of this chapter is as follows. The Ctrl-shrink estimator is defined in Section 4.2. Predictor-specific and group shrinkage estimation are discussed in Section 4.3 and Section 4.4. Section 4.5 provides a simulation exercise to analyze finite-sample risk and variable selection. Section 4.6 presents an application to prostate cancer data. Conclusions and directions for further research are discussed in Section 4.7.

4.2 Ctrl-shrink

4.2.1 Overshrinkage and overselection

Predictor-specific shrinkage estimators apply a shrinkage factor ω_i to each coefficient of the OLS estimator $\tilde{\beta}_i = (1-\omega_i)\hat{\beta}_i$, where $\hat{\beta}_i^{OLS} \sim N(\beta_i, \sigma^2 v_i)$, $v_i = [(\boldsymbol{X}'\boldsymbol{X})^{-1}]_{ii}$ and $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]'$ a $n \times k$ matrix of predictors. If ω_i and the predictors are non-stochastic, the estimation risk is given by

$$\begin{aligned} R(\eta_i, \omega_i) &= \mathrm{E}\left\{\sigma^{-2} v_i^{-1}\left[(1-\omega_i)\hat{\beta}_i^{OLS} - \beta_i\right]^2\right\} \\ &= 1 + \omega_i^2(\eta_i + 1) - 2\omega_i \end{aligned} \tag{4.3}$$

where $\eta_i = \frac{\beta_i^2}{\sigma^2 v_i}$ is the signal-to-noise ratio of individual coefficients. The risk is minimized by the well-known oracle factor $\omega_i^o = \frac{1}{\eta_i+1}$. The maximum shrinkage factor, such that for $\omega_i < \omega_i^m$ the risk is reduced, is $\omega_i^m = 2\omega_i^o$.

Instead of applying a different shrinkage factor to each coefficient, group shrinkage estimators apply a common factor ω to all coefficients of the OLS estimator $\tilde{\boldsymbol{\beta}} = (1-\omega)\,\hat{\boldsymbol{\beta}}^{OLS}$, where $\hat{\boldsymbol{\beta}}^{OLS} \sim N(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1})$. If ω and the matrix of predictors $\boldsymbol{X}$ are non-stochastic, the risk is given by

$$\begin{aligned} R(\eta, \omega, k) &= \mathrm{E}\left\{\sigma^{-2}\left[(1-\omega)\hat{\boldsymbol{\beta}}^{OLS} - \boldsymbol{\beta}\right]' \boldsymbol{X}'\boldsymbol{X}\left[(1-\omega)\hat{\boldsymbol{\beta}}^{OLS} - \boldsymbol{\beta}\right]\right\} \\ &= k + \omega^2(\eta + k) - 2\omega \end{aligned} \tag{4.4}$$

where $\eta = \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}/\sigma^2$ denotes the signal-to-noise ratio. Equation (4.4) is minimized by the oracle weight $\omega^o = \frac{k}{\eta+k}$. The maximum shrinkage factor ω^m is given by $\omega^m = 2\omega^o$.

Equations (4.3) and (4.4) show that the risk is increased beyond that of the OLS estimator when the shrinkage factor is too large. We define *overshrinkage* as estimating $\hat{\omega}_i(\hat{\eta}_i)$ such that $\hat{\omega}_i(\hat{\eta}_i) > \omega_i$, where $\omega_i = \omega_i^o$ or $\omega_i = \omega_i^m$. For group shrinkage estimators, the subscripts i are dropped. *Overselection*, a special case of overshrinkage, occurs when we estimate $\hat{\omega}_i = 1$ while $\omega_i < 1$. Ctrl-shrink is defined such that the rate at which overshrinkage occurs is equal to α. The overselection rate is then smaller or equal to α.

From a risk perspective, one can be inclined to set $\omega_i = \omega_i^m$ in order to bound the rate at which the risk increases over the OLS risk. Both for group shrinkage as for predictor-specific shrinkage there are good reasons not to do so. For group shrinkage, we show that this choice is not sufficient to dominate the risk of the OLS estimator. For predictor-specific shrinkage, the choice between $\omega_i = \omega_i^o$ and $\omega_i = \omega_i^m$ determines the interpretation of overselection. When $\omega_i = \omega_i^o$, overselection implies that a coefficient is set to zero for which $\eta_i \neq 0$. Alternatively, when we choose $\omega_i = \omega_i^m$, overselection occurs when a coefficient is set

to zero while $\eta_i > 1$. Although the latter is sensible from a risk perspective, the former is preferred from a variable selection perspective as it corresponds to standard hypothesis testing procedures.

4.2.2 Definition of Ctrl-shrink

The shrinkage factor used by Ctrl-shrink is defined by setting the probability of overshrinkage equal to a pre-specified level α

$$\Pr\left[\hat{\omega}(\hat{\eta}) > \omega\right] = \Pr\left[\hat{\omega}(\hat{\eta}) > \frac{bk}{\eta + k}\right] = \alpha \tag{4.5}$$

where the dependence of $\hat{\omega}$ on $\hat{\eta}$ is made explicit. Setting $b = 1$ corresponds to $\omega = \omega^o$ and $b = 2$ corresponds to $\omega = \omega^m$. Predictor-specific shrinkage uses (4.5) with $k = 1$. We can rewrite (4.5) as

$$\Pr\left[\eta > k\left(\frac{b}{\hat{\omega}(\hat{\eta})} - 1\right)\right] = \Pr\left[\eta > \nu(\hat{\eta})\right] = \alpha \tag{4.6}$$

This is simply the definition of a one-sided confidence interval. The distribution of the signal-to-noise ratio determines $\nu(\hat{\eta})$ and subsequently $\hat{\omega}(\hat{\eta})$. For individual coefficients, the signal-to-noise ratio is distributed under known variance as

$$\hat{\eta}_i = \frac{\left(\hat{\beta}_i^{OLS}\right)^2}{\sigma^2 v_i} \sim \chi^2\left(1, \frac{\beta_i^2}{\sigma^2 v_i}\right) \tag{4.7}$$

where $\chi^2(k, \eta)$ is the non-central chi-squared distribution with k degrees of freedom and non-centrality parameter η. When the variance is unknown, it can be estimated using

$$\hat{\sigma}^2 = \frac{1}{n-k}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}^{OLS})'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}^{OLS}) \tag{4.8}$$

and the distribution of the signal-to-noise ratio is

$$\hat{\eta}_i = \frac{\left(\hat{\beta}_i^{OLS}\right)^2}{\hat{\sigma}^2 v_i} \sim F\left(1, n-k, \frac{\beta_i^2}{\sigma^2 v_i}\right) \tag{4.9}$$

where $F(k_1, k_2, \eta)$ denotes the non-central F distribution with degrees of freedom k_1, k_2 and non-centrality parameter η.

For group shrinkage, the overall signal-to-noise ratio used to construct a common shrinkage factor is distributed under known variance as

$$\hat{\eta} = \frac{\hat{\boldsymbol{\beta}}^{OLS}\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}}^{OLS}}{\sigma^2} \sim \chi^2\left(k, \frac{\boldsymbol{\beta}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}}{\sigma^2}\right) \tag{4.10}$$

When the variance is unknown, substituting (4.8) gives

$$\hat{\eta} = \frac{1}{k}\frac{\hat{\boldsymbol{\beta}}^{OLS}\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}}^{OLS}}{\hat{\sigma}^2} \sim F\left(k, n-k, \frac{\boldsymbol{\beta}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}}{\sigma^2}\right) \tag{4.11}$$

Following the Neyman construction of a confidence interval, the end-point $\nu(\hat{\eta})$ in (4.6) is determined by taking it as the non-centrality parameter for which $\hat{\eta}$ equals the critical value that corresponds to confidence level α. Under known variance, we set $c_{\alpha,\nu} = \hat{\eta}$ and find ν such that

$$\int_0^{c_{\alpha,\nu}} f(q)dq = \alpha, \qquad q \sim \chi^2(k, \nu) \tag{4.12}$$

where $f(q)$ denotes the pdf of q. Equation (4.6) will hold regardless of the value of η and subsequently (4.5) is invariant with respect to η. Note that if $\hat{\eta} < c_{\alpha,0}$, (4.12) does not have a solution. In this case we set $\nu = 0$, which yields $\hat{\omega} = b$.

If the variance is unknown, (4.12) is adjusted by replacing $\chi^2(k, \nu)$ by $F(k_1, n-k, \nu)$ where $k_1 = 1$ for predictor-specific shrinkage and $k_1 = k$ for group shrinkage.

4.2.3 Properties of the Ctrl-Shrink estimator

Properties of the estimator can be derived assuming σ^2 known. Using the generalized Marcum's Q function, (4.12) can be written as

$$Q_{k/2}(\sqrt{\nu(\hat{\eta})}, \sqrt{\hat{\eta}}) = 1 - \alpha, \qquad \nu(\hat{\eta}) = k\left(\frac{b}{\hat{\omega}(\hat{\eta})} - 1\right) \tag{4.13}$$

To determine how $\hat{\omega}$ depends on $\hat{\eta}$, we take the derivative of (4.13) with respect to $\sqrt{\hat{\eta}}$

$$\begin{aligned}
\frac{\partial Q_{k/2}\left(\sqrt{\nu(\hat{\eta})}, \sqrt{\hat{\eta}}\right)}{\partial\sqrt{\hat{\eta}}} &= \frac{\partial\sqrt{\nu(\hat{\eta})}}{\partial\sqrt{\hat{\eta}}}\frac{\partial Q_{k/2}\left(\sqrt{\nu(\hat{\eta})}, \sqrt{\hat{\eta}}\right)}{\partial\sqrt{\nu(\hat{\eta})}} + \frac{\partial Q_{k/2}\left(\sqrt{\nu(\hat{\eta})}, \sqrt{\hat{\eta}}\right)}{\partial\sqrt{\hat{\eta}}} \\
&= \sqrt{\hat{\eta}}e^{-\frac{\nu(\hat{\eta})+\hat{\eta}}{2}}\left(\frac{\hat{\eta}}{\nu(\hat{\eta})}\right)^{\frac{k}{4}-\frac{1}{2}} \\
&\quad \cdot \left[\frac{\partial\sqrt{\nu(\hat{\eta})}}{\partial\sqrt{\hat{\eta}}}I_{k/2}\left(\sqrt{\nu(\hat{\eta})\hat{\eta}}\right) - I_{k/2-1}\left(\sqrt{\nu(\hat{\eta})\hat{\eta}}\right)\right] \\
&= 0
\end{aligned} \tag{4.14}$$

with $I_{k/2}(\cdot)$ the modified Bessel function of the first kind. The last line follows from the fact that the right-hand side of (4.13) is independent of $\hat{\eta}$. Equation (4.14) holds when

$$\frac{\partial\sqrt{\nu(\hat{\eta})}}{\partial\sqrt{\hat{\eta}}} = \frac{I_{k/2-1}\left(\sqrt{\nu(\hat{\eta})\hat{\eta}}\right)}{I_{k/2}\left(\sqrt{\nu(\hat{\eta})\hat{\eta}}\right)} \tag{4.15}$$

When $\hat{\eta}$ is large, the Bessel function can be approximated by

$$I_{k/2}\left(\sqrt{\nu(\hat{\eta})\hat{\eta}}\right) = \frac{\exp\left(\sqrt{\nu(\hat{\eta})\hat{\eta}}\right)}{\sqrt{2\pi\sqrt{\nu(\hat{\eta})\hat{\eta}}}} + \dots$$

such that the ratio in (4.15) equals one. Then

$$\frac{\partial\nu(\hat{\eta})}{\partial\hat{\eta}} = \sqrt{\frac{\nu(\hat{\eta})}{\hat{\eta}}}$$

which is solved by

$$\nu(\hat{\eta}) \propto \hat{\eta}$$

Therefore, for large values of $\hat{\eta}$

$$\hat{\omega} \propto \hat{\eta}^{-1} \tag{4.16}$$

The results by Magnus and Durbin (1996) show that this property guarantees that the risk of this estimator will converge to the risk of the OLS estimator when $\hat{\eta} \to \infty$.

4.3 Predictor-specific shrinkage

4.3.1 Theoretical properties

To increase the interpretability of the estimates, shrinkage estimators can be used to decide which variables are part of the underlying DGP. In the terminology of Zou (2006), a fitting procedure P is an oracle procedure if $\hat{\boldsymbol{\beta}}(P)$ (asymptotically) identifies the right subset model and converges as $\sqrt{n}\left[\hat{\boldsymbol{\beta}}(P) - \boldsymbol{\beta}\right] \to_d N(\mathbf{0}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma}$ the asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$. The Ctrl-shrink factor $\hat{\omega} = \mathcal{O}(\hat{\eta}^{-1})$ as $\hat{\eta} \to \infty$. The signal-to-noise ratio $\hat{\eta} = \mathcal{O}(n)$ as $n \to \infty$ and therefore the shrinkage factor scales as $\hat{\omega} = \mathcal{O}(n^{-1})$ as $n \to \infty$. Since the OLS estimator is $\sqrt{n}$-consistent, Ctrl-shrink converges to $\boldsymbol{\beta}$ as $n \to \infty$ and satisfies both requirements by Zou (2006).

4.3.2 Computational properties

Computational efficiency is essential for large scale applications. Since the Ctrl-shrink estimator is based on the marginal distribution of each parameter and does not require one to solve a penalized least squares problem, it is computationally inexpensive. The following approximation to the shrinkage factor can be used as the initial value in a line search algorithm for $\hat{\omega}(\hat{\eta})$

Consider shrinking a single predictor distributed as $\hat{\beta}^{OLS} \sim N(\beta, \sigma^2 v)$. The shrinkage estimator can be defined equivalent to (4.12) by using the CDF of the standard normal distribution

$$\Phi\left(\sqrt{\hat{\eta}} - \sqrt{b/\hat{\omega} - 1}\right) - \Phi\left(-\sqrt{\hat{\eta}} - \sqrt{b/\hat{\omega} - 1}\right) = \alpha$$

where α is the chosen significance level. When $\sqrt{\hat{\eta}}$ large, the second term on the left hand side exponentially tends to 0 as $\hat{\eta}$ increases. Neglecting this term gives a closed form expression for the shrinkage factor

$$\hat{\omega} = \frac{b}{[\sqrt{\hat{\eta}} - \Phi^{-1}(\alpha)]^2 + 1} \tag{4.17}$$

which decays as $\mathcal{O}(\hat{\eta}^{-1})$ when $\hat{\eta} \to \infty$ as found in (4.16).

An algorithm to determine the shrinkage constant is readily established. We need to solve $g(\hat{\omega}) - \alpha = 0$ where

$$g(\hat{\omega}) = F\left[\hat{\eta}; 1, n-k, \left(\frac{b}{\hat{\omega}} - 1\right)\right]$$

and

$$\frac{\partial g(\hat{\omega})}{\partial \hat{\omega}} = \frac{1}{2}\left\{g(\hat{\omega}) - F\left[\frac{3}{2}\hat{\eta}; 3, n-k, \left(\frac{b}{\hat{\omega}} - 1\right)\right]\right\}\frac{b}{\hat{\omega}^2} \tag{4.18}$$

Using equation (4.17) as initial value yields an efficient algorithm.

4.3.3 Comparison to existing alternatives

To visualize the shrinkage factor as a function of $\hat{\eta}$ and α, we compare it with the lasso, the pretest estimator and non-negative garotte for shrinking a single parameter. We take $b = 1$ and assume $\sigma^2 = 1$ and known, so that $v = 1$ and $\hat{\beta} = \sqrt{\hat{\eta}}$. All estimators require a single tuning parameter γ to be specified, which is set such that the overselection rates of the estimators are equal.

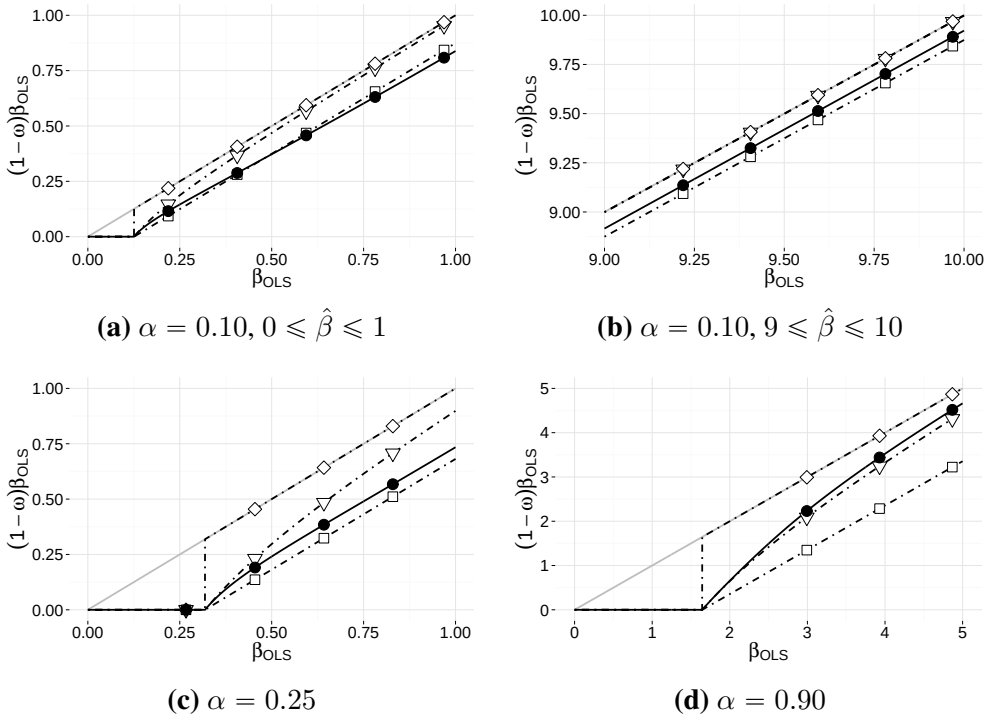
Pretest estimator The pretest estimator is given by

$$\hat{\beta}^{PT} = \begin{cases} 0 & \text{if } |\hat{\beta}^{OLS}| \leqslant \gamma \\ \hat{\beta}^{OLS} & \text{if } |\hat{\beta}^{OLS}| > \gamma \end{cases}$$

Lasso When $k = 1$, the lasso estimator is given as

$$\hat{\beta}^{L} = \begin{cases} 0 & \text{if } |\hat{\beta}^{OLS}| \leqslant \gamma \\ \left(1 - \frac{\gamma}{|\hat{\beta}^{OLS}|}\right)\hat{\beta} & \text{if } |\hat{\beta}^{OLS}| > \gamma \end{cases}$$

Figure 4.1: Univariate regression: shrinkage estimates as a function of $\hat{\beta}_{OLS}$



(a) $\alpha = 0.10, 0 \leqslant \hat{\beta} \leqslant 1$ **(b)** $\alpha = 0.10, 9 \leqslant \hat{\beta} \leqslant 10$

(c) $\alpha = 0.25$ **(d)** $\alpha = 0.90$

Notes: Shrinkage estimator of β as a function of the unbiased estimate $\hat{\beta}$ (—, gray) for Ctrl-shrink (—, ●) as compared with the pretest estimator (–·–, ◇), the lasso (–·–, □), the non-negative Garotte (–·–, ▽).

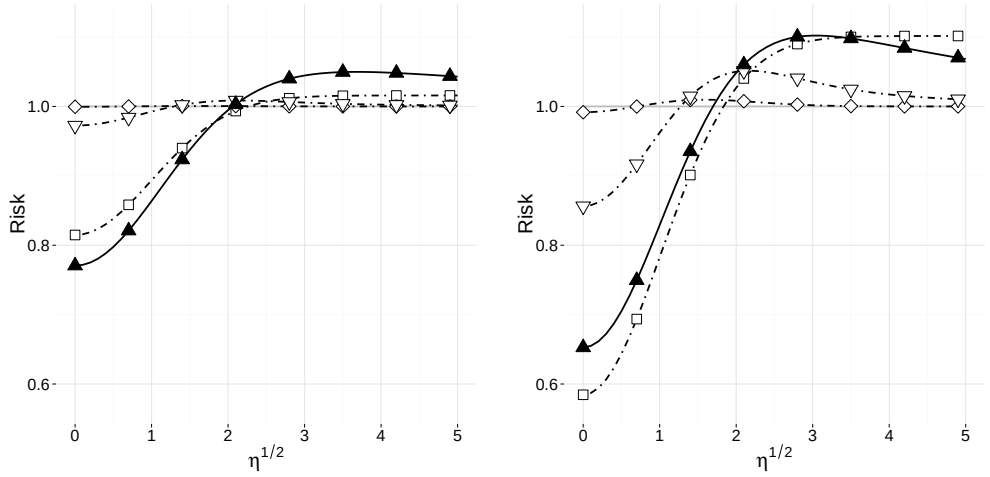
where the second expression shows that in the univariate setting, it is identical to the Burr estimator, see Burr (1942) and Burr and Cislak (1968) and an extensive discussion by Magnus (2002).

Non-negative garotte The non-negative garotte estimator of Breiman (1995) decreases as $\mathcal{O}(\hat{\eta}^{-1})$ offering the same decay rate as Ctrl-shrink when $\hat{\eta} \to \infty$. The estimator is defined as

$$\hat{\beta}^G = \begin{cases} 0 & \text{if } \left(\hat{\beta}^{OLS}\right)^2 \leqslant \gamma^2 \\ \left(1 - \frac{\gamma^2}{\left(\hat{\beta}^{OLS}\right)^2}\right) \hat{\beta}^{OLS} & \text{if } \left(\hat{\beta}^{OLS}\right)^2 > \gamma^2 \end{cases}$$

Figure 4.1 shows the different shrinkage estimators compared with the unbiased estimator. We consider $\alpha \in \{0.10, 0.25, 0.90\}$. The choice of α is important for the behavior of Ctrl-shrink relative to the alternatives. For $\alpha = 0.10$, the Ctrl-shrink estimator fluctuates around the lasso estimator for small values of $\hat{\eta}$. However, when $\hat{\eta}$ increases the quadratic decay rate of the Ctrl-shrink estimator results in smaller shrinkage compared to the lasso. For $\alpha = 0.25$ Ctrl-shrink is in between the lasso and the garotte estimator, while for $\alpha = 0.90$ it converges to the OLS estimator quicker than both the lasso and the garotte estimator.

Figure 4.2: Univariate regression: risk



(a) $\alpha = 0.10$ **(b)** $\alpha = 0.25$

Notes: Risk as a function of the signal-to-noise ratio η for Ctrl-shrink (—, ●) as compared with the pretest estimator (–·–, ◇), the lasso (–·–, □), the non-negative Garotte (–·–, ▽).

The risk functions as a function of the signal-to-noise ratio for the estimators using $\alpha = 0.10$ are depicted in Figure 4.2a. The minimum risk of the Ctrl-shrink estimator is the lowest among these alternatives, while the maximum risk is higher compared with the lasso and Garotte. The pretest estimator is close to 1 and does not offer substantial risk improvements for any value of the signal-to-noise ratio. In Figure 4.2b, the risk functions are plotted when the rate of overshrinkage $\alpha = 0.25$. As expected, the minimum risk of all estimators decreases. This effect is stronger for the lasso estimator than for Ctrl-shrink, and in this case the minimum risk of the lasso is the lowest amongst the estimators. The maximum risk is roughly equal for Ctrl-shrink and the lasso estimator. However, as a consequence of the fact that for large $\hat{\eta}$ the lasso decays as $\mathcal{O}(\hat{\eta}^{-1/2})$, the risk of the lasso estimator does not converge to the risk of the unbiased estimator.

4.4 Group shrinkage estimation

If a researcher is primarily interested in achieving a strictly lower risk than that of the OLS estimator, group shrinkage estimators are preferred over predictor-specific shrinkage estimators. In this section we prove that when the Ctrl-shrink estimator is used for group shrinkage, it dominates the OLS estimator provided that $\alpha \leqslant 0.5$ and $b \leqslant 2\frac{k-2}{k}$. Using $b = 1$ implies that we can uniformly improve over the OLS estimator of $\boldsymbol{\beta}$ if we bound the overshrinkage rate, and then necessarily the rate at which the predictors are erroneously deleted, by $\alpha \leqslant 0.5$ and $k_2 \geqslant 4$.

Define

$$\phi(\hat{\eta}) = \hat{\eta} \cdot \hat{w}(\hat{\eta})$$

To improve upon the risk of the OLS estimator the following condition should be satisfied, Maruyama and Strawderman (2005),

$$\mathrm{E}\left\{ \frac{\phi(\hat{\eta})[2(k-2) - \phi(\hat{\eta})]}{\hat{\eta}} + 4\frac{\partial \phi(\hat{\eta})}{\partial \hat{\eta}} \right\} \geqslant 0 \tag{4.19}$$

which holds for example when $\frac{\partial \phi(\hat{\eta})}{\partial \hat{\eta}} \geqslant 0$ and $0 \leqslant \phi(\hat{\eta}) \leqslant 2(k-2)$

Theorem 1 *The shrinkage estimator defined through* (4.12) *or* (4.13) *satisfies* $\frac{\partial \phi(\hat{\eta})}{\partial \hat{\eta}} \geqslant 0$ *if* $\alpha \leqslant 0.5$

Proof: When $\hat{\eta} \leqslant c_{\alpha,0}$, we have $\hat{\omega} = b$. Then $\frac{\partial \phi(\hat{\eta})}{\partial \hat{\eta}} = b > 0$. For $\hat{\eta} > c_{\alpha,0}$, applying the chain rule to (4.4) and using $\hat{\omega}(\hat{\eta}) = \frac{bk}{\nu(\hat{\eta})+k}$ gives

$$\begin{aligned} \frac{\partial \phi(\hat{\eta})}{\partial \hat{\eta}} &= \hat{w} + \hat{\eta}\frac{\partial \hat{w}}{\partial \hat{\eta}} \\ &= \hat{w}\left[1 - \frac{1}{bk}\hat{w}\hat{\eta}\frac{\partial \nu(\hat{\eta})}{\partial \hat{\eta}} \right] \end{aligned} \tag{4.20}$$

The following inequality provides an upper bound on the weights

$$\frac{\partial \sqrt{\nu(\hat{\eta})}}{\partial \sqrt{\hat{\eta}}} = \frac{I_{k/2-1}\left(\sqrt{\nu(\hat{\eta})\hat{\eta}}\right)}{I_{k/2}\left(\sqrt{\nu(\hat{\eta})\hat{\eta}}\right)} < \frac{\sqrt{\nu(\hat{\eta})\hat{\eta}}}{\sqrt{\nu(\hat{\eta})\hat{\eta} + \frac{k^2}{4}} - \frac{k}{2}}$$

see Laforgia and Natalini (2010). Again using the chain rule we obtain

$$\hat{\eta}\frac{\partial \nu(\hat{\eta})}{\partial \hat{\eta}} \leqslant \frac{\nu(\hat{\eta})\hat{\eta}}{\sqrt{\nu(\hat{\eta})\hat{\eta} + \frac{k^2}{4}} - \frac{k}{2}} \tag{4.21}$$

Equation (4.20) shows that in order for $\frac{\partial \phi(\hat{\eta})}{\partial \hat{\eta}} \geqslant 0$ we need

$$bk - \hat{w}\hat{\eta}\frac{\partial \nu(\hat{\eta})}{\partial \hat{\eta}} \geqslant 0 \tag{4.22}$$

Substituting the upper bound from (4.21) and using that $\nu(\hat{\eta}) = k\left(\frac{b}{\hat{w}(\hat{\eta})} - 1\right)$ gives the cubic equation

$$\hat{\eta}\hat{w}^3 - 2b\left(\frac{k}{2} + \hat{\eta}\right)\hat{w}^2 + b^2\left(2k + \hat{\eta}\right)\hat{w} - b^3 k \leqslant 0 \tag{4.23}$$

This equation has two roots

$$r_1 = b, \quad r_2 = \frac{bk}{\hat{\eta}}$$

If $\hat{w} < \min(r_1, r_2)$ the inequality (4.23) holds so that a sufficient condition on $\hat{w}$ is

$$\hat{w} \leqslant \begin{cases} b & \text{if } \hat{\eta} \leqslant k \\ \frac{bk}{\hat{\eta}} & \text{if } \hat{\eta} > k \end{cases} \tag{4.24}$$

The first inequality in (4.24) is automatically satisfied by the definition of $\hat{\omega}$.

To derive a condition on α for which the second inequality holds, consider first the case where $\hat{\eta} > k$ and $k < c_{\alpha,0}$. This violates (4.24), since $\hat{\omega} = b$ by definition, but $bk/\hat{\eta} < b$. To exclude this possibility, it is sufficient to restrict $\alpha \leqslant 0.5$. The median-mean inequality for the non-central χ^2 distribution proved by Sen (1989) then implies $c_{\alpha,0} < k$ for $\alpha \leqslant 0.5$.

Now consider the case where $\hat{\eta} > k$ and $k > c_{\alpha,0}$. Define again $c_{\alpha,\nu} = \hat{\eta}$ and recall the definition of the Ctrl-shrink estimator

$$\Pr\left[X \leqslant c_{\alpha,\nu}\right] = \alpha, \quad X \sim \chi^2\left(k, \nu\right)$$

where $\hat{\omega}$ is calculated using

$$\nu = k\left(\frac{b}{\hat{\omega}} - 1\right) \tag{4.25}$$

Due to the monotonicity of the CDF it is sufficient to prove that using the bound for the weights given in (4.24) in (4.25), increases the CDF evaluated at $c_{\alpha,\nu}$. In other words, the condition on α is

$$\alpha \leqslant \Pr\left[\tilde{X} \leqslant c_{\alpha,\nu}\right], \quad \tilde{X} \sim \chi^2(k, c_{\alpha,\nu} - k) \tag{4.26}$$

Using again the median-mean inequality for the non-central χ^2 distribution, setting $\alpha \leqslant 0.5$ is sufficient to satisfy (4.26). ■

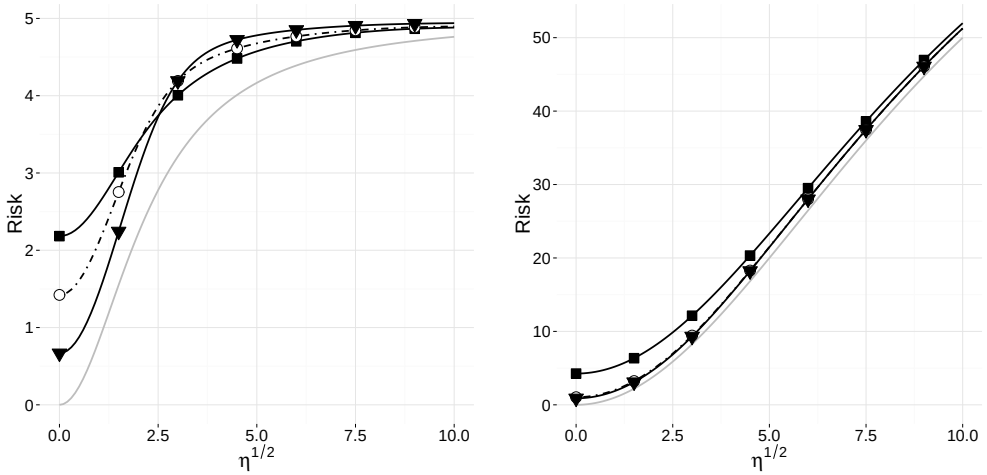
Theorem 2 *Given* $\alpha \leqslant 0.5$ *and* $b \leqslant \frac{2(k-2)}{k}$ *it holds that* $0 \leqslant \phi(\hat{\eta}) \leqslant 2(k-2)$

Proof: Appendix 4.A shows that the limit when $\hat{\eta} \to \infty$ of the weights is $\hat{w} = \frac{bk}{\hat{\eta}}$. In this limit $\phi = bk$ and since we have proven above that ϕ is a nondecreasing function of $\hat{\eta}$, it approaches this limit from below if $\alpha \leqslant 0.5$. Then, if $b \leqslant \frac{2(k-2)}{k}$ and hence, $bk \leqslant 2(k-2)$, Theorem 2 holds. ■

Note that if we choose $b = 1$, this means we improve over the unbiased estimator $\hat{\beta}$ when $k \geqslant 4$. Other choices of b such as $b = \frac{k-2}{2}$ will improve the risk over the unbiased estimator when $k \geqslant 3$ similar to the James-Stein estimator. Controlling the level of overshrinkage from an risk point of view, i.e. choosing $b = 2$, is not sufficient to dominate the OLS estimator.

The risk function under known variance is plotted for $k = 5$ in Figure 4.3a and for $k = 100$ in Figure 4.3b. We evaluated the risk for the Ctrl-shrink estimator using $\alpha \in \{0.10, 0.50\}$. For reference purposes, the risk using the oracle factor is also shown, as well as the James-Stein estimator, $\hat{\boldsymbol{\beta}}^{JS} = \max\left[0, 1 - \frac{k-2}{\hat{\eta}}\right]\hat{\boldsymbol{\beta}}^{OLS}$. In line with the proof above, all

Figure 4.3: Group shrinkage: risk



(a) $k = 5$ **(b)** $k = 100$

Notes: Risk as a function of the signal-to-noise ratio η for Ctrl-shrink with $\alpha = 0.10$ (—, ■), $\alpha = 0.50$ (—, ▼). As a benchmark, the risk of the positive part James-Stein estimator (–·–, ○) and the risk using the oracle shrinkage factor (—, gray) are shown.

estimators have a lower risk than the OLS estimator. Using $\alpha = 0.50$ shows a considerable risk reduction when the signal-to-noise ratio is small, while the increased risk for larger values of η is limited. On the other hand, using $\alpha = 0.10$ offers improvements for larger values of η. When $k = 100$ in Figure 4.3b the risk of the Ctrl-Shrink and James-Stein estimator approaches the oracle risk. The Ctrl-shrink estimator with $\alpha = 0.10$ yields a higher risk and can be considered too conservative from a risk perspective.

4.5 Simulations

4.5.1 Set-up

Recently, Hansen (2015) compared the risk of the James-Stein estimator and the lasso in a simulation exercise. We slightly adjust the simulation set-up presented there to be able to compare the variable selection quality as well as the risk of the estimators. We define accurate variable selection from the perspective of the DGP, such that the Ctrl-shrink estimator takes $b = 1$ in (4.6).

The data is generated using the linear model (4.1). The predictor matrix $\boldsymbol{X}$ is $n \times k$ with $k = 33$ and the sample size is set to $n = 50$. Results for $n = 200$ and a proportional increase in the parameters are similar and provided in Appendix 4.B. The first column of $\boldsymbol{X}$ is a vector of ones and the corresponding parameter is not subject to shrinkage. There are $m = k - 1 = 32$ remaining predictors of which $p = \{16, 32\}$ are non-zero representing a relatively sparse

and a dense model. We set $\beta_i = 1$ for $i = 1, \ldots, m$ and 0 otherwise. Two specifications are considered for the predictors. In the first specification the $\boldsymbol{x}_i$ are generated independently from a $N(0, 1)$ distribution. The predictors in the second specification are equicorrelated, meaning that they are generated as $\boldsymbol{x}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma})$ with $[\boldsymbol{\Sigma}]_{ll} = 1$ and $[\boldsymbol{\Sigma}_{lm}] = 0.5$ for $l \neq m$. This allows us to study the effect of multicollinearity. In this case we also consider $\beta_i = (-1)^i$, which lowers the signal-to-noise ratio. The predictor matrix $\boldsymbol{X}$ is normalized to have zero mean and unit variance.

The following estimates of $\boldsymbol{\beta}$ are considered with abbreviations in parenthesis: predictor-specific Ctrl-shrink with $\alpha = \{0.10, 0.25\}$ (r-CS$_{0.10}$ and r-CS$_{0.25}$), lasso with fivefold cross-validation using the *R* package `glmnet` (lasso), group Ctrl-shrink with $\alpha = 0.50$ (g-CS) and the James-Stein estimator (JS). The variance is estimated using the unbiased estimator (4.8). We report the overall mean squared error relative to the risk of the OLS estimator, the fraction of variables that is erroneously deleted, and the rate at which the method correctly identifies all nonzero coefficients. Based on the univariate simulations in the previous paragraphs, we vary $\frac{1}{p}\eta^{1/2} = \{0.25, 0.5, \ldots, 7.5\}$ using a 30-point grid for the variance of the error term. We consider 45 realizations[1] of the predictor matrix $\boldsymbol{X}$ and for each realization of $\boldsymbol{X}$ 1.000 data sets are generated on which the models are estimated.

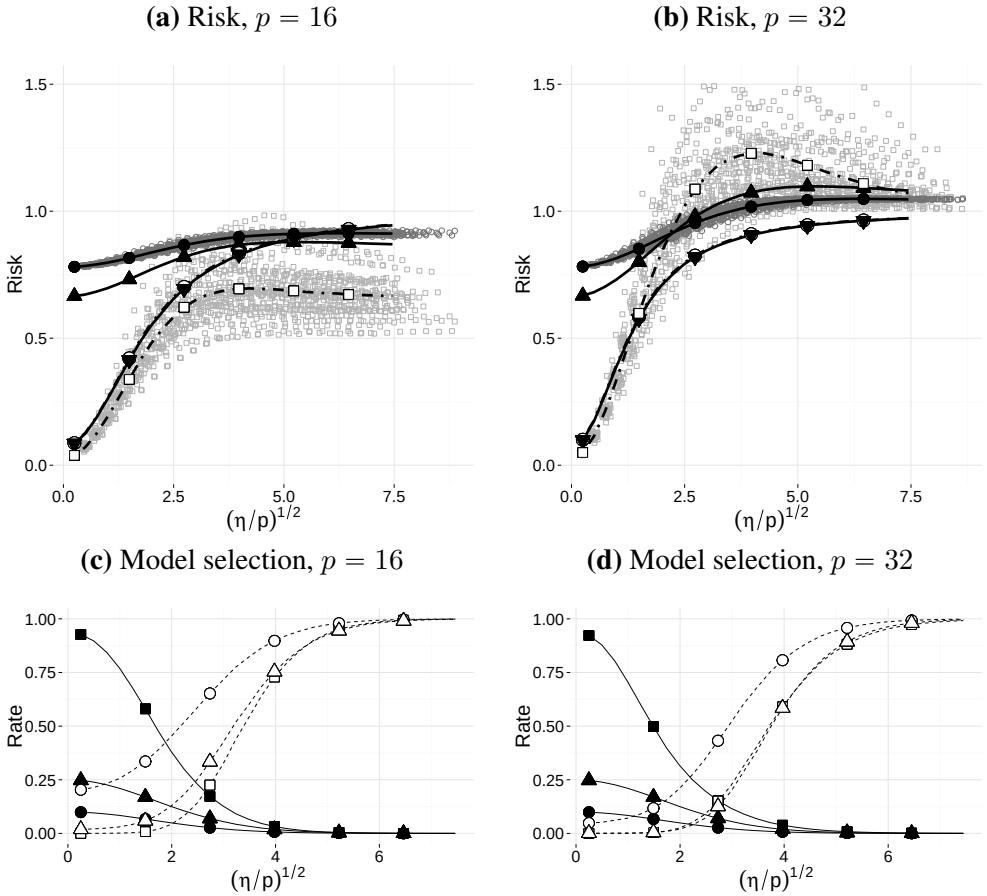
4.5.2 Results

Uncorrelated predictors The risk for uncorrelated predictors is shown in Figure 4.4. In Figure 4.4a, for $p = 16$, all biased estimators have a lower risk than the OLS estimator. The lasso on average has the lowest risk, but its risk function is sensitive to the particular realization of $\boldsymbol{X}$. This is in contrast with r-CS$_{0.10}$ for which the risk shows much less variation. The low variation is also found for r-CS$_{0.25}$, g-CS and JS, but for the clarity of the graph these results are omitted. g-CS has a lower risk than the r-CS estimators when the signal-to-noise ratio is small, but this reverses when the signal-to-noise ratio increases. In Figure 4.4b, for $p = 32$, we see the effect of the 'bet on sparity' that the lasso places. The lasso performs well when the signal-to-noise ratio is small, but the risk increases as the signal-to-noise ratio increases to well above the OLS estimator and Ctrl-shrink estimators. The r-CS estimators are outperformed by the g-CS estimator in this scenario. The difference between the g-CS and the JS estimator is small, but g-CS is consistently lower.

For the predictor-specific shrinkage methods the variable selection accuracy is shown in Figures 4.4c and 4.4d. The Ctrl-shrink estimators by definition bound the fraction of erroneously deleted variables at α. In contrast, the lasso deletes up to 90 percent of the variables which have a nonzero coefficient in the DGP for a small signal-to-noise ratio. This is most likely caused by the fact that the cross-validation procedure focuses on achieving

[1] Computations were performed in parallel on 16 core nodes of which one is left unused to increase computation speed. This makes it convenient to work with multiples of 15.

Figure 4.4: Risk and variable selection, $n = 50$, uncorrelated predictors

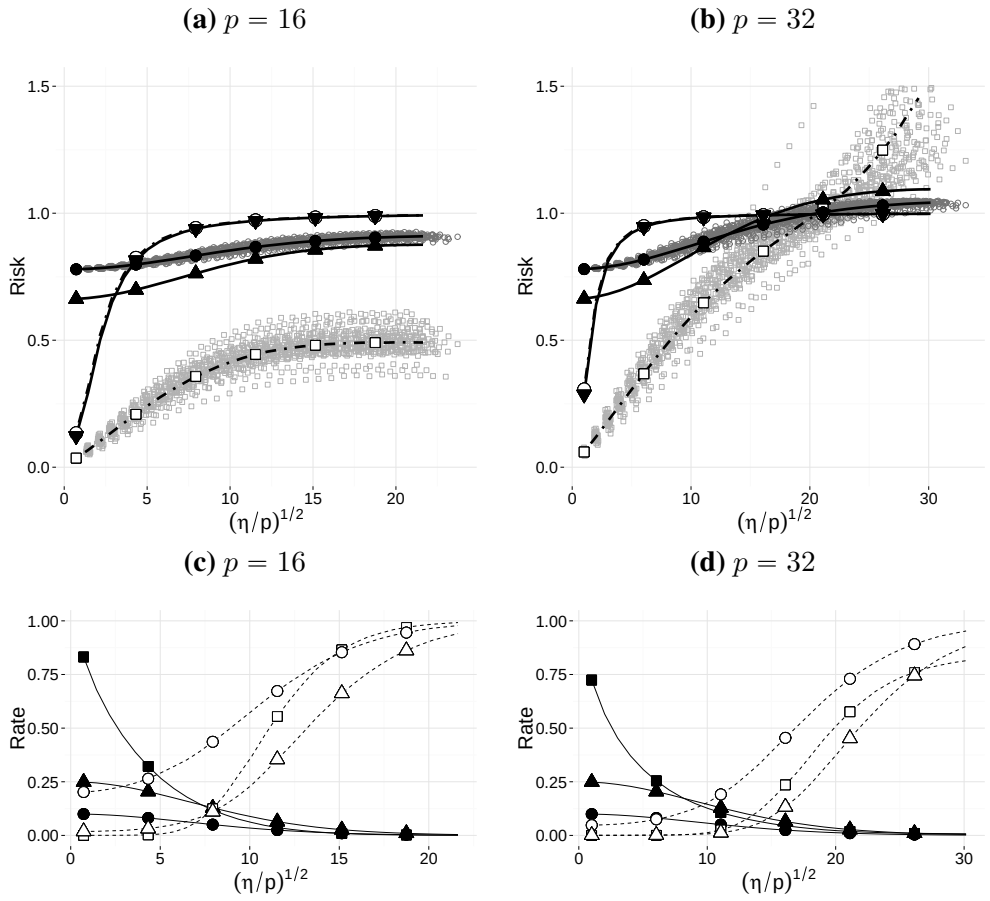


Notes: *DGP* $y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i$, $\varepsilon_i \sim N(0, \sigma^2)$ with $x_i = (1, x_{i1}, \ldots, x_{im})'$ with m = 32 and $x_{ij} \sim N(0, 1)$. Intercept not subject to shrinkage. $\beta_1 = \ldots = \beta_{p+1} = 1$. Signal-to-noise ratio η is varied using a 30-point grid for σ^2. *Upper panel*: solid lines show the risk averaged over 45 realizations of $\boldsymbol{X}$ for r-Cs$_{0.10}$ (—, ●) and r-Cs$_{0.25}$ (—, ▲), the lasso (–·–, □), g-Cs (—, ▼) and JS (–·–, ○, nearly equal to g-Cs). The risk for each realization of $\boldsymbol{X}$ is plotted for r-CS$_{0.10}$ (dark gray) and the lasso (light gray). *Lower panel*: rate of erroneously deleted variables for r-Cs$_{0.10}$ (—, ●), r-Cs$_{0.25}$ (—, ▲) and the lasso (—, ■). Dashed lines: rate at which the method identifies DGP variables.

a low risk, which for small coefficients will not result in a model that is close to the DGP. The difference in variable selection accuracy is also clear from the rate at which all variables that have a nonzero coefficient in the DGP are identified. A substantial difference appears between r-CS$_{0.10}$ and the competing estimators for both $p = 16$ and $p = 32$. For $p = 16$, r-CS$_{0.25}$ also shows substantial improvements over the lasso for small signal-to-noise ratios. For $p = 32$, the lasso is roughly on equal grounds with r-CS$_{0.25}$.

Equicorrelated predictors The results for equicorrelated predictors are given in Figure 4.5. Introducing correlation increases the overall signal-to-noise ratio when the signs of β are equal. The improvements offered by g-CS and JS decrease with the signal-to-noise ratio and in this scenario are outperformed by the lasso and, for larger values of η, also by

Figure 4.5: Risk and variable selection, $n = 50$, $k = 32$, equicorrelated predictors



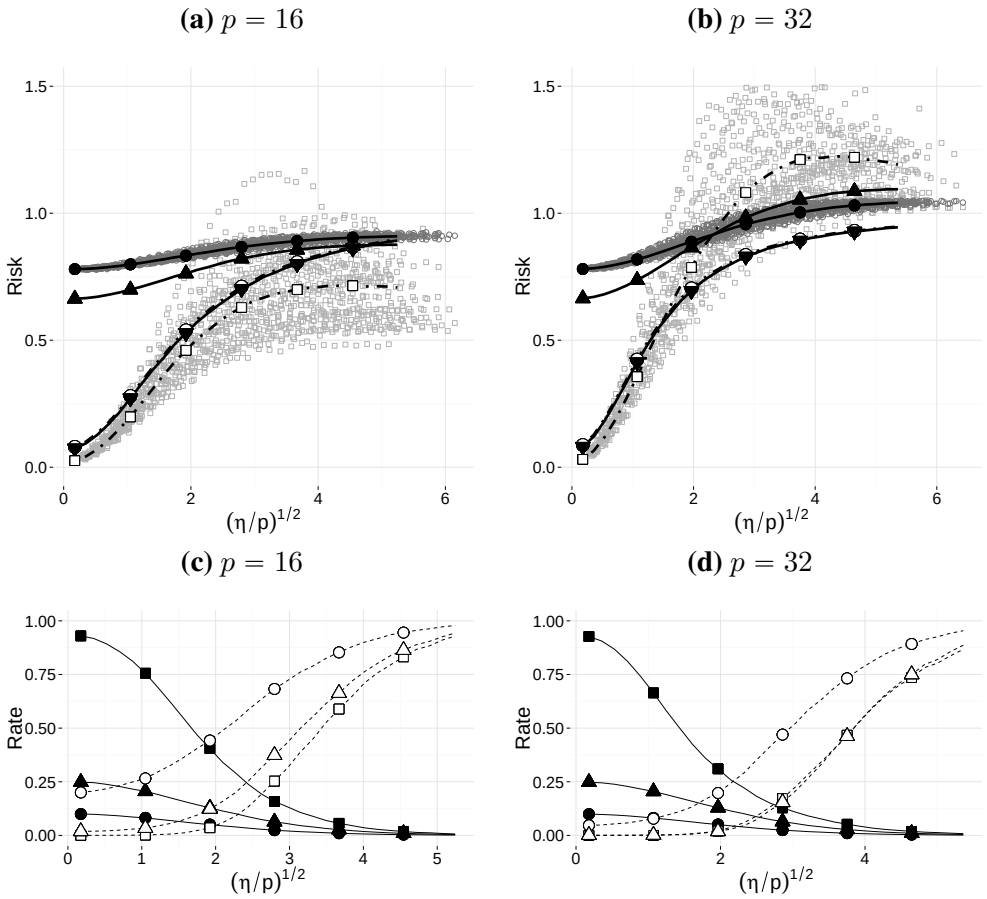
Notes: Correlated predictors: $\text{corr}(x_{ti}, x_{tj}) = 0.5$ for $i \neq j$. For additional information see the *notes* following Figure 4.4

r-CS$_{0.10}$ and r-CS$_{0.25}$. These results might lead to the conclusion that lasso performs well when predictors are correlated. However, in Figure 4.5b for $p = 32$, we see that the risk of the lasso increases too well above the risk of the OLS estimator.

Figures 4.5c and 4.5d show that in terms of variable selection accuracy, r-CS$_{0.10}$ outperforms the alternatives over nearly the entire range of η. Although the lasso deletes a large fraction of relevant variables when η is small, for $p = 16$, there is now a range where the lasso outperforms r-CS$_{0.25}$. For $p = 32$, r-CS$_{0.10}$ offers the best variable selection quality throughout.

Equicorrelated predictors, alternating coefficients Alternating the signs of the coefficients of $\boldsymbol{\beta}$ lowers the overall signal-to-noise ratio compared to the previous setting and makes it comparable to the uncorrelated setting. The results found are similar to the first experiment. The risk of r-CS$_{0.10}$ is robust to different realizations of $\boldsymbol{X}$, while the risk of the lasso is spread out. When $p = 16$ and averaging over $\boldsymbol{X}$ the lasso offers the lowest risk, although the risk for individual realizations of $\boldsymbol{X}$ can be substantially higher. When $p = 32$

Figure 4.6: Risk and variable selection, $n = 50$, $k = 32$, equicorrelated predictors, alternating signs



Notes: Correlated predictors: $\text{corr}(x_{ti}, x_{tj}) = 0.5$ for $i \neq j$. In this experiment, the sign of the coefficients β alternates. For additional information see the *notes* following Figure 4.4

the g-CS and JS estimators yield the lowest risk for the largest part of the parameter region considered. The conclusions concerning variable selection accuracy from the first experimented are maintained in this scenario. r-$CS_{0.10}$ offers the best variable selection accuracy, while the lasso deletes a large fraction of the relevant variables when the signal-to-noise ratio is small.

4.6 Application: prostate cancer data

We use prostate cancer data from Stamey et al. (1989) to analyze the performance of the Ctrl-shrink estimator in an empirical setting. The data measure correlation between the level of prostate specific antigen (lpsa) and eight explanatory variables: log(cancer volume) (lcavol), log(prostate weight) (lweight), age, log(benign prostatic hyperplasia amount) (lbph), seminal vesicle invasion (svi), log(capsular penetration) (lcp), Gleason score (gleason), and percentage Gleason scores 4 or 5 (pgg45). The same data are used to demonstrate the lasso by Tibshirani (1996) and the elastic net by Zou and Hastie (2005). To avoid dependency of the

Table 4.1: Prostate cancer data: risk

	Risk				Relative risk		
	OLS	r-Cs$_{0.10}$	r-Cs$_{0.25}$	lasso	r-Cs$_{0.10}$	r-Cs$_{0.25}$	lasso
Mean	0.568	0.564	0.566	0.612	0.994	0.998	1.100
Std. dev.	0.129	0.126	0.125	0.164	0.029	0.045	0.283
$Q5$	0.361	0.363	0.364	0.385	0.950	0.930	0.781
$Q10$	0.397	0.395	0.397	0.432	0.960	0.947	0.835
$Q50$	0.564	0.561	0.564	0.593	0.993	0.994	1.042
$Q90$	0.733	0.724	0.726	0.799	1.033	1.056	1.400
$Q95$	0.791	0.776	0.773	0.903	1.047	1.079	1.629

Notes: the left panel shows distributional characteristics of the risk obtained using 1000 different split points. $Q5$ denotes the 5% quantile of the risk distribution over the split points. The right panel shows the same characteristics for the risk relative to OLS at each split point.

results on a particular split point, we consider 1,000 different partitions, each with a training sample containing 67 observations and a test set of 30 observations. The lasso tuning parameter is again estimated using fivefold cross-validation.

Summary statistics of the prediction error are reported in Table 4.1. The left panel shows distributional characteristics of risk over the split points. Both Ctrl-shrink estimators improve over OLS in terms of the average risk. In addition, the variance of the error is smaller for these estimators. The lasso is found to perform worse compared to OLS. Table 4.1 shows that the improvements of the Ctrl-shrink estimators over OLS are predominantly found in the upper quantiles. The right panel of Table 4.1 shows the distributional characteristics when for each split point the risk relative to OLS is calculated. The results confirm the conclusions from the simulation exercise. The Ctrl-shrink estimators on average offer a small improvement. Their sensitivity to the use of different split points is small as can be seen from both the variance as from the quantiles. The lasso on the other hand is found to be highly sensitive to different split points, showing potentially large improvements, but even larger losses compared to the OLS estimator.

Table 4.2 provides information on the shrinkage and selection that the different estimators perform. The left panel shows the average rate at which each variable is excluded from the model. The right panel of Table 4.2 shows the average shrinkage factor. In terms of variable selection, the Ctrl-shrink estimators are much more conservative than the lasso estimator. Only *gleason* is frequently omitted by the Ctrl-shrink estimators. The lasso additionally discards *age*, *lbph*, *lcp* and *pgg45*. The Ctrl-shrink estimators also prioritize differently. For example, Ctrl-shrink deletes *lbph* two to three times as often as *age*, while the lasso considers *lbph* to be more relevant.

Table 4.2: Prostate cancer data: variable selection

	Exclusion rate			Average shrinkage factor		
	r-Cs$_{0.10}$	r-Cs$_{0.25}$	lasso	r-Cs$_{0.10}$	r-Cs$_{0.25}$	lasso
Intercept	0.000	0.000	0.000	0.001	0.001	0.000
lcavol	0.000	0.000	0.000	0.023	0.028	0.189
lweight	0.000	0.000	0.055	0.070	0.100	0.446
age	0.005	0.012	0.916	0.127	0.203	0.991
lbph	0.010	0.033	0.606	0.151	0.248	0.891
svi	0.000	0.000	0.026	0.069	0.098	0.500
lcp	0.037	0.108	0.945	0.225	0.384	1.065
gleason	0.161	0.404	0.879	0.427	0.674	0.967
pgg45	0.072	0.171	0.556	0.265	0.433	0.883

4.7 Conclusion and discussion

We provide a new approach to shrinkage and selection based on the distribution of the signal-to-noise ratio of individual predictors as well as of a group of predictors. The definition of the estimator guarantees that exceedingly large shrinkage factors relative to an oracle factor occur at a rate α, which can be set by the researcher. When applied to individual predictors, this ensures that the rate at which variables are erroneously deleted is at most α. When used to find a common shrinkage factor for all coefficients, the estimator dominates the OLS estimator when $k \geqslant 4$ and $\alpha \leqslant 0.50$. Refinements extend the dominance to $k \geqslant 3$. The performance of the predictor-specific estimator in terms of risk and variable selection is consistent and robust to different realizations of the predictor matrix. In the simulation study, the group shrinkage estimator using $\alpha = 0.50$ achieves a slightly lower risk, than the risk of the positive-part James-Stein estimator. An empirical example using prostate cancer data supports the findings of the simulation exercise.

Throughout, we focus on 'Type II' errors which arise naturally as most harmful when considering the risk relative to the unbiased estimator. In some applications 'Type I' errors are more costly, when for example the costs of including variables in subsequent steps is high. One could then alter the definition of Ctrl-shrink such that it controls Type I errors by defining the shrinkage constant through $\Pr[\hat{\omega} < \omega] = \alpha$.

There several possible extensions to the framework discussed here. A practically relevant scenario is when the number of variables exceeds the sample size. In this case, the unbiased OLS estimator, which forms the basis of Ctrl-shrink, does not exist. A possibility is to use $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{+}(\boldsymbol{X}'\boldsymbol{y})$ with $(\boldsymbol{X}'\boldsymbol{X})^{+}$ the Moore-Penrose inverse. Then $\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\mu} = (\boldsymbol{X}'\boldsymbol{X})^{+}(\boldsymbol{X}'\boldsymbol{X})\boldsymbol{\beta}$ and $\boldsymbol{\Sigma} = (\boldsymbol{X}'\boldsymbol{X})^{+}(\boldsymbol{X}'\boldsymbol{X})(\boldsymbol{X}'\boldsymbol{X})^{+}$. Corresponding oracle factors are readily derived. Although in general when $\beta_i = 0$ the coefficients $\mu_i \neq 0$, overselection

is controlled at level α since the signal-to-noise ratio has increased. Consequently the rate at which the corresponding variable is deleted, decreases.

Instead of shrinking the unbiased estimator, we can also use the ridge estimator, which for a known tuning parameter is distributed as $\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Using the same argument as above, overselection is controlled at level α when applying the Ctrl-shrink methodology to this estimator. In this way, a variable selection feature can be added to the Ridge estimator.

When the distribution of the corresponding signal-to-noise ratio is non-standard, bootstrapping can be used to determine the shrinkage constant. However, the Neyman construction requires that the distribution function is simulated for each possible shrinkage constant $\hat{\omega}$. Only then it is possible to find the shrinkage constant for which a fraction α of the bootstrapped empirical distribution function is on the left of the observed value. This is likely to be computationally intensive. Alternatively, one can use the asymptotic distribution of an estimator, which for large classes of estimators is found to be normal.

4.A Approximate normality of the signal-to-noise ratio

For large $\hat{\eta}$ the $\chi^2(k, \nu)$ distribution with $\nu = k\left(\frac{b}{\omega} - 1\right)$ converges to a normal $N\left(\frac{bk}{\hat{\omega}}, 2k\left(\frac{2b}{\hat{\omega}-1}\right)\right)$. To find $\hat{\omega}$ we need to solve

$$\hat{\eta} = bk/\hat{\omega} - z_\alpha\sqrt{2}\sqrt{2b/\hat{\omega} - 1} \tag{4.27}$$

from which we see immediately that for large $\hat{\eta}$ the choice of the significance level α is irrelevant for the weights. The solution to (4.27) is given by

$$\hat{\omega} = \frac{b}{\frac{1}{k^2}\left(\sqrt{k\hat{\eta} + z_\alpha^2 - k^2/2} + z_\alpha\right)^2 + 1/2} \tag{4.28}$$

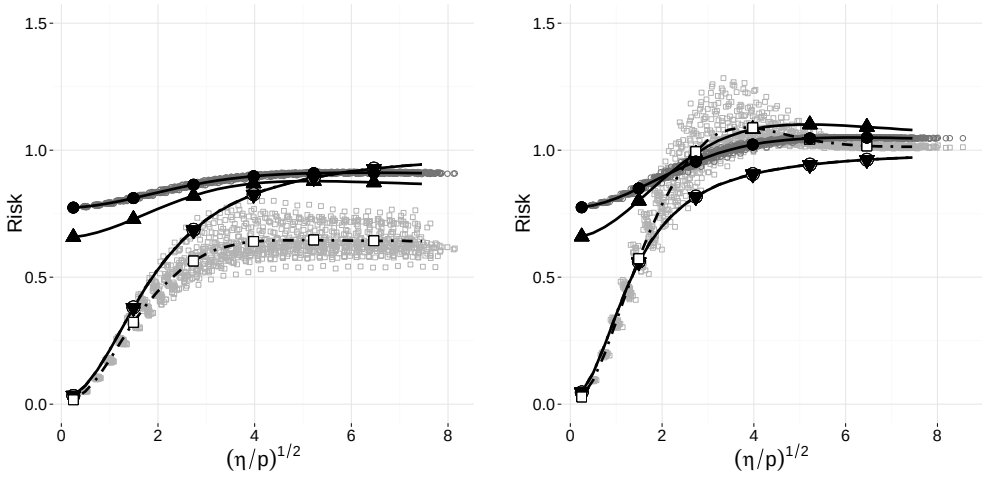
When $\hat{\eta}$ is large, the right-hand side of (4.27) is dominated by the first term, which leads to

$$\lim_{\hat{\eta}\to\infty} \hat{\omega} = \frac{bk}{\hat{\eta}} \tag{4.29}$$

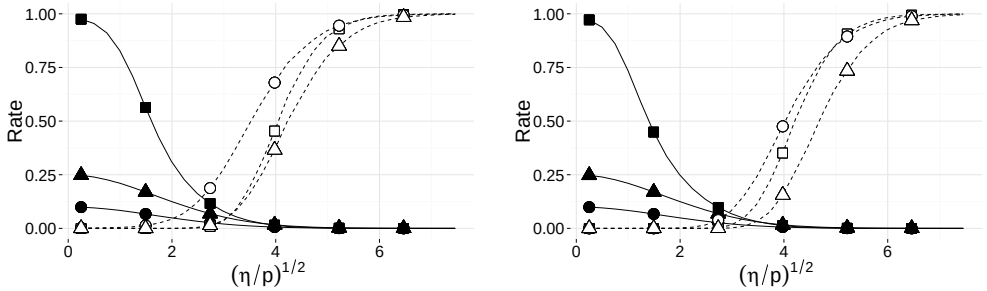
4.B Simulation results for $n = 200$

Figure 4.7: Risk and variable selection, $n = 200$, uncorrelated predictors

(a) Risk, $p = 64$ **(b)** Risk, $p = 128$

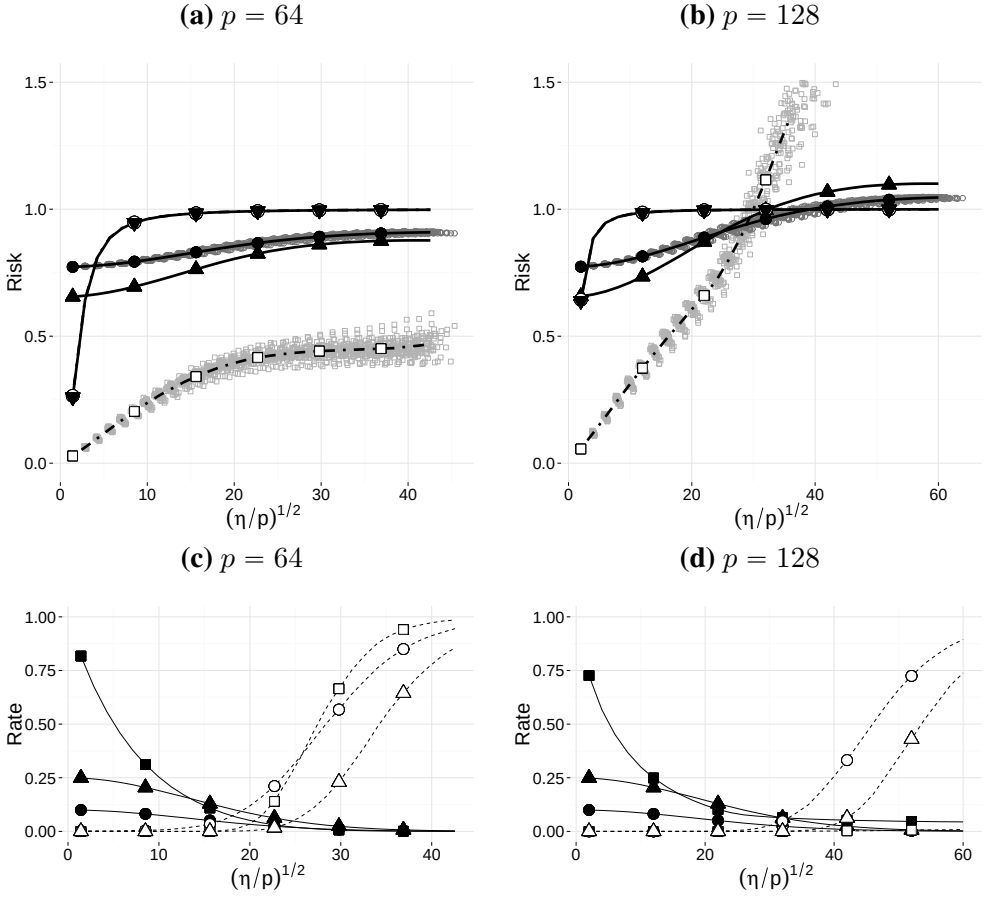


(c) Model selection, $p = 64$ **(d)** Model selection, $p = 128$



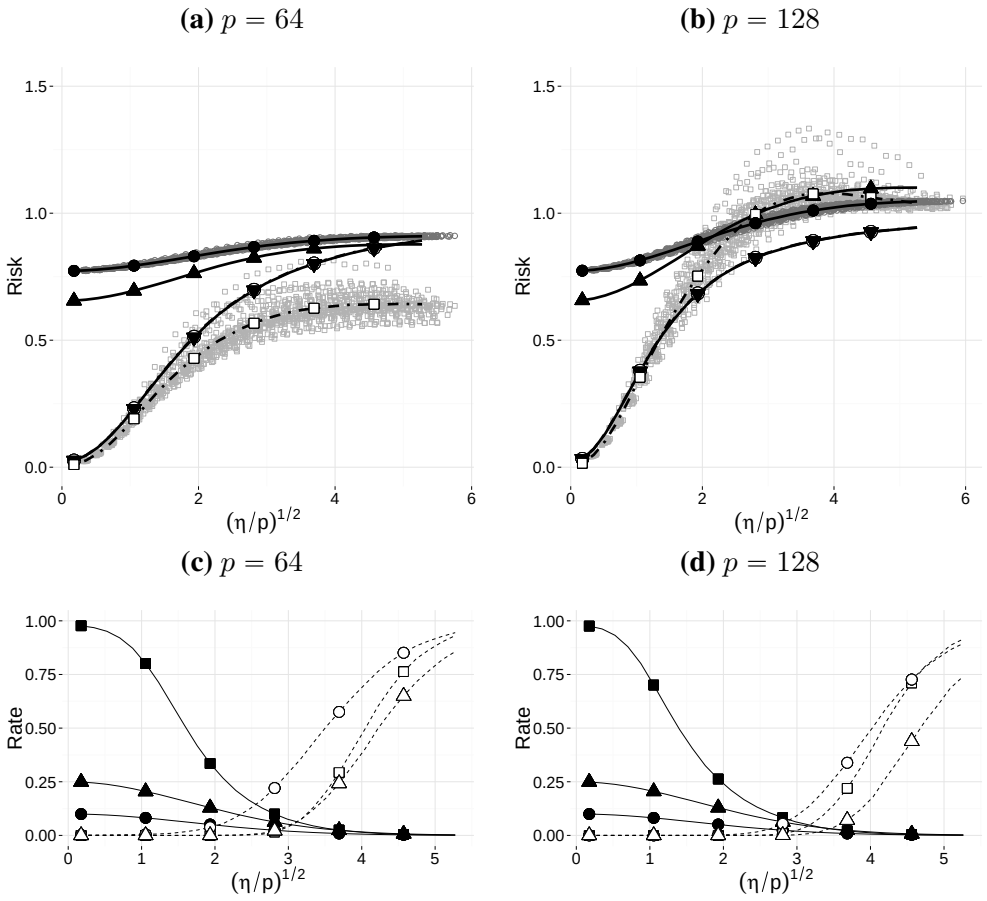
Notes: *DGP* $y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i$, $\varepsilon_i \sim N(0, \sigma^2)$ with $x_i = (1, x_{i1}, \ldots, x_{im})'$ with m = 128 and $x_{ij} \sim N(0, 1)$. Intercept not subject to shrinkage. $\beta_1 = \ldots = \beta_{p+1} = 1$. Signal-to-noise ratio η is varied using a 30-point grid for σ^2. *Upper panel*: risk averaged over 45 realizations of $\boldsymbol{X}$ for r-Cs$_{0.10}$ (—, ●) and r-Cs$_{0.25}$ (—, ▲), the lasso (–·–, □), g-Cs$_{0.50}$ (—, ▼) and JS (–·–, ○, nearly equal to g-Cs$_{0.50}$). The risk for each realization of $\boldsymbol{X}$ is plotted for r-CS using $\alpha = 0.10$ (dark gray) and the lasso (light gray). *Lower panel*: rate of erroneously deleted variables for r-Cs$_{0.10}$ (—, ●), r-Cs$_{0.25}$ (—, ▲) and the lasso (—, ■). Dashed lines: rate at which the method identifies DGP variables.

Figure 4.8: Risk and variable selection, $n = 200$, $k = 128$, equicorrelated predictors



Notes: Correlated predictors: $\text{corr}(x_{ti}, x_{tj}) = 0.5$ for $i \neq j$. For additional information see the *notes* following Figure 4.7

Figure 4.9: Risk and variable selection, $n = 200$, $k = 128$, equicorrelated predictors, alternating signs



Notes: Correlated predictors: $\text{corr}(x_{ti}, x_{tj}) = 0.5$ for $i \neq j$. In this experiment, the sign of the coefficients β alternates. For additional information see the *notes* following Figure 4.7

Chapter 5

Forecasting using random subspace methods

5.1 Introduction

Due to the increase in available macroeconomic data, dimension reduction methods have become an indispensable tool for accurate forecasting. Following Stock and Watson (2002), principal component analysis is widely used to construct a small number of factors from a high-dimensional set of predictors. For an overview of theoretical results and empirical applications, see Stock and Watson (2006).

Instead of combining predictors based on principal component loadings, different combination strategies can be followed. If the underlying factor model is relatively weak, estimation of the factors by principal component analysis is inconsistent as shown by Kapetanios and Marcellino (2010) and one can consider partial least squares as argued by Groen and Kapetanios (2016).

Both principal component regression and partial least squares construct factors by combining the original predictors using data-dependent weights. An intriguing alternative is offered by fully randomized combination strategies. Here, the projection matrix to the low-dimensional subspace is independent of the data and sampled at random from a prespecified probability distribution. In this chapter, we establish theoretical properties of two randomized methods and study their behavior in Monte Carlo simulations and in an extensive application to forecasting monthly macroeconomic data.

The first method we consider is random subset regression, which uses an arbitrary subset of predictors to estimate the model and construct a forecast. The forecasts from many such low-dimensional submodels are then combined in order to lower the mean squared forecast error (MSFE). Previous research by Elliott et al. (2013) focused on the setting where one estimates all possible submodels of fixed dimension. However, when the number of predictors increases, estimating all possible subsets rapidly becomes infeasible. As a practical solution,

Elliott et al. (2013) and Elliott et al. (2015a) propose to draw subsets at random and average over the obtained forecasts. We show that there are in fact strong theoretical arguments for this approach, and establish tight bounds on the resulting MSFE. Using a concentration inequality by Ahlswede and Winter (2002), we also show that it is possible to get arbitrarily close to this bound using a finite and relatively small number of random subsets, explaining why Elliott et al. (2013) find a similar performance when not all subsets are used.

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 normal distribution. Interest in this method sparked by the lemma by Johnson and Lindenstrauss (1984), which states that the geometry of the predictor space is largely preserved under a range of random weighting schemes. This lemma 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). Despite the strong relation to the Johnson-Lindenstrauss lemma, Kabán (2014) shows that in a linear regression model, the underlying assumptions of the lemma are overly restrictive to derive bounds on the in-sample MSFE and that improved bounds can be obtained which eliminate a factor logarithmic in the number of predictors from earlier work by Maillard and Munos (2009). We show that such improved bounds apply to the out-of-sample MSFE as well.

The derived bounds for the two randomized methods can be used to determine in which settings the methods are expected to work well. For random subset regression, the leading bias term depends on the complete eigenvalue structure of the covariance matrix of the data in relation to the non-zero coefficients, while for random projection it depends only on the average of the eigenvalues multiplied by the average coefficient size. This is shown to imply that in settings where the eigenvalues of the population covariance matrix are roughly equal, the difference between both methods will be small. On the other hand, when the model exhibits a factor structure, the methods deviate. If the regression coefficients associated with the most important factors are non-zero, a typical setting for principal component regression, random projection is preferred as the average of the eigenvalues will be small, driving down the MSFE. If on the other hand the relation between the factor structure and the non-zero coefficients is reversed, random subset regression yields more accurate forecasts.

Of practical importance is our finding, both in theory and practice, that the dimension of the subspace should be chosen relatively large. This 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). Instead, in an illustrative example, we find the optimal subspace dimension k^* to be of order $O(\sqrt{ps})$ with p the number of predictors and s the

number of non-zero coefficients. In our empirical setting where $p = 130$, even if $s = 10$, the optimal subspace dimension equals $k^* = 36$.

The theoretical findings are confirmed in a Monte Carlo simulation, which also compares 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). We consider a set-up where the non-zero coefficients are not related to the eigenvalues of the covariance matrix to study the effect of sparsity and signal strength. In addition, we consider two settings where a small number of non-zero coefficients is either associated with the principal components corresponding to large eigenvalues, or to moderately sized eigenvalues.

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, but factors associated with intermediate eigenvalues drive the dependent variable, random subset regression is the only method that outperforms the historical mean of the data.

The theoretical and Monte Carlo findings are empirically tested using the FRED-MD dataset introduced by McCracken and Ng (2015). As the derived theoretical bounds suggest, random subset regression and random projection regression provide similarly accurate forecasts with a clear benefit for random subset regression. This accuracy is shown to be substantially less dependent on the dimension of the reduced subspace than it is in case of principal component regression. In a one-by-one comparison, random subset regression outperforms principal component regression in 88% of the series, partial least squares in 70%, Lasso in 82% and Ridge in 67%. Random projection regression likewise outperforms the benchmarks for a majority of the series and is more accurate than principal component regression in 85% of the series, partial least squares in 56%, Lasso in 82% and Ridge in 57%. Random subset regression is more accurate than random projection regression in 65% of the series, indicating that the factor scenario in the Monte Carlo study where non-zero coefficients are associated with intermediate eigenvalues, is empirically more relevant.

The article is structured as follows. Using results from random matrix theory, Section 5.2 provides tight bounds on the MSFE under random subset regression and random projection regression. A Monte Carlo study is carried out in Section 5.3, which highlights the performance of the techniques under different model specifications. Section 5.4 considers an extensive empirical application using monthly macroeconomic data obtained from the FRED-MD database. Section 5.5 concludes.

5.2 Theoretical results

In this section, we start by setting up a general dimension reduction framework, that naturally fits both deterministic and random methods. We subsequently introduce two different randomized reduction methods: random subset regression and random projection regression. We derive bounds on the MSFE under general projection matrices, after which we specialize to the case where these matrices are random. The resulting bounds turn out to be highly informative on scenarios where the methods can be expected to work well.

Consider the data generating process (DGP)

$$y_{t+1} = \boldsymbol{x}_t'\boldsymbol{\beta} + \varepsilon_{t+1} \tag{5.1}$$

for $t = 1, \ldots, T$, and where $\boldsymbol{x}_t'$ is a vector of predictors in $\mathbb{R}^p$. We assume that the errors satisfy $\varepsilon_t \sim i.i.d.(0, \sigma^2)$. We regard the predictors $\boldsymbol{x}_t$ as weakly exogenous, which is not overly restrictive as one typically does not average over lagged terms of the dependent variable. The DGP in (5.1) can be straightforwardly adjusted to the situation where some predictors always need to be included.

Since the variance of ordinary least squares (OLS) estimates increases with the number of estimated coefficients, forecasts can get inaccurate when large numbers of predictors are available. As a solution, we project the p-dimensional vector of predictors $\boldsymbol{x}_t$ on a k-dimensional subspace using a matrix $\boldsymbol{R}_i \in \mathbb{R}_i^{p\times k}$

$$\tilde{\boldsymbol{x}}_t' = \boldsymbol{x}_t'\boldsymbol{R}_i \tag{5.2}$$

A frequently used choice for $\boldsymbol{R}_i$ in order to reduce the number of predictors, is to take the matrix of principal component loadings corresponding to the k largest eigenvalues from the sample covariance matrix $\frac{1}{T-1}\sum_{t=1}^{T-1}\boldsymbol{x}_t\boldsymbol{x}_t'$. Instead of using a single deterministic matrix, randomized methods sample a large number of different realizations of $\boldsymbol{R}_i$ from a prespecified probability distribution.

As mentioned above, we consider two different methods to generate $\boldsymbol{R}_i$: random subset regression and random projection regression, which are defined as follows.

Random subset regression In random subset regression, the matrix $\boldsymbol{R}_i$ is a random permutation matrix that selects a random set of k predictors out of the original p available

predictors. For example, if $p = 5$ and $k = 3$, a possible realization of $\boldsymbol{R}_i$ is

$$\boldsymbol{R}_i = \sqrt{\frac{5}{3}} \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} \tag{5.3}$$

For a single realization of $\boldsymbol{R}_i$, the probability that a diagonal element is non-zero equals $k/p = 3/5$. The scaling factor thus ensures that $\mathrm{E}[\boldsymbol{R}_i \boldsymbol{R}_i'] = \boldsymbol{I}$, which is required in the following sections. More formally, define an index $l = 1, \dots k$ with k the dimension of the subspace, and a scalar $c(l)$ such that $1 \leqslant c(l) \leqslant p$. Denote by $e_{c(l)}$ the p-dimensional unit vector with the $c(l)$-th entry equal to one, then random subset regression is based on random projection matrices of the form

$$\boldsymbol{R}_i = \sqrt{\frac{p}{k}} \left[\boldsymbol{e}^i_{c(1)}, \dots, \boldsymbol{e}^i_{c(k)} \right] \quad e^i_{c(m)} \neq e^i_{c(n)} \text{ if } m \neq n \tag{5.4}$$

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 chooses the weights at random from a normal distribution. In this case, each entry of $\boldsymbol{R}_i$ is independent and identically distributed as

$$[\boldsymbol{R}_i]_{mn} \sim N\left(0, \frac{1}{k}\right) \quad 1 \leqslant m \leqslant p,\ 1 \leqslant n \leqslant k \tag{5.5}$$

where the scaling is again introduced to ensure $\mathrm{E}[\boldsymbol{R}_i \boldsymbol{R}_i'] = \boldsymbol{I}_p$. In fact, a broader class of sampling distributions is allowed. For the results below, it is only required that the entries have zero mean and finite fourth moment.

5.2.1 Mean squared forecast error bound

We now derive a bound on the mean squared forecast error for general projection matrices $\boldsymbol{R}_i$, which can be deterministic or random. Following the ideas set out by Kabán (2014), we rewrite the data generating process (5.1) as

$$y_{t+1} = \boldsymbol{x}_t' \boldsymbol{R}_i \boldsymbol{R}_i' \boldsymbol{\beta} + \boldsymbol{x}_t' (\boldsymbol{I} - \boldsymbol{R}_i \boldsymbol{R}_i') \boldsymbol{\beta} + \varepsilon_{t+1} \tag{5.6}$$

Instead of (5.6) we estimate the low-dimensional model

$$y_{t+1} = \boldsymbol{x}_t' \boldsymbol{R}_i \boldsymbol{\gamma}_i + \tilde{\varepsilon}_{t+1} \tag{5.7}$$

where $\boldsymbol{\gamma}_i \in \mathbb{R}^k$ denotes the optimal parameter vector in the k-dimensional subproblem, that is

$$\boldsymbol{\gamma}_i = \arg\min_{\boldsymbol{u}} \sum_{t=1}^{T-1} \left(\boldsymbol{x}_t'\boldsymbol{\beta} - \boldsymbol{x}_t'\boldsymbol{R}_i\boldsymbol{u}\right)^2 \tag{5.8}$$

The least squares estimator of $\boldsymbol{\gamma}_i$ is denoted by $\hat{\boldsymbol{\gamma}}_i$ and given by

$$\hat{\boldsymbol{\gamma}}_i = \left(\sum_{t=1}^{T-1} \boldsymbol{R}_i'\boldsymbol{x}_t\boldsymbol{x}_t'\boldsymbol{R}_i\right)^{-1} \left(\sum_{t=1}^{T-1} \boldsymbol{R}_i'\boldsymbol{x}_t y_{t+1}\right) \tag{5.9}$$

Using this estimate, we construct a forecast as

$$\hat{y}_{T+1}^i = \boldsymbol{x}_T'\boldsymbol{R}_i\hat{\boldsymbol{\gamma}}_i \tag{5.10}$$

If $\boldsymbol{R}_i$ is random, then intuitively, relying on a single realization of the random matrix $\boldsymbol{R}_i$ is suboptimal. By Jensen's inequality, we indeed find that averaging over different realizations of $\boldsymbol{R}_i$ will improve the accuracy

$$\begin{aligned}
&\mathrm{E}\left[(\mathrm{E}_{R_i}\left[\hat{y}_{T+1}^i\right] - \boldsymbol{x}_T'\boldsymbol{\beta})^2\right] = \\
&\quad = \mathrm{E}\left[\mathrm{E}_{R_i}[\hat{y}_{T+1}^i]^2\right] - 2\mathrm{E}\left[\mathrm{E}_{R_i}[\hat{y}_{T+1}^i]\boldsymbol{x}_T'\boldsymbol{\beta}\right] + \mathrm{E}\left[(\boldsymbol{x}_T'\boldsymbol{\beta})^2\right] \\
&\quad \leqslant \mathrm{E}_{R_i}\left[\mathrm{E}\left[\hat{y}_{T+1}^i\right]^2\right] - 2\mathrm{E}_{R_i}\left[\mathrm{E}[\hat{y}_{T+1}^i]\boldsymbol{x}_T'\boldsymbol{\beta}\right] + \mathrm{E}\left[(\boldsymbol{x}_T'\boldsymbol{\beta})^2\right] \\
&\quad \leqslant \mathrm{E}_{R_i}\left[\mathrm{E}\left[(\hat{y}_{T+1}^i - \boldsymbol{x}_T'\boldsymbol{\beta})^2\right]\right]
\end{aligned} \tag{5.11}$$

where E_{R_i} denotes the expectation with respect to the random variable $\boldsymbol{R}_i$. For ease of exposition we ignore the variance term ε_{T+1}.

Following (5.11), we consider the MSFE after averaging over different realizations of the projection matrix $\boldsymbol{R}_i$. For a single, deterministic projection matrix, this expectation is obviously superfluous. The following bound can be established on the mean squared forecast error

Theorem 1 Let $\boldsymbol{x}_t$ a vector of predictors for which $\frac{1}{T}\sum_{t=1}^{T-1} \boldsymbol{x}_t\boldsymbol{x}_t' \xrightarrow{p} \boldsymbol{\Sigma}_X$ and $\mathrm{E}[\boldsymbol{x}_t\boldsymbol{x}_t'] = \boldsymbol{\Sigma}_X$ for all t, then

$$\begin{aligned}
&E\left[\left(\boldsymbol{x}_T'\boldsymbol{\beta} - \boldsymbol{x}_T'\mathrm{E}_{R_i}\left[\boldsymbol{R}_i\hat{\boldsymbol{\gamma}}_i\right]\right)^2\right] = \\
&\quad \leqslant \sigma^2\frac{k}{T} + \mathrm{E}_{R_i}\left[\boldsymbol{\beta}'(\boldsymbol{I} - \boldsymbol{R}_i\boldsymbol{R}_i')\boldsymbol{\Sigma}_X(\boldsymbol{I} - \boldsymbol{R}_i\boldsymbol{R}_i')\boldsymbol{\beta}\right] + o_p(T^{-1})
\end{aligned} \tag{5.12}$$

A proof is presented in Appendix 5.A.

The first term of (5.12) represents the variance of the estimates. This can be compared to the variance that is achieved by forecasting using OLS estimates for $\boldsymbol{\beta}$, which is $\sigma^2\frac{p}{T}$.

The second term reflects the bias that arises by estimating $\boldsymbol{\beta}$ in a low-dimensional subspace. Loosely speaking, if in (5.12) the product $\boldsymbol{R}_i\boldsymbol{R}_i'$ concentrates tightly around $\boldsymbol{I}$ under a particular choice of sampling distribution, then the bias term will be small. It is exactly this concentration that underlies the power of randomized methods.

The effect of the choice of k on the bias, can be anticipated from (5.12). The elements of the matrix $\boldsymbol{R}_i\boldsymbol{R}_i'$ are averages of k products of random entries. Intuitively, as k increases, the concentration of $\boldsymbol{R}_i\boldsymbol{R}_i'$ around its expected value $\boldsymbol{I}$ will tighten. Indeed, we show below that the bias is a decreasing function of k, emphasizing the bias-variance trade-off governed by the choice of the subspace dimension k.

We now specialize to the two different randomized methods, in which case analytic expression are available for the expectation in the bias term.

MSFE bound for random subset regression

For random subset regression, the dimension of the original data space is reduced using a random permutation matrix $\boldsymbol{R}_i$ defined in (5.4). For this type of matrices we have the following result by Tucci and Wang (2011)

Theorem 2: Let $\boldsymbol{R}_i \in \mathbb{R}^{p\times k}$ be a random permutation matrix, scaled such that $\mathrm{E}[\boldsymbol{R}_i\boldsymbol{R}_i'] = \boldsymbol{I}$. Then

$$\begin{aligned} &\mathrm{E}_{R_i}^{RS}\left[(\boldsymbol{I}-\boldsymbol{R}_i\boldsymbol{R}_i')\boldsymbol{\Sigma}_X(\boldsymbol{I}-\boldsymbol{R}_i\boldsymbol{R}_i')\right] = \\ &\quad = \frac{p}{k}\left(\left[\frac{k-1}{p-1}-\frac{k}{p}\right]\boldsymbol{\Sigma}_X + \frac{p-k}{p-1}\boldsymbol{D}_{\Sigma_X}\right) \end{aligned} \tag{5.13}$$

where $[\boldsymbol{D}_{\Sigma_X}]_{ii} = [\boldsymbol{\Sigma}_X]_{ii}$, and $[\boldsymbol{D}_{\Sigma_X}]_{ij} = 0$ if $i \neq j$.

Substituting this expression into (5.12), we obtain that for random subset regression

$$\begin{aligned} &E\left[\left(\boldsymbol{x}_T'\boldsymbol{\beta} - \boldsymbol{x}_T'\mathrm{E}_{R_i}^{RS}\left[\boldsymbol{R}_i\hat{\boldsymbol{\gamma}}_i\right]\right)^2\right] = \\ &\quad \leqslant \frac{\sigma^2 k}{T} + \frac{p-k}{k}\frac{p}{p-1}\left[\boldsymbol{\beta}'\boldsymbol{D}_{\Sigma_X}\boldsymbol{\beta} - \frac{1}{p}\boldsymbol{\beta}'\boldsymbol{\Sigma}_X\boldsymbol{\beta}\right] + o_p(T^{-1}) \end{aligned} \tag{5.14}$$

We observe that as $k \to p$, the bias decreases and we obtain the variance formula for the OLS estimates of $\boldsymbol{\beta}$ when $k = p$. In many high-dimensional settings, we expect $p \gg k$ and $p, k \gg 1$, such that the leading bias term is $\frac{p}{k}\boldsymbol{\beta}'\boldsymbol{D}_{\Sigma_X}\boldsymbol{\beta}$. We will discuss this term in more depth in an illustrating example below.

MSFE bound for random projection regression

For random projection defined in (5.5), the following theorem is derived by Kabán (2014)

Theorem 3 For $\boldsymbol{R}_i \in \mathbb{R}^{p\times k}$ and $[\boldsymbol{R}_i]_{mn} = N\left(0, \frac{1}{\sqrt{k}}\right)$ and $\boldsymbol{\Sigma_X}$ a positive semi-definite matrix

$$\begin{aligned} &\mathrm{E}_{R_i}^{RP}\left[(\boldsymbol{I} - \boldsymbol{R}_i\boldsymbol{R}_i')\boldsymbol{\Sigma}_X(\boldsymbol{I} - \boldsymbol{R}_i\boldsymbol{R}_i')\right] = \\ &\quad = \frac{p}{k}\left[\left(\frac{k+1}{p} - \frac{k}{p}\right)\boldsymbol{\Sigma}_X + \frac{1}{p}\mathrm{trace}(\boldsymbol{\Sigma}_X)\boldsymbol{I}\right] \end{aligned} \tag{5.15}$$

This result holds when the assumption on the entries of the random matrix is weakened, requiring only that they are drawn from a symmetric distribution with zero mean and finite fourth moments.

Substituting (5.15) into (5.12), the mean squared forecast error that follows from random projection regression satisfies the following bound

$$\begin{aligned} &E\left[\left(\boldsymbol{x}_T'\boldsymbol{\beta} - \boldsymbol{x}_T'\mathrm{E}_{R_i}^{RP}\left[\boldsymbol{R}_i\hat{\boldsymbol{\gamma}}_i\right]\right)^2\right] = \\ &\quad \leqslant \frac{\sigma^2 k}{T} + \frac{1}{k}\left[\boldsymbol{\beta}'\boldsymbol{\Sigma}_X\boldsymbol{\beta} + \mathrm{trace}(\boldsymbol{\Sigma}_X)\boldsymbol{\beta}'\boldsymbol{\beta}\right] + o_p(T^{-1}) \end{aligned} \tag{5.16}$$

A notable difference with random subset regression is that the bias term remains non-zero even when $p = k$. The reason is that the columns of the projections matrix are not exactly orthogonal, and therefore might span a smaller space than the original predictor matrix. Indeed, when the columns are orthogonalized, the following theorem by Marzetta et al. (2011) guarantees that the bias is identically zero when $k = p$.

Theorem 4 Let $\boldsymbol{R}_i$ a random matrix with i.i.d. normal entries such that $\boldsymbol{R}_i'\boldsymbol{R}_i = \frac{p}{k}\boldsymbol{I}_k$ and $\boldsymbol{\Sigma}_X$ a positive semi-definite matrix, then

$$\begin{aligned} &\mathrm{E}_{R_i}^{ORP}\left[(\boldsymbol{I} - \boldsymbol{R}_i\boldsymbol{R}_i')\boldsymbol{\Sigma}_X(\boldsymbol{I} - \boldsymbol{R}_i\boldsymbol{R}_i')\right] = \\ &\quad = \frac{p}{k}\left[\left(\frac{pk-1}{p^2-1} - \frac{k}{p}\right)\boldsymbol{\Sigma}_X + \frac{p-k}{p^2-1}\mathrm{trace}(\boldsymbol{\Sigma}_X)\boldsymbol{I}\right] \end{aligned} \tag{5.17}$$

Hence, the MSFE after orthogonalization is bounded by

$$\begin{aligned} &E\left[\left(\boldsymbol{x}_T'\boldsymbol{\beta} - \boldsymbol{x}_T'\mathrm{E}_{R_i}^{ORP}\left[\boldsymbol{R}_i\hat{\boldsymbol{\gamma}}_i\right]\right)^2\right] = \\ &\quad \leqslant \frac{\sigma^2 k}{T} + \frac{p-k}{k}\frac{p^2}{p^2-1}\left[\frac{\mathrm{trace}(\boldsymbol{\Sigma}_X)}{p}\boldsymbol{\beta}'\boldsymbol{\beta} - \frac{1}{p}\boldsymbol{\beta}'\boldsymbol{\Sigma}_X\boldsymbol{\beta}\right] + o_p(T^{-1}) \end{aligned} \tag{5.18}$$

where the second term equals zero when $p = k$. Orthogonalization leads to an improved bound compared to (5.16), since the difference in MSFE between random projection and its orthogonalized form satisfies

$$E\left[\left(\boldsymbol{x}_T'\boldsymbol{\beta} - \boldsymbol{x}_T'\mathrm{E}_{R_i}^{RP}\left[\boldsymbol{R}_i\hat{\boldsymbol{\gamma}}_i\right]\right)^2\right] - E\left[\left(\boldsymbol{x}_T'\boldsymbol{\beta} - \boldsymbol{x}_T'\mathrm{E}_{R_i}^{ORP}\left[\boldsymbol{R}_i\hat{\boldsymbol{\gamma}}_i\right]\right)^2\right] \geqslant 0 \tag{5.19}$$

which is derived in Appendix 5.B. However, orthogonalization is computationally costly and in many examples the dimensions of the problem are such that the gain in predictive accuracy will be negligible.

A second important difference with the results for random subset regression, is that when $p \gg k$ and $p, k \gg 1$, the leading bias term equals $\frac{\text{trace}(\boldsymbol{\Sigma}_X)}{k}\boldsymbol{\beta}'\boldsymbol{\beta}$. For random subset regression the leading term was found to be $\frac{p}{k}\boldsymbol{\beta}'\boldsymbol{D}_{\Sigma_X}\boldsymbol{\beta}$. This points out a conceptual difference between the two methods that is further analyzed in the next section.

Comparison between the MSFE of OLS, RS, and RP

To gain intuition for the performance of the randomized methods compared with unrestricted estimation by ordinary least squares (OLS), and to show when one of the randomized methods is preferred over the other, we consider a simplified setting. This setting nevertheless brings out the main features we observe in the more sophisticated set-up studied in the Monte Carlo simulations described in Section 5.3.

Suppose $p \gg k$ and $p, k \gg 1$, then from (5.14) we have that the leading bias term for random subset regression is $\frac{p}{k}\boldsymbol{\beta}'\boldsymbol{D}_{\Sigma_X}\boldsymbol{\beta}$. For random projection, we have from (5.16) that the leading bias term equals $\frac{\text{trace}(\boldsymbol{\Sigma}_X)}{k}\boldsymbol{\beta}'\boldsymbol{\beta}$. Suppose that the population covariance matrix is given by

$$\boldsymbol{\Sigma}_X = \begin{pmatrix} 1+\alpha & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix} \tag{5.20}$$

For notational convenience, assume that $\frac{\sigma^2}{T} = 1$. In this setting, the MSFE for random subset regression is given by

$$E\left[\left(\boldsymbol{x}_T'\boldsymbol{\beta} - \boldsymbol{x}_T'\mathrm{E}_{R_i}^{RS}\left[\boldsymbol{R}_i\hat{\boldsymbol{\gamma}}_i\right]\right)^2\right] \leqslant k + \frac{p}{k}(\alpha\beta_1^2 + \boldsymbol{\beta}'\boldsymbol{\beta}) \tag{5.21}$$

This expression depends explicitly on the size of the coefficient β_1. This in contrast with random projection regression, for which the MSFE is given by

$$E\left[\left(\boldsymbol{x}_T'\boldsymbol{\beta} - \boldsymbol{x}_T'\mathrm{E}_{R_i}^{RP}\left[\boldsymbol{R}_i\hat{\boldsymbol{\gamma}}_i\right]\right)^2\right] \leqslant k + \frac{p+\alpha}{k}\boldsymbol{\beta}'\boldsymbol{\beta} \tag{5.22}$$

RS and RP versus OLS The simplest scenario is when $\alpha = 0$ in (5.20), and $\beta_i = c$ for $i = 1, \ldots, s$, with $s \leqslant p$, and zero otherwise. We refer to s as the sparsity of the coefficient vector $\boldsymbol{\beta}$. Both for RS and RP the bound on the MSFE reduces to

$$E\left[\left(\boldsymbol{x}_T'\boldsymbol{\beta} - \boldsymbol{x}_T'\mathrm{E}_{R_i}\left[\boldsymbol{R}_i\hat{\boldsymbol{\gamma}}_i\right]\right)^2\right] \leqslant \frac{\sigma^2}{T}\left[k + \frac{ps}{k}c^2\right] \tag{5.23}$$

When using the optimal value of k derived in Appendix 5.C, $k^* = c\sqrt{ps}$, this reduces to

$$E\left[\left(\boldsymbol{x}_T'\boldsymbol{\beta} - \boldsymbol{x}_T'\mathrm{E}_{R_i}\left[\boldsymbol{R}_i\hat{\boldsymbol{\gamma}}_i\right]\right)^2\right] \leqslant 2c\sqrt{ps} \tag{5.24}$$

Note that the optimal size is of order $O(\sqrt{ps})$, which can be much larger than what one might expect based on findings when forecasting using factor models where typically around 5 factors are selected, as for example in Stock and Watson (2012). In the empirical setting of Section 5.4, we have $p = 130$ such that even at a sparsity level of 10%, the optimal model size is $k^* = 36$.

Under the optimal value of k, the relative performance compared to OLS is given by $2c\sqrt{\frac{s}{p}}$. As one might expect, the increase in accuracy of the randomized methods compared to OLS is larger when the coefficient size and the number of non-zero coefficients are small.

RS versus RP To examine the relative performance of RS and RP, we analyze the difference in MSFE obtained from (5.21) and (5.22)

$$\Delta = \frac{p}{k}\alpha\left[\beta_1^2 - \frac{\boldsymbol{\beta}'\boldsymbol{\beta}}{p}\right] \tag{5.25}$$

If all coefficients are of the same size, then $\beta_1^2 \approx \frac{\boldsymbol{\beta}'\boldsymbol{\beta}}{p}$ and the methods are expected to perform equally well. The same happens if the covariance matrix is well-conditioned, i.e. $\alpha \to 0$ and all eigenvalues of the covariance matrix are of the same size.

For non-zero α, two things can happen. First, consider a typical principal component regression setting where β_1 is large while all other coefficients are close or equal to zero. Here, the MSFE for random projection is only affected by the large coefficient β_1 through the inner product $\frac{\boldsymbol{\beta}'\boldsymbol{\beta}}{p}$. Random subset regression on the other hand suffers, as the MSFE depends explicitly on the product $\alpha\boldsymbol{\beta}_1^2$. This setting therefore favors random projection. The difference between the two methods increases as β_1 and/or α grow larger.

In contrast with the previous setting, it is also possible that the factor associated with the largest eigenvalue of $\boldsymbol{\Sigma}_X$ is not associated with the dependent variable. This is the case when α is large, while $\beta_1 = 0$. If any signal is present in the remaining factors, random subset regression will outperform random projection.

In addition to the contrast in MSFE, there is also a difference in the optimal subspace dimension. We have

$$\begin{aligned} k^*_{RS} &= \sqrt{p(\alpha\beta_1^2 + \boldsymbol{\beta}'\boldsymbol{\beta})} \\ k^*_{RP} &= \sqrt{(p+\alpha)\boldsymbol{\beta}'\boldsymbol{\beta}} \end{aligned} \tag{5.26}$$

In the factor setting where both α and β_1 are large, the optimal dimension for random subset regression can be much larger. If on the other hand β_1 is close to or equal to zero, random projection chooses a larger subspace dimension when $\alpha > 0$.

5.2.2 Feasibility of the MSFE bounds

The bounds from the previous section are calculated using expectations over the random matrix $\boldsymbol{R}_i$. In reality we have to settle for a finite number of draws. We therefore need the average over these draws to concentrate around the expectation, i.e. with high probability it should hold that

$$\Delta = \left|\left| \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{R}_i\boldsymbol{R}_i'\boldsymbol{\Sigma}_X\boldsymbol{R}_i\boldsymbol{R}_i' - \mathrm{E}\left[\boldsymbol{R}_i\boldsymbol{R}_i'\boldsymbol{\Sigma}_X\boldsymbol{R}_i\boldsymbol{R}_i'\right] \right|\right| < e \tag{5.27}$$

where $||\cdot||$ denotes the Euclidean norm and e is some small, positive number. Such a concentration can be proven both for random projections and for random subset regression using the following theorem by Ahlswede and Winter (2002)

Theorem 5 Let $\boldsymbol{X}_i$, $i = 1, \ldots, N$ be a $p \times p$ independent random positive semi-definite matrix with $||\boldsymbol{X}_i|| \leqslant 1$ almost surely. Let $\boldsymbol{S}_N = \sum_{i=1}^{N} \boldsymbol{X}_i$ and $\Omega = \sum_{i=1}^{N} ||\mathrm{E}[\boldsymbol{X}_i]||$, then for all $\epsilon \in (0, 1)$

$$\mathrm{P}\left(||\boldsymbol{S}_N - \mathrm{E}[\boldsymbol{S}_N]|| \geqslant \epsilon\Omega\right) \leqslant 2p\exp(-\epsilon^2\Omega/4) \tag{5.28}$$

Since this holds for all $\epsilon \in (0, 1)$, we can make $\epsilon\Omega$ arbitrarily small, which we use to show that (5.27) holds with high probability for small e. Using the same approach, it is then straightforward to show that

$$\tilde{\Delta} = \left|\left| \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{R}_i\boldsymbol{R}_i' - \mathrm{E}\left[\boldsymbol{R}_i\boldsymbol{R}_i'\right] \right|\right| < e \tag{5.29}$$

for some finite number N.

Random subset regression Consider random permutation matrices $\boldsymbol{R}_i \in \mathbb{R}^{p\times k}$ suitably scaled by a factor $\sqrt{\frac{p}{k}}$ to ensure that $\mathrm{E}[\boldsymbol{R}_i\boldsymbol{R}_i'] = \boldsymbol{I}$. Let $\boldsymbol{Q}_i = \boldsymbol{R}_i\boldsymbol{R}_i'\boldsymbol{\Sigma}_X\boldsymbol{R}_i\boldsymbol{R}_i'$, then

$$||\boldsymbol{Q}_i|| \leqslant ||\boldsymbol{\Sigma}_X|| \cdot ||\boldsymbol{R}_i\boldsymbol{R}_i'||^2 = \left(\frac{p}{k}\right)^2 ||\boldsymbol{\Sigma}_X|| \tag{5.30}$$

using that for any draw of $\boldsymbol{R}_i$, the Euclidean norm of the outer product satisfies $||\boldsymbol{R}_i\boldsymbol{R}_i'|| = \frac{p}{k}$. Define now $\boldsymbol{X}_i = \boldsymbol{Q}_i/||\boldsymbol{Q}_i||$. Then

$$\Omega = N\frac{||\mathrm{E}\left[\boldsymbol{R}_i\boldsymbol{R}_i'\boldsymbol{\Sigma}_X\boldsymbol{R}_i\boldsymbol{R}_i'\right]||}{\left(\frac{p}{k}\right)^2||\boldsymbol{\Sigma}_X||} \tag{5.31}$$

where we use that $||\mathrm{E}\left[\boldsymbol{R}_i\boldsymbol{R}_i'\boldsymbol{\Sigma}_X\boldsymbol{R}_i\boldsymbol{R}_i'\right]||$ is independent of i which can be observed from (5.13). We can simply plug this expression into (5.28) to obtain

$$\begin{aligned}\mathrm{P}\left(||\boldsymbol{\Delta}|| \geqslant \epsilon\frac{||\mathrm{E}\left[\boldsymbol{R}_i\boldsymbol{R}_i'\boldsymbol{\Sigma}_X\boldsymbol{R}_i\boldsymbol{R}_i'\right]||}{\left(\frac{p}{k}\right)^2||\boldsymbol{\Sigma}_X||}\right) = \\ \leqslant 2p\exp\left(-\epsilon^2 N\frac{||\mathrm{E}\left[\boldsymbol{R}_i\boldsymbol{R}_i'\boldsymbol{\Sigma}_X\boldsymbol{R}_i\boldsymbol{R}_i'\right]||}{4\left(\frac{p}{k}\right)^2||\boldsymbol{\Sigma}_X||}\right)\end{aligned} \tag{5.32}$$

Now, to satisfy (5.27) with high probability, we need the right hand side to be close to zero. If we require for some $\delta \in (0,1)$ that

$$2p\exp\left(-\epsilon^2 N\frac{||\mathrm{E}\left[\boldsymbol{R}_i\boldsymbol{R}_i'\boldsymbol{\Sigma}_X\boldsymbol{R}_i\boldsymbol{R}_i'\right]||}{4\left(\frac{p}{k}\right)^2||\boldsymbol{\Sigma}_X||}\right) \leqslant \delta \tag{5.33}$$

then we should choose the number of samples

$$N \geqslant \frac{4||\boldsymbol{\Sigma}_X||}{\epsilon^2||\mathrm{E}\left[\boldsymbol{R}_i\boldsymbol{R}_i'\boldsymbol{\Sigma}_X\boldsymbol{R}_i\boldsymbol{R}_i'\right]||}\left(\frac{p}{k}\right)^2\log\left(\frac{2p}{\delta}\right) \tag{5.34}$$

For the term in the denominator we know by Theorem 2 that

$$||\mathrm{E}\left[\boldsymbol{R}_i\boldsymbol{R}_i'\boldsymbol{\Sigma}_X\boldsymbol{R}_i\boldsymbol{R}_i'\right]|| = O\left(\frac{p}{k}\right) \tag{5.35}$$

Hence, we need

$$N = O(p\log p) \tag{5.36}$$

draws of the random matrix to obtain results that are close to the bounds of the previous paragraph. 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. (2015a), where it was found that little prediction accuracy is lost by using a finite number of random draws of the subsets.

Random projection regression For random projection regression, similar bounds to the ones we found for random subset regression have been established when $\boldsymbol{R}_i$ is a random projection matrix. The proof in this case is somewhat more involved as one needs additional concentration inequalities to bound the Euclidean norm $||\boldsymbol{R}_i\boldsymbol{R}_i'||$ with high probability. A complete proof of the following theorem can be found in Kabán et al. (2015)

Theorem 6: Let $\boldsymbol{\Sigma}_X$ be a positive semi-definite matrix of size $p \times p$ and rank r. Furthermore, let $\boldsymbol{R}_i$, $i = 1, \dots N$ be independent random projections with $[\boldsymbol{R}_i]_{jk} \sim \frac{1}{\sqrt{k}} N(0,1)$. Define $\boldsymbol{\Delta}$ as in (5.27), then for all $\epsilon \in (0,1)$

$$\begin{aligned} P\left(\boldsymbol{\Delta} \geqslant \epsilon \frac{||\mathrm{E}\left[\boldsymbol{R}_i\boldsymbol{R}_i'\boldsymbol{\Sigma}_X\boldsymbol{R}_i\boldsymbol{R}_i'\right]||}{K}\right) \\ \leqslant 2p \exp\left(-\epsilon^2 N \frac{||\mathrm{E}\left[\boldsymbol{R}_i\boldsymbol{R}_i'\boldsymbol{\Sigma}_X\boldsymbol{R}_i\boldsymbol{R}_i'\right]||}{4K}\right) + 4N \exp\left(-\frac{N^{1/3}}{2}\right) \end{aligned} \tag{5.37}$$

where

$$K = ||\boldsymbol{\Sigma}_X|| \left[\left(1 + \sqrt{\frac{p}{k}}\right) + \frac{1}{\sqrt{k}}\right]^2 \left[\left(\sqrt{\frac{r}{k}} + \sqrt{\frac{p}{k}}\right) + \frac{1}{\sqrt{k}}\right]^2 \tag{5.38}$$

If we neglect the last term of (5.37), then by the same arguments as above it can be shown that the required order of draws is the same as for random subset regression, i.e. $N = O(p \log p)$. The additional term on the right-hand side of (5.37) implies that we need a slightly larger number of draws for random projection regression. In practice however, we found no difference in the behavior for a finite number of draws between the two methods.

5.3 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 mean squared forecast error, 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.

5.3.1 Monte Carlo set-up

The set-up we employ is similar to the one by Elliott et al. (2015a). The data generating process takes the form

$$y_{t+1} = \boldsymbol{x}_t'\boldsymbol{\beta} + \varepsilon_{t+1}, \tag{5.39}$$

where $\boldsymbol{x}_t$ is a $p \times 1$ vector with predictors, $\boldsymbol{\beta}$ a $p \times 1$ coefficient vector, and ε_{t+1} an error term with $\varepsilon_{t+1} \sim N(0, \sigma_\varepsilon^2)$.

In each replication of the Monte Carlo simulations, predictors are generated by drawing $\boldsymbol{x}_t \sim N(\mathbf{0}, \boldsymbol{\Sigma}_X)$, after which we standardize the predictor matrix. The covariance matrix of the predictors equals $\boldsymbol{\Sigma}_X = \frac{1}{p}\boldsymbol{P}'\boldsymbol{P}$, where $\boldsymbol{P}$ is a $p \times p$ matrix whose elements are independently and randomly drawn from a standard normal distribution. As argued by Elliott et al. (2015a), 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 $\boldsymbol{\beta} = \sqrt{\sigma_\varepsilon^2/T} \cdot b\boldsymbol{\iota}_s$ for a fixed constant b The vector $\boldsymbol{\iota}_s$ 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 ε_{T+1} is set to zero, as this only yields an additional noise term σ^2 which is incurred by all forecasting methods.

We employ two sets of experimental designs, which mimick the high-dimensional setting in the empirical application by choosing the number of predictors $p = 100$ and the sample size $T = 200$. Results are based on $M = 10,000$ replications of the data generating process (5.39).

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 5.2.1. In this case we replace $\boldsymbol{x}_t$ in the DGP (5.39) by a subset of the factors extracted from the sample covariance matrix $\frac{1}{T}\sum_{t=1}^{T} \boldsymbol{x}_t\boldsymbol{x}_t'$ using principal component analysis. Denote by $\boldsymbol{f}_i$ for $i = 1, \ldots, p$ 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 $\{\boldsymbol{f}_{46}, \ldots, \boldsymbol{f}_{55}\}$, 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 (5.7) in which we vary the subspace dimension over $k = \{1, \ldots, p\}$. The subspace methods, as well as the benchmark models discussed below, estimate (5.39) 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 5.2.2 which suggest to use $O(p \log p) = O(100 \cdot \log 100) = O(460)$ draws.

Benchmark models We compare the performance of the random methods with principal component regression, and partial least squares regression introduced by Wold (1982). Both methods approximate the data generating process (5.39) as

$$y_{t+1} = \boldsymbol{z}_t' \boldsymbol{\delta}^f + \sum_{i=1}^{k} f_{ti} \beta_i^f + \eta_t \tag{5.40}$$

where $k \in \{1, \dots, p\}$. The methods differ in their construction of the factors f_{ti}. Principal component regression is implemented by extracting the factors from the standardized predictors $\boldsymbol{x}_t$ with $t = 1, \dots, T$ using principal component analysis. We then estimate (5.40) and generate a forecast as $\hat{y}_{T+1} = \boldsymbol{z}_t' \hat{\boldsymbol{\delta}}^f + \sum_{i=1}^{k} f_{Ti} \hat{\beta}_i^f$. Note that for the top factor setting in the second set of experiments, the principal component regression model is thus correctly specified.

Partial least squares uses a two-step procedure to construct the factors, as described by Groen and Kapetanios (2016). We orthogonalize both the standardized predictors $\boldsymbol{x}_t$ and the dependent variable y_{t+1} with respect to $\boldsymbol{z}_t$ for $t = 1, \dots, T-1$. We then calculate the covariance of each predictor x_{it} with y_{t+1} which yields weights $\boldsymbol{w} = \{w_1, \dots, w_p\}$. The first factor is readily constructed as $f_{t1} = \boldsymbol{x}_t' \boldsymbol{w}$. We then orthogonalize x_{it} and y_{t+1} with respect to this factor and repeat the procedure with the corresponding residuals until the required number of factors $f_{t1}, \dots, f_{tk}$ is obtained. To construct a forecast we require $\boldsymbol{f}_T$ for which the above procedure is repeated now taking $t = 1, \dots, T$. Calculating the covariance with y_{T+1} naturally is infeasible, such that the same weights w_i are used as obtained before.

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

$$(\hat{\boldsymbol{\delta}}_k, \hat{\boldsymbol{\beta}}_k) = \arg\min_{\delta, \beta} \left(\frac{1}{T-1} \sum_{t=1}^{T-1} (y_{t+1} - \boldsymbol{z}_t' \boldsymbol{\delta} - \boldsymbol{x}_t' \boldsymbol{\beta})^2 + k P(\boldsymbol{\beta}) \right), \tag{5.41}$$

where $\boldsymbol{z}_t$ includes an intercept. The penalty term $P(\boldsymbol{\beta}) = \sum_{j=1}^{p} \frac{1}{2} \beta_j^2$ in case of ridge regression and $P(\boldsymbol{\beta}) = \sum_{j=1}^{p} |\beta_j|$ 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; $\ln k \in \{-30, \dots, 0\}$ for lasso and $\ln k \in \{-15, \dots, 15\}$ for ridge regression. In general, we expect lasso to do well when the model contains a small number of large

Table 5.1: Monte Carlo simulation: MSFE under optimal subspace dimension

b	RP	RS	PC	PL	RI	LA
			$s = 10$			
0.5	0.966 (2)	0.966 (2)	1.259 (1)	9.698 (1)	0.969 (-3.8)	1.000 (-30.0)
1.0	0.866 (8)	0.867 (8)	1.052 (1)	3.087 (1)	0.860 (-2.3)	0.960 (-28.2)
2.0	0.630 (22)	0.629 (22)	0.953 (7)	0.962 (1)	0.632 (-1.1)	0.648 (-27.6)
			$s = 50$			
0.5	0.831 (10)	0.829 (10)	1.049 (1)	2.492 (1)	0.829 (-2.0)	0.974 (-28.2)
1.0	0.574 (25)	0.574 (25)	0.869 (14)	0.796 (1)	0.579 (-0.8)	0.724 (-27.6)
2.0	0.289 (46)	0.290 (46)	0.428 (43)	0.372 (2)	0.304 (0.5)	0.369 (-26.7)
			$s = 100$			
0.5	0.715 (16)	0.714 (16)	0.998 (1)	1.383 (1)	0.712 (-1.4)	0.872 (-27.9)
1.0	0.436 (35)	0.436 (35)	0.667 (25)	0.535 (1)	0.438 (-0.2)	0.569 (-27.3)
2.0	0.195 (56)	0.195 (56)	0.277 (61)	0.236 (3)	0.200 (0.8)	0.259 (-26.4)

Note: this table reports the MSFE relative to the benchmark of the prevailing mean, for the optimal value of k corresponding to the minimum MSFE which is given in brackets. For additional information, see the note following Figure 5.5

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 mean squared forecast error relative to that of the prevailing mean model that takes $\bar{y}_{T+1} = \frac{1}{T-1}\sum_{t=1}^{T-1} y_{t+1}$. The mean squared forecast error is computed as

$$MSFE = \frac{1}{M}\sum_{j=1}^{M}(y_{T+1}^{(j)} - \hat{y}_{T+1}^{(j)})^2, \tag{5.42}$$

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

5.3.2 Simulation results

Sparsity and signal strength

Table 5.1 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 5.5 in the appendix. The predictive performance of each forecasting method is reported relative to the prevailing mean. Values below one indicate that the benchmark model is outperformed.

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 5.5 in the appendix, show that in case of a weak signal, increasing the subspace dimension 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 decreases 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 projections and random subsets is almost the same. Table 5.1 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, when $b = 2$ the random subspace methods perform better, whereas for very weak signals with $b = 0.5$ ridge regression appears to have a slight edge.

Table 5.1 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 exactly the same subspace dimension. Principal components is observed to select less factors 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 all together. Note that where the parameter k has a 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 5.2. Table 5.6 in the appendix shows results for different values of k. We observe precisely what was anticipated based on the discussion in Section 5.2.1. 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

Table 5.2: Monte Carlo Simulation: optimal subspace dimension under a factor design

b	RP	RS	PC	PL	RI	LA
	Top factor setting					
0.5	0.713 (10)	0.959 (9)	0.952 (3)	2.466 (1)	0.712 (-2.0)	0.861 (-28.2)
1.0	0.421 (21)	0.853 (27)	0.297 (10)	0.501 (1)	0.419 (-1.1)	0.474 (-27.9)
2.0	0.202 (33)	0.573 (60)	0.075 (10)	0.133 (1)	0.202 (-0.5)	0.147 (-27.6)
	Intermediate factor setting					
0.5	1.010 (1)	0.998 (1)	1.489 (1)	16.766 (1)	1.000 (-15.0)	1.000 (-29.7)
1.0	1.002 (1)	0.982 (4)	1.181 (1)	7.034 (1)	1.000 (-6.5)	1.000 (-29.4)
2.0	1.001 (1)	0.916 (16)	1.063 (1)	2.894 (1)	1.000 (-15.0)	1.000 (-30.0)

Note: this table shows the out-of-sample performance of random projection (RP), random subset (RS), principal component (PC), partial least squares (PL), ridge (RI), and lasso (LA) in the Monte Carlo simulations using a factor design and selecting the value of k that yields the minimum MSFE compared to forecasting using the prevalent mean. For additional information, see the note following Table 5.6.

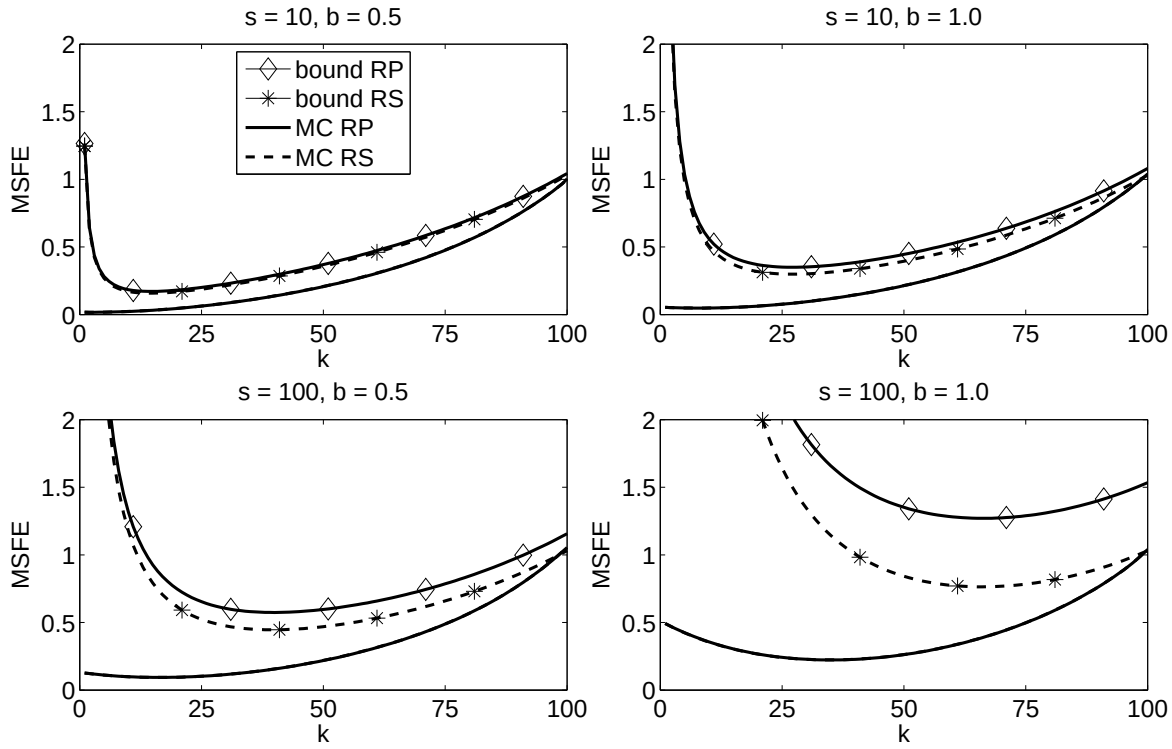
are in a setting where we have a small number of large coefficients. As expected, this favors lasso, although not to the extend that it outperforms principal component regression. The findings are almost completely reversed in the intermediate factor setting, when the nonzero coefficients are associated with factors $\boldsymbol{f}_{46}, \ldots, \boldsymbol{f}_{55}$. 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 brackets in Table 5.2. 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.

5.3.3 Relation between theoretical bounds and Monte Carlo experiments

The qualitative correspondence between the results from the Monte Carlo experiments and the theoretical results show that the bounds are useful to determine settings where the random

Figure 5.1: Monte Carlo simulation: comparison with theoretical bounds



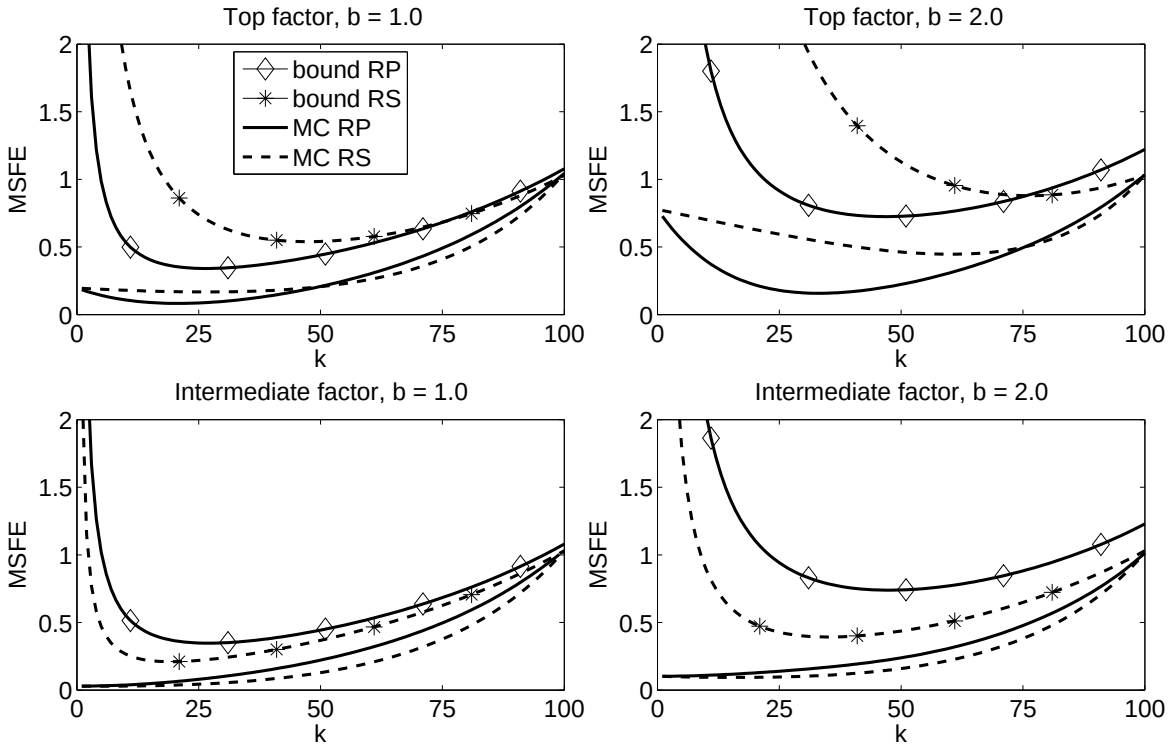
Note: this figure shows the MSFE for different values of the subspace dimension k, along with the theoretical upper bounds on the MSFE derived in Section 5.2.1 after a small sample size correction. The different lines correspond to the upper bound for random projections (bound RP, diamond marker), upper 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 top panels correspond to settings in which the sparsity $s = 10$, while in the bottom panels $s = 100$. The signal to noise ratio parameter $b = 0.5$ in the left panels and $b = 1$ in the right panels.

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 5.1 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 5.2.1, for the first set of experiments described above. As we found in Table 5.5, the values of the MSFE of the random subspace methods are almost identical to each other over the whole range of k. The bounds are closest to the exact MSFE from the Monte Carlo experiments when the signal is not too strong and for large values of k. The bound for random subset regression is tighter than the bound for random projection regression due to the lack of exact orthogonality of the projection matrix. From the Monte Carlo results, it appears that this lack of orthogonality is not a driving force behind the difference between both methods.

In Figure 5.2 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 panel, corresponding to the top factor structure, shows the bound for

Figure 5.2: Monte Carlo simulation: 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 5.2.1 for the top and intermediate factor settings. For additional information, see the note following 5.1.

random projection to be lower. In line with our theoretical results, the optimal subspace dimension for random projection regression is found to be lower. In the lower panel displays the MSFE in the intermediate factor setting. We observe that both the bounds and the exact Monte Carlo results indicate that random subset regression is best suited in this case.

5.4 Empirical application

This section evaluates the predictive performance of the discussed methods in a macroeconomic application.

5.4.1 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 (2015). 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.

5.4.2 Forecasting framework

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

$$y_{t+1} = \boldsymbol{z}_t'\boldsymbol{\delta} + \boldsymbol{x}_t'\boldsymbol{R}_i\boldsymbol{\gamma}_i + u_{t+1},$$

where $\boldsymbol{z}_t$ is a $q \times 1$ vector with predictors which are always included in the model and not subject to the dimension reduction methods, $\boldsymbol{x}_t$ a $p \times 1$ vector with possible predictors, and $\boldsymbol{R}_i$ a $p \times k$ projection matrix. In this application y_{t+1} is one of the macroeconomic time series, $\boldsymbol{z}_t$ includes an intercept along with twelve lags of the dependent variable y_{t+1}, and $\boldsymbol{x}_t$ consists of all 129 remaining variables in the database. The predictors in $\boldsymbol{x}_t$ are projected on a low-dimensional subspace using four different projection methods whose projection matrices are discussed in Section 5.2: random projection regression (RP), random subset regression (RS), principal component regression (PC) and partial least squares (PL). In addition, we again compare the performance to lasso (LA) and ridge regression (RI) as described in Section 5.3.1, as well as to the baseline AR(12) model (AR). Predictive accuracy is measured by the MSFE defined in (5.42).

We use an expanding window to produce 348 forecasts, from January 1985 to December 2014. The initial estimation sample contains 312 observations and runs from January 1960 to December 1984. 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 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(12) model.

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.

Table 5.3: FRED-MD: percentage best predictive performance

		percentage loss							
		RP	RS	PC	PL	RI	LA	AR	All
percentage wins	RP		34.62	84.62	82.31	56.92	56.15	72.31	5.38
	RS	65.38		87.69	81.54	66.92	70.00	73.08	42.31
	PC	15.38	12.31		46.92	16.15	22.31	50.77	5.38
	PL	17.69	17.69	53.08		16.92	20.00	39.23	4.62
	RI	43.08	33.08	83.85	83.08		58.46	72.31	3.85
	LA	43.85	30.00	77.69	80.00	41.54		69.23	20.00
	AR	27.69	26.15	49.23	50.00	27.69	30.77		18.46

Note: this table shows the percentage wins of a method in terms of lowest MSFE compared to other methods separately, and with respect to all other methods (last column). Ties can 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. The percentages are calculated over forecasts for all 130 series in FRED-MD generated by random projections (RP), random subsets (RS), principal components (PC), partial least squares (PL), lasso (LA), ridge regression (RI), and an AR(12) model (AR). The numbers represent the percentage wins of the method listed in the rows over the method listed in the columns.

5.4.3 Empirical results

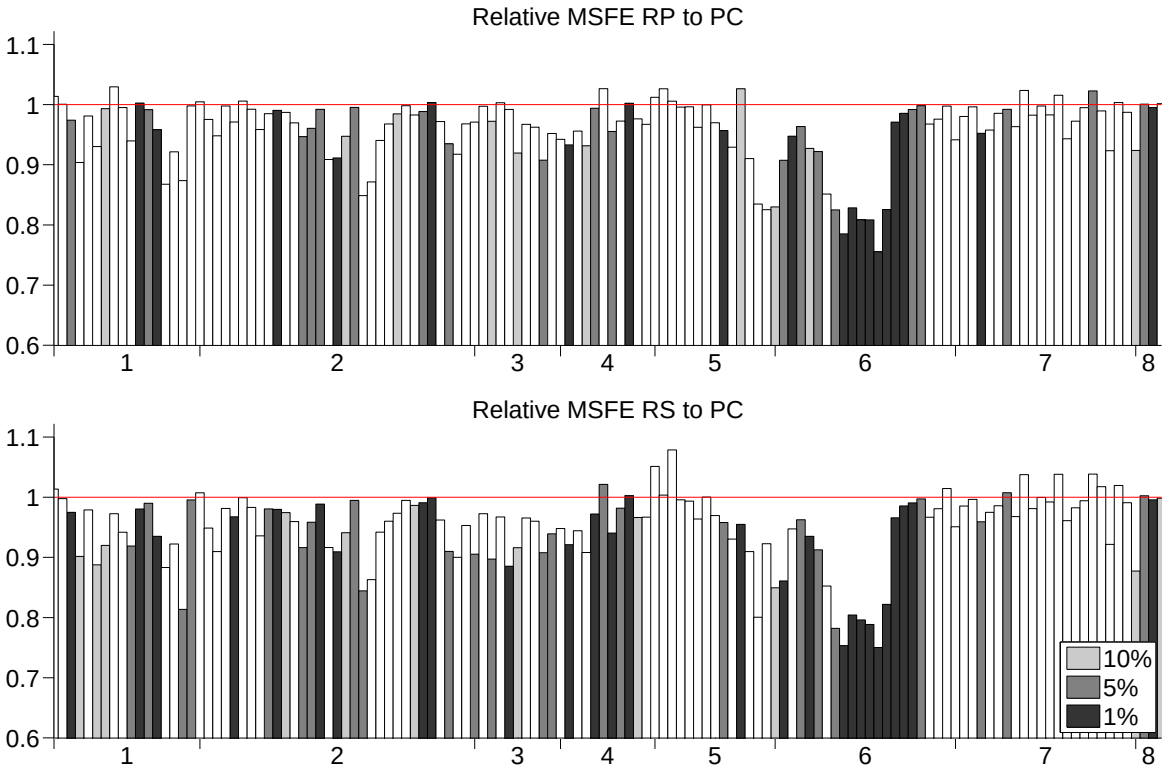
Aggregate statistics

We obtain series of forecasts for 130 macroeconomic variables generated by six different methods. Table 5.3 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 42% of the series. This is a substantial difference with lasso and the AR(12) model that win in approximately 20% of the cases. Random projection, principal component regression, ridge regression and partial least squares score approximately equally well at 5%.

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 5.3 shows again that random subset regression achieves the best results, outperforming the alternatives for at least 65% of the series. Interestingly, its closest 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 predictive performance of the random methods sorted by category, Figure 5.3 shows relative predictive performance compared with princi-

Figure 5.3: FRED-MD: predictive accuracy of random subspace methods compared with PCR



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 (2015). 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.

pal component regression, for all series available in the FRED-MD dataset over the period from January 1985 through December 2014. The MSFE is calculated for the subspace dimension as determined by past predictive performance. The upper panel shows the relative MSFE of random subset regression to principal component regression and the lower panel compares random projection to principal component regression. Values below one, indicate that the random method is preferred over the benchmark. As found in Table 5.3, the random methods outperform the deterministic principal components in most of the cases. For random subset regression this happens in 88% of the cases, which is slightly lower for random projections with 85%. Figure 5.3 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 almost never significant, even at a 10% level. The random methods show the largest improvements in forecast performance in category 6, which contains the interest rate and exchange rate series.

Table 5.4: FRED-MD: predictive accuracy relative to the AR(12)-model

	INDP	UNR	CPI	3TB	Avg.		INDP	UNR	CPI	3TB	Avg.
k	Random projection regression					k	Random subset regression				
k_R	0.955	0.884	0.899	1.123	0.969	k_R	0.912	0.863	0.915	1.255	0.962
1	0.987	0.982	0.993	0.969	0.990	1	0.984	0.976	0.992	0.966	0.987
5	0.955	0.936	0.974	0.934	0.969	5	0.942	0.921	0.974	0.929	0.964
10	0.935	0.906	0.954	0.954	0.962	10	0.917	0.892	0.958	0.952	0.957
15	0.926	0.891	0.938	1.001	0.963	15	0.905	0.878	0.943	0.993	0.957
30	0.921	0.879	0.900	1.184	0.987	30	0.894	0.860	0.908	1.133	0.972
50	0.946	0.902	0.883	1.434	1.049	50	0.902	0.875	0.887	1.323	1.017
100	1.109	1.111	0.976	2.016	1.324	100	1.061	1.083	0.950	1.913	1.278
k	Principal component regression					k	Partial least squares				
k_R	1.027	0.922	0.938	1.360	1.017	k_R	1.027	0.917	0.949	1.224	1.011
1	0.953	0.933	1.014	0.974	1.003	1	0.964	0.917	0.998	0.997	1.011
5	0.955	0.921	0.969	1.136	1.007	5	1.110	1.013	0.943	2.066	1.254
10	0.976	0.924	0.932	1.426	1.019	10	1.162	1.143	0.988	2.285	1.357
15	0.973	0.891	0.946	1.585	1.040	15	1.190	1.181	1.002	2.328	1.415
30	1.007	0.888	0.932	1.732	1.102	30	1.209	1.257	1.030	2.359	1.507
50	1.049	0.961	0.918	1.864	1.178	50	1.243	1.287	1.033	2.447	1.541
100	1.192	1.163	1.012	2.290	1.417	100	1.248	1.305	1.045	2.462	1.541
$\ln k$	Ridge regression					$\ln k$	Lasso				
k_R	0.953	0.881	0.898	1.140	0.974	k_R	0.963	0.888	0.905	1.100	0.979
-6	0.997	0.995	0.998	0.990	0.997	-28	0.956	0.934	0.962	0.953	0.979
-4	0.983	0.973	0.989	0.957	0.985	-27	0.917	0.883	0.891	1.127	0.971
-2	0.936	0.907	0.954	0.956	0.962	-26	0.927	0.901	0.901	1.435	1.024
0	0.927	0.881	0.887	1.287	1.008	-25	1.004	0.979	0.924	1.694	1.126
4	1.118	1.118	0.983	2.056	1.341	-22	1.227	1.280	1.038	2.369	1.514
8	1.261	1.324	1.058	2.464	1.592	-15	1.305	1.390	1.079	2.612	1.639
12	1.305	1.392	1.079	2.606	1.641	-5	1.305	1.392	1.080	2.613	1.641

Note: this table shows the out-of-sample performance of random projections, random subsets, principal components, lasso, and Ridge regression relative to the benchmark of an autoregressive model of order twelve, for different values of subspace dimension k and the recursively selected optimal value of k denoted by k_R. For lasso and ridge regression, the penalty parameter runs over a grid of values k. The predictive accuracy is reported for the dependent variables industrial production (INDP), unemployment rate (UNR), inflation (CPI), three month treasury bill rate (3TB), and the average over the mean squared forecast errors for all series. The predictive accuracy is measured by relative MSFE, which equals values below one when the particular method outperforms the benchmark model.

A case study of four key macroeconomic indicators

We look more closely into the predictive performance of the different methods on four key macroeconomic indicators: industrial production index (INDP), unemployment rate (UNR), inflation (CPI), and the three-month Treasury Bill rate (3mTB). In Table 5.4 we show the MSFE relative to the AR(12) model for different values of the subset 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. 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 the three month Treasury bill rate, where it outperforms random subset, but not random projection regression.

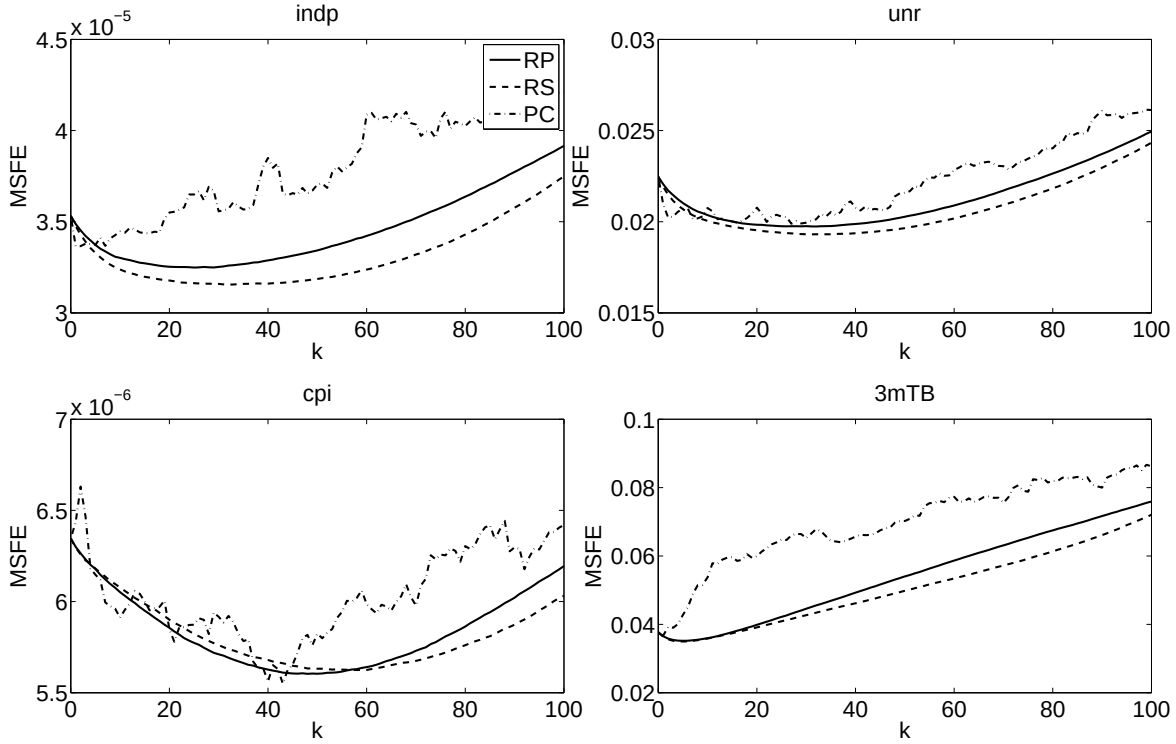
With regard to the lasso and ridge regression benchmarks, the results show that on average, these methods are outperformed by both random subset and random projection regression. For the individual series reported here, the evidence is mixed. Random subset regression outperforms both lasso and ridge on industrial production and the unemployment rate series, while the situation is reversed on the inflation and treasury bill rate. Random projection has a slight edge when predicting the treasury bill rate, but is close to ridge regression, which is in line with our findings in Section 5.3, and lasso on all four series.

Table 5.4 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(12) benchmark model for almost all subspace dimensions, even for very large values of k. Compared to PC and PL, we again see that the random methods select much larger values of k.

To visualize the dependence on k for the different projection methods, Figure 5.4 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 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 5.4 confirms that the random methods reach their minimum for relatively large values of k as discussed in Section 5.2. 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 a single factor, while the random methods reach their minimum when using a subspace of dimension 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 methods seems to be lower for

Figure 5.4: FRED-MD: predictive accuracy for different subspace dimensions



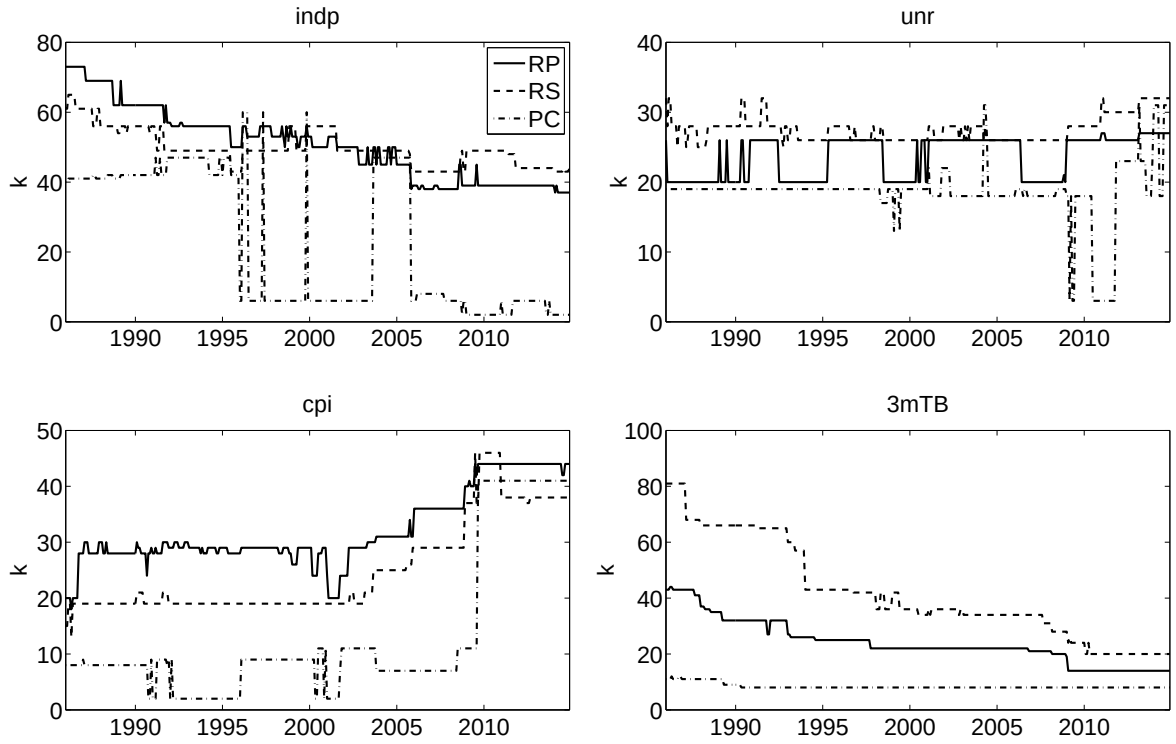
Note: this figure shows the MSFE for different values of the subspace dimension k. The different lines correspond to the evaluation criterium for the dimension reduction methods random projection (RP, solid), random subset (RS, dashed), and principal component regression (PC, dotted). The models at $k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

most values of k, except for inflation where a large principal component model yields more accurate results.

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. We found in Figure 5.4 that the minimum MSFE is lower for random subset than for random projection regression for all four series but inflation. However, the MSFE of the treasury bill rate corresponding to the recursively selected optimal value of k is lower for random projections while for all fixed k random subsets perform better. This shows that the selection of k plays an important role in the practical predictive performance of the methods.

Figure 5.5 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 has been gradually decreasing over time. While starting at a very large

Figure 5.5: FRED-MD: recursive selection of subspace dimensions

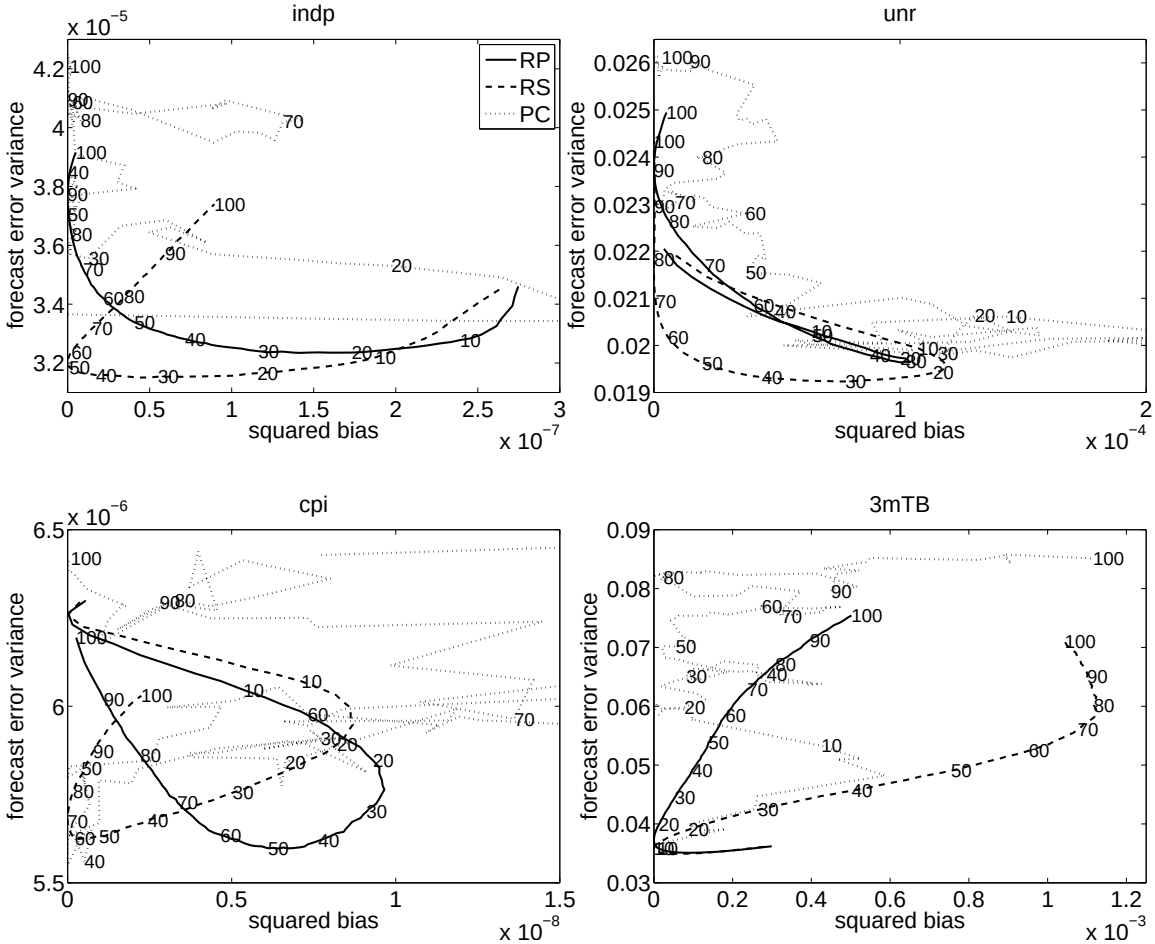


Note: this figure shows the selection of subset dimension k. The different lines correspond to the dimension reduction methods random projection (RP, solid), random subset (RS, dashed), and principal component regression (PC, dotted). At each point in time the subset dimension is selected based on its 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 (3mTB).

dimension around 70 in 1985, this has since dropped to values around 40. A minor effect of the global financial crisis is observed on random subset regression. 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 large 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.

The dimension reduction methods are expected to trade of bias and variance when the subspace dimension k varies. One would typically expect the forecast variance to be decreasing with k, while the bias is increasing with k. Figure 5.6 plots the bias-variance trade-off of the dimension reduction methods. It is immediately clear that for PC, the behavior is very erratic. Although in general a large number of factors translates into a larger forecast

Figure 5.6: FRED-MD: bias-variance trade-off



Note: this figure plots the forecast error variance against the squared bias for different values of the subspace dimension k. The different lines correspond to the dimension reduction methods random projection (RP, solid), random subset (RS, dashed), and principal component (PC, dotted) regression. The four panels correspond to four dependent variables, industrial production (INDP), unemployment rate (UNR), inflation (CPI), and the three month treasury bill rate (3mTB).

variance, this increase is by no means uniform. For random subset regression and random projection regression, we find values for k where both the variance and the bias are smaller relative to principal components, explaining the better performance of the random method. The relationship between forecast error variance and squared bias follows a much smoother pattern over k for the random methods. Nevertheless, it is striking that also for both random methods the forecast error variance does not monotonically increase in k, and the bias not automatically declines with increasing subspace dimension. This observation is explained by the fact that the forecasts are constructed as averages over draws of projection matrices. The reported forecast error variance only includes the 'explained' part of the variance, the variance over the averaged predictions. However, there is also an unexplained part, due to the variance over the predictions within the averages. Appendix 5.D shows that the sum of the explained and unexplained part, the total forecast error variance, increases in the subspace

dimension, but due to the variance from the draws of the projection matrix, the observed forecast error variance can be decreasing in k.

5.5 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 tight bounds on their 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 compare the predictive accuracy to several widely used benchmarks: principal component regression, partial least squares, lasso regularization and ridge regression. The performance increases for nearly all settings under consideration compared to principal component regression and lasso regularization. Compared to ridge regression, we find large differences when we impose a factor structure on the model. When nonzero coefficients are associated with factors that explain most of the variance, random projection regression gives results very similar to ridge regression, but random subset regression is clearly outperformed. On the other hand, when the nonzero coefficients are associated with intermediate factors, random subset regression is the only method that is capable of beating the historical mean.

In the application, it seems this last scenario is prevalent, with random subset regression providing more accurate forecasts in 45% of the series. In method-by-method comparison, it outperforms the benchmarks in no less than 67% of the series. It also outperforms random projection regression in 65% of the cases. Random projection regression itself is more accurate than the benchmarks in at least 56% of the series.

5.A Proof of Theorem 1

We start by noting that by Jensen's inequality

$$\mathrm{E}\left[\left(\boldsymbol{x}_T'\boldsymbol{\beta} - \boldsymbol{x}_T'\mathrm{E}_{R_i}\left[\boldsymbol{R}_i\hat{\boldsymbol{\gamma}}_i\right]\right)^2\right] \leqslant \mathrm{E}_{R_i}\mathrm{E}\left[\left(\boldsymbol{x}_T'\boldsymbol{\beta} - \boldsymbol{x}_T'\boldsymbol{R}_i\hat{\boldsymbol{\gamma}}_i\right)^2 \middle| \boldsymbol{R}_i\right] \tag{5.43}$$

Furthermore, since by assumption $\mathrm{E}[\boldsymbol{x}_T\boldsymbol{x}_T'] = \boldsymbol{\Sigma}_X$ and $\hat{\gamma}_i$ is independent of $\boldsymbol{x}_T$ we have that

$$\begin{aligned}&\mathrm{E}\left[\left(\boldsymbol{x}_T'\boldsymbol{\beta} - \boldsymbol{x}_T'\boldsymbol{R}_i\hat{\boldsymbol{\gamma}}_i\right)^2\right] \\ &\quad = \mathrm{E}\left[\left(\boldsymbol{\beta} - \boldsymbol{R}_i\hat{\boldsymbol{\gamma}}_i\right)'\boldsymbol{\Sigma}_X\left(\boldsymbol{\beta} - \boldsymbol{R}_i\hat{\boldsymbol{\gamma}}_i\right)\right] + o_p(T^{-1})\end{aligned} \tag{5.44}$$

For the MSFE, we now have

$$\begin{aligned}&\mathrm{E}\left[\left(\boldsymbol{x}_T'\left(\boldsymbol{\beta} - \mathrm{E}_{R_i}\left[\boldsymbol{R}_i\hat{\boldsymbol{\gamma}}_i\right]\right)\right)^2\right] = \\ &\quad \leqslant \mathrm{E}_{R_i}\mathrm{E}\left[\left|\left|\boldsymbol{\Sigma}_X^{1/2}\left(\boldsymbol{\beta} - \boldsymbol{R}_i\hat{\boldsymbol{\gamma}}_i\right)\right|\right|^2\middle|\,\boldsymbol{R}_i\right] + o_p(T^{-1}) \\ &\quad = \mathrm{E}_{R_i}\mathrm{E}\left[\left|\left|\boldsymbol{\Sigma}_X^{1/2}\left(\boldsymbol{\beta} - \boldsymbol{R}_i\boldsymbol{\gamma}_i - \boldsymbol{R}_i(\hat{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_i)\right)\right|\right|^2\middle|\,\boldsymbol{R}_i\right] + o_p(T^{-1}) \\ &\quad = \mathrm{E}_{R_i}||\boldsymbol{\Sigma}_X^{1/2}(\boldsymbol{\beta} - \boldsymbol{R}_i\boldsymbol{\gamma}_i)||^2 + \mathrm{E}_{R_i}\mathrm{E}\left[\left|\left|\boldsymbol{\Sigma}_X^{1/2}\boldsymbol{R}_i(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_i)\right|\right|^2\middle|\,\boldsymbol{R}_i\right] \\ &\qquad - 2\mathrm{E}_{R_i}\mathrm{E}\left[\left(\boldsymbol{\beta} - \boldsymbol{R}_i\boldsymbol{\gamma}_i\right)'\boldsymbol{\Sigma}_X\boldsymbol{R}_i(\hat{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_i)\middle|\,\boldsymbol{R}_i\right] + o_p(T^{-1})\end{aligned} \tag{5.45}$$

The parameter $\boldsymbol{\gamma}_i$ is estimated by OLS and we have

$$\boldsymbol{X}\boldsymbol{R}_i(\hat{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_i) = \boldsymbol{P}_{XR_i}\boldsymbol{X}(\boldsymbol{\beta} - \boldsymbol{R}_i\boldsymbol{\gamma}_i) + \boldsymbol{P}_{XR_i}\boldsymbol{\varepsilon} \tag{5.46}$$

where $\boldsymbol{P}_{XR_i}$ denotes the projection matrix on the subspace spanned by the columns of $\boldsymbol{X}\boldsymbol{R}_i$. The crucial step, observed in Kabán (2014), is that $\boldsymbol{\gamma}_i$ is the optimal parameter vector in the low-dimensional subproblem, defined as

$$\boldsymbol{\gamma}_i = \arg\min_{\boldsymbol{u}} \sum_{t=1}^{T-1}\left(x_t\beta - \boldsymbol{x}_t'\boldsymbol{R}_i\boldsymbol{u}\right)^2 \tag{5.47}$$

This implies the following inequality

$$||\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{X}\boldsymbol{R}_i\boldsymbol{\gamma}_i||^2 \leqslant ||\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{X}\boldsymbol{R}_i\boldsymbol{R}_i'\boldsymbol{\beta}||^2 \tag{5.48}$$

Substituting (5.46) and (5.48) into (5.45) and using that $\frac{1}{T}\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{\Sigma}_X + o_p(T^{-1})$ we obtain

$$\begin{aligned}&E\left[\left(\boldsymbol{x}_T'\boldsymbol{\beta} - \boldsymbol{x}_T'\mathrm{E}_{R_i}\left[\boldsymbol{R}_i\hat{\boldsymbol{\gamma}}_i\right]\right)^2\right] \\ &\quad \leqslant \sigma^2\frac{k}{T} + \mathrm{E}_{R_i}\left[\left(\boldsymbol{\beta} - \boldsymbol{R}_i\boldsymbol{\gamma}_i\right)'\boldsymbol{\Sigma}_X\left(\boldsymbol{\beta} - \boldsymbol{R}_i\boldsymbol{\gamma}_i\right)\right] \\ &\qquad - \mathrm{E}_{R_i}\left[\left|\left|\boldsymbol{P}_{\boldsymbol{\Sigma}_X^{1/2}R_i}\boldsymbol{\Sigma}_X^{1/2}\left(\boldsymbol{\beta} - \boldsymbol{R}\boldsymbol{\gamma}_i\right)\right|\right|^2\right] + o_p(T^{-1}) \\ &\quad \leqslant \sigma^2\frac{k}{T} + \mathrm{E}_{R_i}\left[\boldsymbol{\beta}'(\boldsymbol{I} - \boldsymbol{R}_i\boldsymbol{R}_i')\boldsymbol{\Sigma}_X(\boldsymbol{I} - \boldsymbol{R}_i\boldsymbol{R}_i')\boldsymbol{\beta}\right] \\ &\qquad - \mathrm{E}_{R_i}\left[\left|\left|\boldsymbol{P}_{\boldsymbol{\Sigma}_X^{1/2}R_i}\boldsymbol{\Sigma}_X^{1/2}\left(\boldsymbol{\beta} - \boldsymbol{R}_i\boldsymbol{\gamma}_i\right)\right|\right|^2\right] + o_p(T^{-1})\end{aligned} \tag{5.49}$$

Finally, (5.47) has a simple solution

$$\boldsymbol{\gamma}_i = \left(\frac{1}{T-1} \sum_{t=1}^{T-1} \boldsymbol{R}_i' \boldsymbol{x}_t \boldsymbol{x}_t' \boldsymbol{R}_i \right)^{-1} \left(\frac{1}{T-1} \sum_{t=1}^{T-1} \boldsymbol{R}_i' \boldsymbol{x}_t \boldsymbol{x}_t' \boldsymbol{\beta} \right) \tag{5.50}$$

Hence

$$\boldsymbol{\Sigma}_X^{1/2} (\boldsymbol{\beta} - \boldsymbol{R}_i \boldsymbol{\gamma}_i) = \left(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{\Sigma}_X^{1/2} R_i} \right) \boldsymbol{\Sigma}_X^{1/2} \boldsymbol{\beta} \tag{5.51}$$

which shows that the last term of (5.49) is identically zero.

5.B Derivation of equation (5.19)

For the difference between the MSFE under random projection and orthogonalized random projection we have that

$$\begin{aligned}
\Delta &= \frac{1}{k} \left[\boldsymbol{\beta}' \boldsymbol{\Sigma}_X \boldsymbol{\beta} + \text{trace}(\boldsymbol{\Sigma}_X) \boldsymbol{\beta}' \boldsymbol{\beta} \right] \\
&\quad - \frac{p-k}{k} \frac{1}{p^2-1} \left[p \cdot \text{trace}(\boldsymbol{\Sigma}_X) \boldsymbol{\beta}' \boldsymbol{\beta} - \boldsymbol{\beta}' \boldsymbol{\Sigma}_X \boldsymbol{\beta} \right] \\
&= \frac{1}{k} \left[\boldsymbol{\beta}' \boldsymbol{\Sigma}_X \boldsymbol{\beta} \left(1 + \frac{p-k}{p^2-1} \right) - \text{trace}(\boldsymbol{\Sigma}_X) ||\boldsymbol{\beta}||^2 \frac{kp-1}{p^2-1} \right] \\
&\geqslant \frac{\boldsymbol{\beta}' \boldsymbol{\Sigma}_X \boldsymbol{\beta}}{k} \frac{p^2 - 1 + p - k - kp + 1}{p^2 - 1} \\
&= \frac{\boldsymbol{\beta}' \boldsymbol{\Sigma}_X \boldsymbol{\beta}}{k} \frac{p-k}{p-1} \\
&\geqslant 0
\end{aligned} \tag{5.52}$$

In the third line we use the fact that $\boldsymbol{\Sigma}_X = \boldsymbol{U} \boldsymbol{\Lambda} \boldsymbol{U}'$ with $\boldsymbol{U}$ an orthogonal matrix and Λ a diagonal matrix consisting of the non-negative eigenvalues of $\boldsymbol{\Sigma}_X$. Then

$$\begin{aligned}
\boldsymbol{\beta}' \boldsymbol{\Sigma}_X \boldsymbol{\beta} - \text{trace}(\boldsymbol{\Sigma}_X) ||\boldsymbol{\beta}||^2 &= \boldsymbol{\beta}' (\boldsymbol{\Sigma}_X - \text{trace}(\boldsymbol{\Sigma}_X)) \boldsymbol{\beta} \\
&= \boldsymbol{\beta}' \boldsymbol{U} \left[\boldsymbol{\Lambda} - \left(\sum_{i=1}^{p} \lambda_i \right) \boldsymbol{I} \right] \boldsymbol{U}' \boldsymbol{\beta} \\
&\leqslant 0
\end{aligned} \tag{5.53}$$

where the last inequality holds since each term on the diagonal satisfies $\lambda_i - \sum_{j=1}^{p} \lambda_j = -\sum_{j \neq i} \lambda_j < 0$.

5.C Optimal bounds

The optimal, but infeasible, choice of k that minimizes the bounds is given by

$$\begin{aligned} k^*_{RSR} &= \left[\frac{T}{\sigma^2} p \frac{p}{p-1}\left(\boldsymbol{\beta}'\boldsymbol{D}_{\Sigma_X}\boldsymbol{\beta} - \frac{1}{p}\boldsymbol{\beta}\boldsymbol{\Sigma}_X\boldsymbol{\beta}\right)\right]^{1/2} \\ k^*_{RP} &= \left[\frac{T}{\sigma^2}\left(\boldsymbol{\beta}'\boldsymbol{\Sigma}_X\boldsymbol{\beta} + \text{trace}(\boldsymbol{\Sigma}_X)\boldsymbol{\beta}'\boldsymbol{\beta}\right)\right]^{1/2} \end{aligned} \tag{5.54}$$

The optimal choice of k leads to the following bound for random subset regression

$$\begin{aligned} &E\left[\left(\boldsymbol{x}_T'\boldsymbol{\beta} - \boldsymbol{x}_T'\mathrm{E}^{RS}_{R_i}\left[\boldsymbol{R}_i\hat{\boldsymbol{\gamma}}_i\right]\right)^2\right] = \\ &\quad\leqslant 2\left[\frac{\sigma^2}{T} p \frac{p}{p-1}\left(\boldsymbol{\beta}'\boldsymbol{D}_{\Sigma_X}\boldsymbol{\beta} - \frac{1}{p}\boldsymbol{\beta}\boldsymbol{\Sigma}_X\boldsymbol{\beta}\right)\right]^{1/2} \\ &\qquad - \frac{p}{p-1}\sum_{j=1}^{p}\left(\boldsymbol{\beta}'\boldsymbol{D}_{\Sigma_X}\boldsymbol{\beta} - \frac{1}{p}\boldsymbol{\beta}\boldsymbol{\Sigma}_X\boldsymbol{\beta}\right) + o_p(T^{-1}) \end{aligned} \tag{5.55}$$

For random projection regression under $k = k^*_{RP}$ we have

$$\begin{aligned} &E\left[\left(\boldsymbol{x}_T'\boldsymbol{\beta} - \boldsymbol{x}_T'\mathrm{E}^{RP}_{R_i}\left[\boldsymbol{R}_i\hat{\boldsymbol{\gamma}}_i\right]\right)^2\right] = \\ &\quad\leqslant 2\left[\frac{\sigma^2}{T}\left(\boldsymbol{\beta}'\boldsymbol{\Sigma}_X\boldsymbol{\beta} + \text{trace}(\boldsymbol{\Sigma}_X)\boldsymbol{\beta}'\boldsymbol{\beta}\right)\right]^{1/2} + o_p(T^{-1}) \end{aligned} \tag{5.56}$$

5.D Application: bias-variance tradeoff

The mean squared forecast error can be decomposed in a bias and a variance component:

$$\mathrm{E}\left[(y_{t+1} - \mathrm{E}_{R_i}[\hat{y}^i_{T+1}])^2\right] = \mathrm{E}\left[y_{t+1} - \mathrm{E}_{R_i}[\hat{y}^i_{T+1}]\right]^2 + \mathrm{Var}\left[y_{t+1} - \mathrm{E}_{R_i}[\hat{y}^i_{T+1}]\right]$$

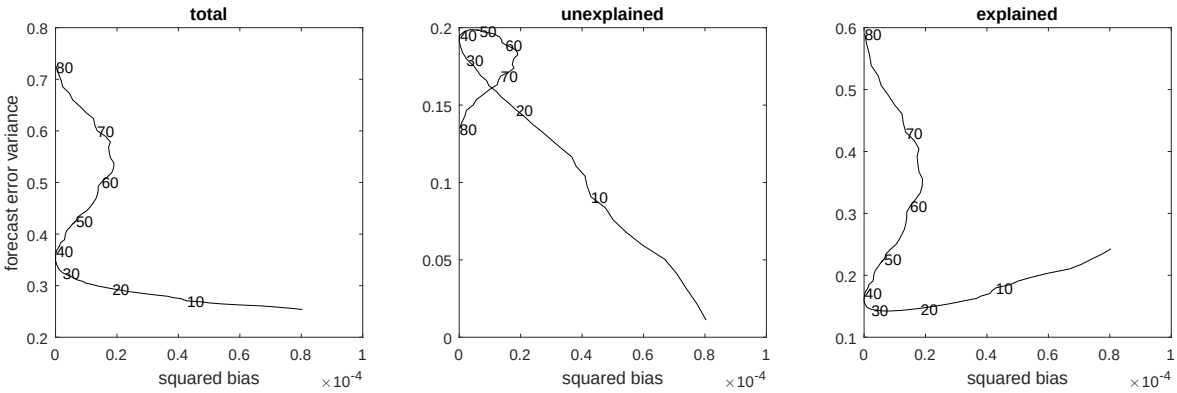
The first term equals the squared bias of the forecasts and the second term the forecast error variance. However, since we average over realizations of R_i, the second term only includes the explained component of the forecast error variance. This can be illustrated by applying the law of total variance on the forecast error of all generated predictions:

$$\mathrm{Var}\left[y_{t+1} - \hat{y}^i_{T+1}\right] = \mathrm{E}\left[y_{t+1} - \mathrm{Var}_{R_i}[\hat{y}^i_{T+1}]\right] + \mathrm{Var}\left[y_{t+1} - \mathrm{E}_{R_i}[\hat{y}^i_{T+1}]\right]$$

where the left term equals the unexplained and the right term the explained component of the forecasts error variance.

Because of computational constraints, we do not store predictions for all different projection matrices in the empirical application. Hence, we setup a Monte Carlo experiment

Figure 5.7: Bias-variance Trade-off



Note: this figure plots the forecast error variance against the squared bias for different values of the subspace dimension k. The three panels show the total forecasts error variance, the unexplained and the explained part. The forecasts are generated by random projections in a simulation design as discussed in Section 5.3.1, where we us $M = 1000$ replications with $b = 1$ and $s = 50$.

to investigate the behaviour of the unexplained and explained components of the forecast error variance. The simulation design is a small scale version of the experiments explained in Section 5.3.1, where we use $M = 1000$ replications with $b = 1$ and $s = 50$ to generate forecasts with random projection regressions. Figure 5.7 shows the bias-variance trade-off for the total variance and the unexplained and explained variance components. The total variance behaves as expected; the forecast error variance increase with the subspace dimension k. The unexplained component shows unpredictable behaviour, which causes that the explained variance is not always increasing in k. The third panel of Figure 5.7 shows similar patterns as we find in Figure 5.6, which shows the bias-variance trade-off in the empirical example. The empirical findings can be explained by the fact that the reported forecast error variance leaves out the unexplained part, leading to a forecast error variance that can decrease for larger subspace dimensions.

Table 5.5: Monte Carlo simulation: MSFE relative to prevailing mean

		Random projections - k				Random subsets - k			
s	b	1	10	25	50	1	10	25	50
10	0.5	0.977	1.291	3.584	11.861	0.977	1.301	3.626	11.938
	1.0	0.968	0.875	1.382	3.873	0.967	0.876	1.396	3.889
	2.0	0.964	0.732	0.635	1.091	0.964	0.729	0.635	1.096
50	0.5	0.965	0.831	1.188	3.160	0.965	0.829	1.196	3.174
	1.0	0.963	0.716	0.574	0.885	0.962	0.714	0.574	0.889
	2.0	0.962	0.682	0.408	0.293	0.961	0.679	0.406	0.293
100	0.5	0.964	0.756	0.781	1.668	0.963	0.753	0.782	1.673
	1.0	0.962	0.697	0.473	0.512	0.962	0.693	0.472	0.513
	2.0	0.961	0.678	0.386	0.202	0.961	0.674	0.384	0.202
		Principal components - k				Partial least squares - k			
s	b	1	10	25	50	1	10	25	50
10	0.5	1.259	3.883	8.929	19.732	9.698	41.613	50.135	52.279
	1.0	1.052	1.696	3.143	6.385	3.087	13.005	15.610	16.278
	2.0	0.990	0.961	1.085	1.733	0.962	3.455	4.192	4.408
50	0.5	1.049	1.477	2.584	5.231	2.492	10.157	12.189	12.732
	1.0	0.979	0.886	0.941	1.416	0.796	2.781	3.371	3.525
	2.0	0.960	0.733	0.518	0.438	0.438	0.679	0.821	0.864
100	0.5	0.998	1.097	1.493	2.761	1.383	5.241	6.326	6.642
	1.0	0.971	0.783	0.667	0.790	0.535	1.345	1.621	1.703
	2.0	0.959	0.690	0.451	0.287	0.371	0.335	0.424	0.444
		Ridge regression - $\ln k$				Lasso - $\ln k$			
s	b	-6	-4	-2	0	-28	-27	-26	-25
10	0.5	0.993	0.972	1.370	7.359	1.526	6.449	16.136	27.703
	1.0	0.990	0.948	0.873	2.370	0.998	2.239	4.950	8.326
	2.0	0.989	0.937	0.707	0.818	0.677	0.818	1.475	2.378
50	0.5	0.990	0.945	0.829	1.981	1.006	1.953	4.111	6.877
	1.0	0.988	0.934	0.685	0.689	0.760	0.803	1.257	1.911
	2.0	0.985	0.917	0.599	0.306	0.516	0.374	0.404	0.521
100	0.5	0.989	0.940	0.741	1.115	0.872	1.197	2.225	3.585
	1.0	0.988	0.929	0.648	0.449	0.671	0.569	0.720	1.007
	2.0	0.982	0.900	0.546	0.218	0.425	0.281	0.262	0.300

Note: this table shows the MSFE divided by that of the prevailing mean forecast, for random projection regression, random subset regression, principal component regression, partial least squares, lasso, and ridge regression under the data generating process (5.39) 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 5.6: 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
	0.5	0.942	0.713	1.217	3.872	0.992	0.959	1.145	2.599
Top	1.0	0.936	0.552	0.438	1.062	0.991	0.917	0.854	1.053
	2.0	0.935	0.510	0.230	0.287	0.990	0.903	0.764	0.595
	0.5	1.010	1.834	5.749	19.192	0.998	1.213	2.797	11.190
Int.	1.0	1.002	1.299	2.735	7.629	0.993	1.015	1.497	4.435
	2.0	1.001	1.068	1.363	2.336	0.990	0.929	0.947	1.558
		Principal components - k				Partial least squares - k			
s	b	1	10	25	50	1	10	25	50
	0.5	0.976	1.082	2.774	6.390	2.466	13.681	16.341	17.293
Top	1.0	0.901	0.297	0.745	1.719	0.501	3.704	4.393	4.602
	2.0	0.883	0.075	0.192	0.449	0.133	0.936	1.125	1.181
	0.5	1.489	5.917	14.398	32.184	16.766	66.078	78.234	82.060
Int.	1.0	1.181	2.943	6.388	12.876	7.034	24.611	29.615	31.166
	2.0	1.063	1.637	2.722	4.077	2.894	7.410	8.587	8.970
		Ridge regression - $\ln k$				Lasso - $\ln k$			
s	b	-6	-4	-2	0	-28	-27	-26	-25
	0.5	0.983	0.908	0.712	2.296	2.367	5.358	9.181	13.919
Top	1.0	0.981	0.891	0.517	0.680	0.737	1.516	2.582	3.879
	2.0	0.976	0.867	0.417	0.226	0.201	0.391	0.648	0.968
	0.5	1.001	1.025	1.931	11.236	10.792	25.880	45.308	68.811
Int.	1.0	1.000	1.007	1.340	4.761	4.749	10.264	17.290	25.895
	2.0	1.000	1.002	1.083	1.774	1.772	3.053	4.856	7.162

Note: this table shows the out-of-sample performance of random projection regression (RP), random subset regression (RS), principal component regression (PC), partial least squares (PL), ridge regression (RI), and lasso (LA) in the Monte Carlo simulations when the underlying model has a factor structure. In the experiments referred to with 'Top', 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 $\{\boldsymbol{f}_{46}, \ldots, \boldsymbol{f}_{55}\}$. For additional information, see the note following Table 5.5.

Nederlandse Samenvatting (Summary in Dutch)

Anticiperen op economische ontwikkelingen is essentieel voor beleidsmakers, ondernemers, investeerders en andere economische spelers. De afgelopen jaren hebben opnieuw bewezen dat het lastig is deze ontwikkelingen nauwkeurig te voorspellen op basis van historische data. Steeds weer lijkt de economie in een unieke, nieuwe fase terecht te komen. Zo zijn de huidige kapitaalinjecties van centrale banken historisch ongeëvenaard, evenals het recente voornemen van Groot-Brittannië om uit de Europese Unie te stappen. De vraag is dan ook: hoe kunnen we in deze instabiele wereld betrouwbare voorspellingen construeren?

Een optie is om economische modellen steeds flexibeler te maken. Zo kunnen modellen er rekening mee houden dat macro-economische relaties over de tijd veranderen. Zelfs als de complexe modellen beter bij de data passen, leidt een toename in flexibiliteit echter niet in alle gevallen tot een toename van de voorspelnauwkeurigheid. Hoe complexer het model, hoe groter namelijk ook de statistische onzekerheid rondom de voorspellingen. Het vinden van een optimale uitruil van door de data gevraagde modelcomplexiteit en de resulterende statistische onzekerheid, speelt de hoofdrol in dit proefschrift.

Het eerste deel van dit proefschrift richt zich op voorspellingen uit modellen die wisselende economische fases beschrijven. Dit kan gaan om terugkerende fases, zoals recessies en expansies, maar ook om eenmalige structurele veranderingen. Hoofdstuk 2 beschrijft hoe de voorspelnauwkeurigheid vergroot kan worden door datapunten anders te wegen. In hoofdstuk 3 testen we of een model dat de structurele verandering expliciet modelleert, betere voorspellingen op zal leveren dan een simpel model dat deze verandering negeert.

Met de toenemende hoeveelheid data is een belangrijke vraag: helpt dit met het beschrijven en voorspellen van economische ontwikkeling? Voor macro-economische data is dit niet meteen duidelijk. Nu er steeds meer variabelen beschikbaar zijn, is de vraag: welke van deze variabelen zijn ook echt relevant? Alle variabelen meenemen lijkt onverstandig, maar een verkeerde selectie heeft ook een negatief effect op de voorspelnauwkeurigheid. In het tweede deel van dit proefschrift worden verschillende technieken geïntroduceerd, die de mogelijkheid bieden om met een grote hoeveelheid data toch nauwkeurige schattingen te maken (hoofdstuk 4) en voorspellingen te doen (hoofdstuk 5).

Deel 1: voorspellen in wisselende economische omstandigheden

Een typische eigenschap van macro-economische data is dat de ontwikkeling door de tijd heen niet constant is. Het bruto binnenlands product kent periodes van groei, afgewisseld door, meestal kortere, periodes van krimp. In andere gevallen veranderen economische omstandigheden door een eenmalige gebeurtenis, zoals een natuurramp, een wisseling in overheidsbeleid of de ontdekking van een waardevolle grondstof.

Er zijn verschillende modellen ontwikkeld die deze veranderingen nauwkeurig lijken te kunnen vatten. Helaas voorspellen deze modellen in de praktijk vaak minder goed dan simpeler varianten, die de variatie over de tijd compleet negeren. Een potentiële verklaring is dat we met deze modellen veel vragen van de data. We willen niet alleen informatie over de gemiddelde groei in de afgelopen decennia, maar ook over het verschil in groei in recessies en in expansies. Omdat we niet precies weten wanneer de economie van een expansieperiode overgaat in een recessie, moet het model ook zelf bepalen wanneer wisselingen plaatsvinden. Deze toename in complexiteit leidt tot meer onzekerheid in de voorspellingen. Het is essentieel om deze toegenomen onzekerheid mee te wegen in een voorspelprocedure.

Het meestgebruikte model om overgangen tussen recessies en expansies te beschrijven, is het Markov switching model. Een mooie eigenschap van Markov switching modellen is dat ze een waarschijnlijkheid aangeven waarmee de economie zich op een bepaald moment in de tijd in een recessie of expansie bevond. Hoofdstuk 2 van dit proefschrift laat zien hoe deze waarschijnlijkheid gebruikt kan worden om de voorspelnauwkeurigheid te vergroten. Het blijkt effectief om de waarschijnlijkheid te overdrijven, en daarmee de verschillende economische periodes te benadrukken. Dit gebeurt door waarnemingen anders te wegen. Met het gebruik van deze nieuwe wegingsmethode blijken de Markov switching modellen wel degelijk in staat beter te voorspellen dan simpeler alternatieven.

De methode in hoofdstuk 2 past achteraf de voorspellingen van het model aan. Een andere optie is om vooraf te testen of je het complexe model wel nodig hebt. In hoofdstuk 3 ontwikkelen we daarom een statistische test die informatie geeft of we een structurele verandering nauwkeurig genoeg kunnen modelleren, zodat de resulterende voorspellingen beter zijn dan wanneer we de verandering simpelweg negeren. Dit verschilt van de huidige testen, die vragen of er überhaupt een verandering plaatsvindt. We vinden dat zelfs als er een verandering plaatsvindt, deze niet altijd groot genoeg is om relevant te zijn voor de voorspellingen. De onzekerheid over de precieze timing van de verandering blijkt hierin cruciaal. Is deze onzekerheid groot, dan leidt dit ook tot een grote onzekerheid in de voorspellingen. Er kan dan beter een simpel model gebruikt worden. In een macro-economische toepassing, vinden we dat twee tot drie keer vaker dan bestaande testen suggereren, een simpel model nauwkeuriger voorspellingen oplevert.

Deel 2: schatten en voorspellen met hoog-dimensionale data

De laatste jaren komen steeds rijkere datasets beschikbaar. We kunnen daarom van meer en meer variabelen de invloed meten op economische ontwikkelingen. Als we de data voor ons zien als een spreadsheet waarbij de rijen waarnemingen over de tijd bevatten, en de kolommen de verschillende variabelen, dan zien we bij macroeconomische data vooral een groei in de breedte. De reden is dat de frequentie waarmee de variabelen gemeten wordt vaak maandelijks is, of zelfs op kwartaalbasis. Het aantal verschillende statistieken dat wordt bijgehouden stijgt echter sneller. Het is dan ingewikkeld om van individuele variabelen nauwkeurig het effect te schatten.

Hoofdstukken 5 en 6 beschrijven twee verschillende technieken die deze nauwkeurigheid vergroten. In hoofdstuk 5 wordt een techniek ontwikkeld die de schattingen zodanig aanpast dat met een vooraf ingestelde minimale waarschijnlijkheid, de nauwkeurigheid van individuele schattingen vergroot wordt. Deze waarschijnlijkheid impliceert een trade-off: hoe hoger je de waarschijnlijkheid kiest om de nauwkeurigheid te vergroten, hoe kleiner de potentiële winst. Als we geïntereseerd zijn in de gemiddelde nauwkeurigheid over een groep schattingen, dan gaat met deze methode de nauwkeurigheid echter in alle gevallen omhoog.

Hoofdstuk 6 richt zich weer op voorspellingen. Het analyseert een nieuwe dimensiereductie techniek, die erop gericht is om de brede datasets ‘smaller’ te maken. Stel we hebben een dataset met 100 variabelen. Bestaande methodes proberen de informatie in deze data zo goed mogelijk samen te vatten in een veel kleiner aantal variabelen. Een veel simpeler methode om deze dataset te versmallen is om willekeurig 10 van de 100 variabelen te kiezen. Op basis van de informatie in deze 10 variabelen construeer je vervolgens een voorspelling. Door dit te herhalen met steeds willekeurig 10 variabelen eindig je met een reeks voorspellingen. Verrassend genoeg blijkt dat het gemiddelde van deze reeks een zeer nauwkeurige voorspelling oplevert. Hoofdstuk 6 geeft een theoretische verklaring voor dit op het eerste gezicht tegen-intuïtieve resultaat. In een grootschalige toepassing op 130 macroeconomische tijdreeksen laten we bovendien zien dat de theoretische nauwkeurigheid inderdaad ook in de praktijk bevestigd wordt.

Bibliography

Ahlswede, R. and Winter, A. (2002). Strong converse for identification via quantum channels. *IEEE Transactions on Information Theory*, 48(3):569–579.

Andrews, D. W. (1993). Tests for parameter instability and structural change with unknown change point. *Econometrica*, 61(4):821–856.

Andrews, D. W. and Ploberger, W. (1994). Optimal tests when a nuisance parameter is present only under the alternative. *Econometrica*, 62(6):1383–1414.

Ang, A. and Bekaert, G. (2002). Regime switches in interest rates. *Journal of Business & Economic Statistics*, 20(2):163–182.

Bai, J. and Perron, P. (1998). Estimating and testing linear models with multiple structural changes. *Econometrica*, 66(1):47–78.

Boot, T. (2015). Controlled shrinkage and variable selection. *Working paper*.

Boot, T. and Nibbering, D. (2016). Forecasting using random subspace methods. *Working paper*.

Boot, T. and Pick, A. (2016a). A near-optimal test for structural breaks when forecasting under mean squared error loss. *Working paper*.

Boot, T. and Pick, A. (2016b). Optimal forecasts from Markov switching models. *Journal of Business & Economic Statistics*, forthcoming.

Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics*, 37(4):373–384.

Burr, I. W. (1942). Cumulative frequency functions. *Annals of Mathematical Statistics*, 13(2):215–232.

Burr, I. W. and Cislak, P. J. (1968). On a general system of distributions: I. Its curve-shape characteristics; II. The sample median. *Journal of the American Statistical Association*, pages 627–635.

Chiong, K. X. and Shum, M. (2016). Random projection estimation of discrete-choice models with large choice sets. *USC-INET Research Paper*, (16-14).

Chow, G. C. (1960). Tests of equality between sets of coefficients in two linear regressions. *Econometrica: Journal of the Econometric Society*, pages 591–605.

Clark, T. E. and McCracken, M. W. (2001). Tests of equal forecast accuracy and encompassing for nested models. *Journal of Econometrics*, 105(1):85–110.

Clark, T. E. and McCracken, M. W. (2012). In-sample tests of predictive ability: A new approach. *Journal of Econometrics*, 170(1):1–14.

Clark, T. E. and McCracken, M. W. (2013). Advances in forecast evaluation. In Elliott, G. and Timmermann, A., editors, *Handbook of Forecasting*, volume 2, pages 1107–1201. Elsevier.

Clements, M. P. and Krolzig, H.-M. (1998). A comparison of the forecast performance of Markov-switching and threshold autoregressive models of US GNP. *Econometrics Journal*, 1(1):47–75.

Crawford, G. W. and Fratantoni, M. C. (2003). Assessing the forecasting performance of regime-switching, ARIMA and GARCH models of house prices. *Real Estate Economics*, 31(2):223–243.

Dacco, R. and Satchell, S. (1999). Why do regime-switching models forecast so badly? *Journal of Forecasting*, 18(1):1–16.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, B*, 39(1):1–38.

Deschamps, P. J. (2008). Comparing smooth transition and Markov switching autoregressive models of US unemployment. *Journal of Applied Econometrics*, 23(4):435–462.

Diebold, F. X. and Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3):253–263.

Donoho, D. L. and Johnstone, J. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455.

Efron, B. and Morris, C. (1973). Stein's estimation rule and its competitors: an empirical bayes approach. *Journal of the American Statistical Association*, 68(341):117–130.

Elliott, G., Gargano, A., and Timmermann, A. (2013). Complete subset regressions. *Journal of Econometrics*, 177(2):357–373.

Elliott, G., Gargano, A., and Timmermann, A. (2015a). Complete subset regressions with large-dimensional sets of predictors. *Journal of Economic Dynamics and Control*, 54:86–110.

Elliott, G. and Müller, U. K. (2007). Confidence sets for the date of a single break in linear time series regressions. *Journal of Econometrics*, 141(2):1196–1218.

Elliott, G. and Müller, U. K. (2014). Pre and post break parameter inference. *Journal of Econometrics*, 180(2):141–157.

Elliott, G., Müller, U. K., and Watson, M. W. (2015b). Nearly optimal tests when a nuisance parameter is present under the null hypothesis. *Econometrica*, 83(2):771–811.

Engel, C. (1994). Can the Markov switching model forecast exchange rates? *Journal of International Economics*, 36(1):151–165.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360.

Giacomini, R. and Rossi, B. (2009). Detecting and predicting forecast breakdowns. *Review of Economic Studies*, 76(2):669–705.

Groen, J. J. and Kapetanios, G. (2016). Revisiting useful approaches to data-rich macroeconomic forecasting. *Computational Statistics & Data Analysis*, 100:221–239.

Guhaniyogi, R. and Dunson, D. B. (2015). Bayesian compressed regression. *Journal of the American Statistical Association*, 110(512):1500–1514.

Guidolin, M. (2011). Markov switching models in empirical finance. *Advances in Econometrics*, 27:1–86.

Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, 57(2):357–384.

Hamilton, J. D. (1994). *Time Series Analysis*. Princeton University Press, Princeton.

Hansen, B. E. (2015). The risk of James–Stein and Lasso shrinkage. *Econometric Reviews*, pages 1–15.

Hansen, B. E. (2016). Efficient shrinkage in parametric models. *Journal of Econometrics*, 190(1):115–132.

Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica*, 46(6):1251–1271.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Hüsler, J. (1990). Extreme values and high boundary crossings of locally stationary Gaussian processes. *Annals of Probability*, 18(3):1141–1158.

James, W. and Stein, C. (1961). Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 361–379.

Johnson, W. B. and 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. In *AISTATS*, pages 448–456.

Kabán, A., Bootkrajang, J., and Durrant, R. J. (2015). Toward large-scale continuous EDA: A random matrix theory perspective. *Evolutionary computation*.

Kapetanios, G. and Marcellino, M. (2010). Factor-gmm estimation with large sets of possibly weak instruments. *Computational Statistics & Data Analysis*, 54(11):2655–2675.

Kim, C.-J. (1994). Dynamic linear models with Markov-switching. *Journal of Econometrics*, 60(1):1–22.

Kim, C.-J. and Nelson, C. R. (1999). Has the US economy become more stable? a Bayesian approach based on a Markov-switching model of the business cycle. *Review of Economics & Statistics*, 81(4):608–616.

Klaassen, F. (2005). Long swings in exchange rates: Are they really in the data? *Journal of Business & Economic Statistics*, 23(1):87–95.

Koop, G., Korobilis, D., and Pettenuzzo, D. (2016). Bayesian compressed vector autoregressions. *Available at SSRN 2754241*.

Krolzig, H.-M. (1997). *Markov-switching Vector Autoregressions: Modelling, Statistical Inference, and Application to Business Cycle Analysis*. Springer Verlag, Berlin.

Krolzig, H.-M. (2000). Predicting Markov-switching vector autoregressive processes. Nuffield College *Working Paper*, W31.

Laforgia, A. and Natalini, P. (2010). Some inequalities for modified Bessel functions. *Journal of Inequalities and Applications*, 2010(1):253035.

Long, J. S. and Ervin, L. H. (2000). Using heteroscedasticity consistent standard errors in the linear regression model. *American Statistician*, 54(3):217–224.

MacKinnon, J. G. and White, H. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics*, 29(3):305–325.

Magnus, J. R. (2002). Estimation of the mean of a univariate normal distribution with known variance. *Econometrics Journal*, 5(1):225–236.

Magnus, J. R. and Durbin, J. (1996). A classical problem in linear regression or how to estimate the mean of a univariate normal distribution with known variance. *CENTER Working paper No. 9660*.

Maillard, O. and Munos, R. (2009). Compressed least-squares regression. In *Advances in Neural Information Processing Systems*, pages 1213–1221.

Maruyama, Y. and Strawderman, W. E. (2005). Necessary conditions for dominating the James-Stein estimator. *Annals of the Institute of Statistical Mathematics*, 57(1):157–165.

Marzetta, T. L., Tucci, G. H., and 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. and Ng, S. (2015). FRED-MD: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics*, (forthcoming).

Meese, R. A. and Rogoff, K. (1983). Empirical exchange rate models of the seventies: Do they fit out of sample? *Journal of international economics*, 14(1):3–24.

Paye, B. S. and Timmermann, A. (2006). Instability of return prediction models. *Journal of Empirical Finance*, 13(3):274–315.

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.

Perez-Quiros, G. and Timmermann, A. (2001). Business cycle asymmetries in stock returns: Evidence from higher order moments and conditional densities. *Journal of Econometrics*, 103(1):259–306.

Pesaran, M. H., Pick, A., and Pranovich, M. (2013). Optimal forecasts in the presence of structural breaks. *Journal of Econometrics*, 177(2):134–152.

Pesaran, M. H. and Timmermann, A. (2005). Small sample properties of forecasts from autoregressive models under structural breaks. *Journal of Econometrics*, 129(1):183–217.

Pesaran, M. H. and Timmermann, A. (2007). Selection of estimation window in the presence of breaks. *Journal of Econometrics*, 137(1):134–161.

Piterbarg, V. I. (1996). *Asymptotic Methods in the Theory of Gaussian Processes and Fields*, volume 148. American Mathematical Soc.

Ploberger, W., Krämer, W., and Kontrus, K. (1989). A new test for structural stability in the linear regression model. *Journal of Econometrics*, 40(2):307–318.

Rapach, D. and Zhou, G. (2013). Forecasting stock returns. In Elliott, G. and Timmermann, A., editors, *Handbook of Economic Forecasting,* Vol. 2A, chapter 6, pages 328–383. Elsevier, Amsterdam.

Rapach, D. E. and Wohar, M. E. (2006). Structural breaks and predictive regression models of aggregate U.S. stock returns. *Journal of Financial Econometrics*, 4(2):238–274.

Rossi, B. (2006). Are exchange rates really random walks? some evidence robust to parameter instability. *Macroeconomic Dynamics*, 10(1):20–38.

Rossi, B. and Inoue, A. (2012). Out-of-sample forecast tests robust to the choice of window size. *Journal of Business & Economics Statistics*, 30(3):432–453.

Schneider, M. J. and 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.

Sen, P. K. (1989). The mean-median-mode inequality and noncentral chi square distributions. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 106–114.

Stamey, T. A., Kabalin, J. N., McNeal, J. E., Johnstone, I. M., Freiha, F., Redwine, E. A., and Yang, N. (1989). Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate. II. radical prostatectomy treated patients. *The Journal of Urology*, 141(5):1076–1083.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley Symposium on mathematical statistics and probability*, volume 1, pages 197–206.

Stock, J. H. and Watson, M. W. (1996). Evidence on structural instability in macroeconomic time series relations. *Journal of Business & Economic Statistics*, 14(1):11–30.

Stock, J. H. and 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. and Watson, M. W. (2006). Forecasting with many predictors. *Handbook of economic forecasting*, 1:515–554.

Stock, J. H. and Watson, M. W. (2012). Generalized shrinkage methods for forecasting using many predictors. *Journal of Business & Economic Statistics*, 30(4):481–493.

Thompson, J. R. (1968). Some shrinkage techniques for estimating the mean. *Journal of the American Statistical Association*, 63(321):113–122.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 58(1):267–288.

Toro-Vizcarrondo, C. and Wallace, T. D. (1968). A test of the mean square error criterion for restrictions in linear regression. *Journal of the American Statistical Association*, 63(322):558–572.

Trenkler, G. and Toutenburg, H. (1992). Pre-test procedures and forecasting in the regression model under restrictions. *Journal of Statistical Planning and Inference*, 30(2):249–256.

Tucci, G. H. and Wang, K. (2011). New methods for handling singular sample covariance matrices. *arXiv preprint arXiv:1111.0235*.

Wallace, T. D. (1972). Weaker criteria and tests for linear restrictions in regression. *Econometrica*, 40(4):689–698.

Welch, I. and Goyal, A. (2008). A comprehensive look at the empirical performance of equity premium prediction. *Review of Financial Studies*, 21(4):1455–1508.

Wold, H. (1982). Soft modelling: the basic design and some extensions. *Systems under indirect observation, Part II*, pages 36–37.

Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.

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:

627. X. CAI, Essays in Labor and Product Market Search
628. L. ZHAO, Making Real Options Credible: Incomplete Markets, Dynamics, and Model Ambiguity
629. K. BEL, Multivariate Extensions to Discrete Choice Modeling
630. Y. ZENG, Topics in Trans-boundary River sharing Problems and Economic Theory
631. M.G. WEBER, Behavioral Economics and the Public Sector
632. E. CZIBOR, Heterogeneity in Response to Incentives: Evidence from Field Data
633. A. JUODIS, Essays in Panel Data Modelling
634. F. ZHOU, Essays on Mismeasurement and Misallocation on Transition Economies
635. P. MULLER, Labor Market Policies and Job Search
636. N. KETEL, Empirical Studies in Labor and Education Economics
637. T.E. YENILMEZ, Three Essays in International Trade and Development
638. L.P. DE BRUIJN, Essays on Forecasting and Latent Values
639. S. VRIEND, Profiling, Auditing and Public Policy: Applications in Labor and Health Economics
640. M.L. ERGUN, Fat Tails in Financial Markets
641. T. HOMAR, Intervention in Systemic Banking Crises
642. R. LIT, Time Varying Parameter Models for Discrete Valued Time Series
643. R.H. KLEIJN, Essays on Bayesian Model Averaging using Economic Time Series
644. S. MUNS, Essays on Systemic Risk
645. B.M. SADABA, Essays on the Empirics of International Financial Markets

646. H. KOC, Essays on Preventive Care and Health Behaviors

647. V.V.M. MISHEVA, The Long Run Effects of a Bad Start

648. W. LI, Essays on Empirical Monetary Policy

649. J.P. HUANG, Topics on Social and Economic Networks

650. K.A. RYSZKA, Resource Extraction and the Green Paradox: Accounting for Political Economy Issues and Climate Policies in a Heterogeneous World

651. J.R. ZWEERINK, Retirement Decisions, Job Loss and Mortality

652. M. K. KAGAN, Issues in Climate Change Economics: Uncertainty, Renewable Energy Innovation and Fossil Fuel Scarcity

653. T.V. WANG, The Rich Domain of Decision Making Explored: The Non-Triviality of the Choosing Process

654. D.A.R. BONAM, The Curse of Sovereign Debt and Implications for Fiscal Policy

655. Z. SHARIF, Essays on Strategic Communication

656. B. RAVESTEIJN, Measuring the Impact of Public Policies on Socioeconomic Disparities in Health

657. M. KOUDSTAAL, Common Wisdom versus Facts; How Entrepreneurs Differ in Their Behavioral Traits from Others

658. N. PETER, Essays in Empirical Microeconomics

659. Z. WANG, People on the Move: Barriers of Culture, Networks, and Language

660. Z. HUANG, Decision Making under Uncertainty-An Investigation from Economic and Psychological Perspective

661. J. CIZEL, Essays in Credit Risk, Banking, and Financial Regulation

662. I. MIKOLAJUN, Empirical Essays in International Economics

663. J. BAKENS, Economic Impacts of Immigrants and Ethnic Diversity on Cities

664. I. BARRA, Bayesian Analysis of Latent Variable Models in Finance

665. S. OZTURK, Price Discovery and Liquidity in the High Frequency World

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