A Tradeoff in Econometrics

A Tradeoff in Econometrics

Een evenwichtsoefening in econometrie

Thesis

to obtain the degree of Doctor from the
Erasmus Universiteit 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 defence shall be held on
Friday, June 8, 2018 at 09:30 hrs

by

Victor Hoornweg
born in Amsterdam, The Netherlands.

Erasmus University Rotterdam

Doctoral Committee

Promotors:	Prof. dr. P.H.B.F. Franses
	Prof. dr. R. Paap
Other members:	Prof. dr. H.P. Boswijk
	Prof. dr. M.J.C.M. Verbeek
	Dr. A. Pick

Tinbergen Institute
ISBN 978 90 3610 513 2

Design: Crasborn Graphic Designers bno, Valkenburg a.d. Geul.

This book is no. 717 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.

Contents

1

Introduction

In the accumulative process of science, man's knowledge of the underlying truth is continually refined by confronting theoretical conjectures to empirical data. An essential task of Statistics is to enable researchers to anticipate and control how hypotheses will be influenced by data-optimized estimates of a new experiment. As I will illustrate below, current statistical procedures make it difficult for researchers to balance the in-sample accuracy of data-optimization with the simplicity of sticking to prior hypotheses. One of the main goals of this thesis is to present a general approach towards controlling such an Accuracy-Simplicity Tradeoff ('AST').

This topic will be explored within the field of Econometrics, because this discipline is primarily concerned with developing techniques for estimating parameters of the underlying data generating process. The linear regression model is the workhorse of Econometrics. The model posits that a dependent variable y and independent variables X are linearly related through unknown coefficients β and an error term ϵ.

In truth, that is, data are generated according to

$$y = X\beta + \epsilon,$$

with an $N \times 1$ vector y of dependent observations, an $N \times K$ matrix of regressors X, a $K \times 1$ vector of coefficients β, and an $N \times 1$ vector of disturbances ϵ. The latter term captures the effects on y that cannot be explained through $X\beta$. For a given sample of data, the true but unknown parameters β can be estimated by b to give

$$y = Xb + e,$$

where $e = y - Xb$ represent the residuals. Estimates of β that are fully dependent on the data can be obtained by minimizing the sum of squared residuals $e'e$.

In the next three chapters I will focus on how the complexity of data-optimized solutions can be reduced when estimating linear regression coefficients. Complex methods are more flexible in selecting parameters and are therefore more likely to capture random noise rather than the actual underlying parameters. This could worsen forecasting performance as well as our understanding of the true model. At the expense of in-sample accuracy, a model's simplicity can be increased by shrinking parameters towards prior hypotheses β_0. This is the first AST that I spoke of just now. When regressors are highly correlated, their parameters can also be stimulated to have a similar deviance from β_0. In this second AST, in-sample accuracy is balanced with the simplicity of grouping parameters together.

Bayesian and Frequentist statistics hardly enable a researcher to control these ASTs. Their tuning parameters for making the first tradeoff between the data-optimized parameters and the prior β_0 can have values ranging from zero to infinity, and it is often unclear to what degree parameter estimates change in case a value of 0.1 is used instead of a 1000, for example. In Frequentist methods like Ridge regression, it only becomes evident *a posteriori* what degree of shrinkage towards β_0 is associated with a given tuning parameter. Bayesians can try to better anticipate how a prior will be balanced with a data-optimized solution by rescaling each regressor, but they often resort to 'uninformative' priors to avoid this cumbersome process. Regarding the second AST, methods have been developed which either emphasize subset selection, grouping of correlated parameters, or both; but none of these techniques differentiate between high and low cross-correlations. As a result, the deletion of irrelevant regressors and the grouping of highly correlated regressors are not performed effectively. The added effect of small cross-correlations can lead irrelevant regressors to deviate considerably from $\beta_0 = 0$, for example.

As an alternative, I will develop an *ast*imator whereby the researcher can directly indicate through a tuning parameter λ how much influence data-optimization should have relative to a reference setup of prior hypotheses. When regressors are uncorrelated, the prior coefficient will at least have an influence of $\lambda \cdot 100\%$ in estimating regression coefficients. The degree to which a parameter is further shrunk towards the prior is determined by the regressor's contribution to R^2 accuracy. With a second tuning parameter $c_{\min}$, the researcher will be able to specify how high cross-correlations between regressors need to be for their parameters to be grouped together. Next to establishing an effective grouping, this also ensures that irrelevant deviations from β_0 are not permitted. The astimator that I will develop in Chapter 2 makes use of an ℓ_2 norm in measuring deviations from β_0 and has an analytic solution.

In Chapter 3, astimators with an ℓ_1 norm will be constructed that enable the researcher to perform *exact* subset selection, which means that parameters of irrelevant regressors are equated exactly to β_0 even when λ has not reached its maximum value yet. I will provide astimated versions of well-known frequentist shrinkage methods with an ℓ_1 norm. The interpretation of the moment that a regressor is activated (allowed to deviate from β_0) has been an enigma for the latter techniques, as a result of which the researcher has not been able to anticipate and influence to what extent data-optimized solutions are penalized. I will show that these transition points are directly related to a regressor's contribution to the R^2 measure of fit when regressors are uncorrelated. I will introduce an ℓ_1 astimator that effectively performs grouping and exact subset selection and combine this astimator with an ℓ_2 norm to further promote grouping.

The out-of-sample performances of the different estimators and astimators will be assessed in Chapter 4. Here, I will discuss how the tuning parameter λ can be selected with the help of cross-validation and information criteria. In the former case, it will be shown that a researcher's own λ_0 can easily be balanced with a cross-validated alternative. When applying information criteria, one has to specify the model's effective number of parameters $\mathscr{K}$, or the 'effective degrees of freedom' as it has also been called (Hastie et al., 2009). Since there is no undisputed method available for measuring $\mathscr{K}$, I will offer a plain but effective solution. Astimators penalize in-sample accuracy with a relative simplicity term, and I will argue that this relative simplicity term can be used as an astimator's measure for the effective number of parameters. To apply cross-validation or an information criterion, the researcher must also specify a set of candidate λ values from which the optimal one is chosen. Up till now, such candidate sets often had to be readjusted *a posteriori*. Astimators help to overcome this obstacle as well, because they make the effect of λ easier to anticipate.

Until Chapter 5, it is assumed that there are no breaks in the underlying data generating process. How should model parameters be estimated if we relax this restriction of coefficients being fixed over time? One strategy is to estimate the break date and use post-break data. The best starting point method ('SPB') makes use of cross-validation to determine the timing of the break point. The data is split in a validation sample of recent observations and a training sample of more distant observations. Model parameters are estimated with the training sample, and these estimates are used to 'predict' the outcomes in the validation sample. By varying the starting point of a data set with which the model is estimated, one can select the starting point with the best pseudo predictions. SPB can conveniently be applied to a broad range of techniques, but it also has a number of drawbacks. It is slow to respond to a new break, it discards old information too easily, and it only considers assigning positive weights to post-break observations.

In Chapter 5, I will attempt to improve upon these three aspects. In the process, I will develop an algorithm which adaptively combines discrete and exponential weights to give robust estimates of the underlying breaks and parameters. The algorithm selects multiple candidate break points in the first step, assigns weights to the resulting periods in the second step, and shrinks these weights to equal or exponential weights in the third step. Forecast errors in the validation window are weighed exponentially to respond more quickly to recent forecasting errors. Central to the method is that deviations from equal weights are intuitively penalized with the same Accuracy Simplicity Tradeoffs as before. I will explain the difference between using an ℓ_1 and an ℓ_2 norm in penalizing deviations from equal weights and derive a measure for the effective number of parameters that can be used when applying an information criterion.

In Chapter 6, I will further study how techniques for finding optimal configurations, like cross-validation and information criteria, can be performed more efficiently. In a typical grid search, configurations are equally spread across the given dimensions after which all combinations of configurations between the dimensions are evaluated. A random search aims at distributing configurations equally across the configuration space by selecting candidates from a uniform distribution. A more sophisticated approach starts with a random search, and then iteratively estimates which set of configurations results in the largest Expected Improvement with the help of a stochastic model. The grid and random searches are inefficient because they do not take into account that groups of configurations may result in highly similar forecasts and because they fail to focus on known good areas. The Expected Improvement approach is inefficient because it takes a long time to estimate the stochastic model.

As an alternative, I will present a global to local approach towards choosing candidate configurations that is simple, quick, and accurate. The basic idea is to start by selecting the middle of two configurations whose forecasts are on average the most dissimilar and to gradually tip the balance towards choosing configurations based on the average accuracy between neighboring configurations. This search procedure can be applied when there are multiple statistical choices to be optimized over and when there are multiple (local) minima.

Astimating regression parameters, weighing observations, and efficiently selecting configurations are the main applications that this thesis about tradeoffs in econometrics entails. The dissertation consists of single-authored chapters only. It has benefited from the comments of my supervisors, Prof. dr. Philip Hans Franses and Prof. dr. Richard Paap, for which I owe them my gratitude.

Although econometric approaches may vary in how they make claims about the underlying truth, there is widespread agreement with regard to the general steps that ought to be taken when doing research. The scientist starts with a research question, derives hypotheses based on previous knowledge, and defines methods for evaluating the theoretical conjectures. Next, he collects a random sample of data and applies the methods on the data to assess the main hypotheses, while holding auxiliary hypotheses fixed. Finally, the researcher discusses which inferences can and cannot be drawn and suggests how future research could overcome possible limitations of the study at hand.

The above procedure is known as the scientific method. This conception of science is not without its problems and these will be further examined in my forthcoming book *Science: Under Submission*. The chapters of the current PhD thesis are written in accordance with predominant norms of science. The statistical techniques that are presented here will make it easier for researchers to specify in advance how they wish to balance prior hypotheses with the possible findings of a new data set.

2

Accuracy-Simplicity Tradeoffs and the Linear Regression Model: b_2 Astimators

2.1. Introduction

A simple model has few parameters to be estimated at a given point in time and parameters that do not alter across time. Complex models are more flexible in optimizing over in-sample accuracy and are therefore more liable to confuse the underlying process with random noise. At the cost of in-sample accuracy, simplicity can be achieved by penalizing deviations from a given scheme. This *Accuracy-Simplicity Tradeoff* ('AST') is fundamental to statistics and I aim to control it when selecting parameters of the linear regression model, so that the researcher can better specify and anticipate how parameters will be estimated. To simplify the choice of linear regression parameters, one can urge them to stay close to a prior coefficient β_0 or to stay close to each other. I will explore both possibilities.

Bayesian and Frequentist estimators make it difficult to control the first AST of balancing the in-sample accuracy of a data-optimized regression coefficient and the simplicity of a prior coefficient β_0. In Bayesian regression, the researcher has to rescale each regressor in some sensible manner to turn the prior variance into a measure of trust regarding β_0. Alternatives to this strenuous process are to use uninformative priors or Zellner's g-prior. In the former, the AST is completely nullified; and in the latter, the degree of shrinkage towards prior coefficients is controlled with no regard for the data. Frequentist shrinkage techniques have been developed as well, like Ridge regression, Lasso, Adaptive Lasso, and the Elastic Net. These methods are typically sensitive to the choice of parameterization (Smith and Campbell, 1980, Leamer, 1981), so that one cannot even remotely anticipate how a tuning parameter influences the AST.

As an alternative, I will introduce a class of methods called linear regression '*ast*imators'. Astimators allow a researcher to specify through λ how large a relative increase in accuracy must be for a relative decrease in simplicity to be allowed. When regressors are uncorrelated, $\lambda \cdot 100\%$ specifies the minimum degree of shrinkage towards β_0 in percentage terms. Relative accuracy is directly defined in terms of R^2, which corresponds to the well-known 'coefficient of determination' for $\beta_0 = 0$. The lower a regressor's contribution to R^2 accuracy, the more it will be shrunk towards β_0.

In this way, the first AST promotes subset selection, so that only those parameters are allowed to deviate from β_0 whose contribution to R^2 accuracy is sufficiently large. This can be contrasted to Ridge regression (Hoerl and Kennard, 1970), which does not perform subset selection at all. When regressors are uncorrelated, this method shrinks the unrestricted solutions towards β_0 by the same degree; and when regressors are correlated, its parameters are stimulated to have a similar deviance from β_0. Such a grouping of parameters is another way of reducing a model's dimensionality and helps to diversify risks among correlated regressors.

Bayesian and Frequentist estimators can be refined in dealing with this second instigation of an AST, where the freedom to optimize over in-sample accuracy is restricted by the simplicity of grouping parameters together. These estimators do not differentiate between high and low cross-correlations among regressors. The implication for estimators that are mainly oriented to the first AST, like the Adaptive Lasso, is that they will only select a single regressor from a group of highly correlated regressors. Such a risky strategy might deteriorate forecasting performance and could prevent researchers from identifying truly relevant regressors (Chapter 3). Estimators that indiscriminately emphasize grouping of parameters, like Ridge regression, have a tendency to let irrelevant regressors substantially deviate from β_0 even when cross-correlations are low.

In this chapter, the main goal is to reduce the complexity of the linear regression model by shrinking coefficients towards β_0 and by grouping parameters of highly correlated regressors together. I will focus on procedures that employ an ℓ_2 norm in penalizing deviations from prior coefficients. One astimator will be introduced that performs subset selection, one that groups regressors, and one that does both. The latter is called a b_{2c} astimator, where the c stands for *correlated* variables being controlled and the 2 refers to an ℓ_2 norm being used. The b_{2c} astimator has a straightforward analytic expression. The tuning parameter $\lambda \in [0,1]$ controls deviations from β_0, and through a second tuning parameter $c_{\min} \in [0,1]$, the researcher can specify how high cross-correlations need to be for parameters to be grouped together.

Regarding the organization of this chapter, Section 2.2 discusses benchmark estimators with an ℓ_2 norm and Section 2.3 presents astimators with an ℓ_2 norm. The behavior of astimators is illustrated with simulation studies and an empirical application in Section 2.4. Section 2.5 concludes.

2.2. Bayesian Regression, Zellner's g-prior, and Ridge Regression

The linear regression model can be defined as

$$y = X\beta + \epsilon,$$

where y is an $N \times 1$ dependent variable, X is an $N \times K$ matrix of independent variables, β is a $K \times 1$ vector of parameters and ϵ is an $N \times 1$ vector of disturbances. Individual observations will be marked by $n = 1, 2, \ldots, N$ and a subscript k refers to the k^{th} parameter. The estimated model is given by

$$y = Xb + e,$$

for residuals $e = y - Xb$ and parameter estimates b. In ordinary least squares, the sum of squared residuals $(e'e)$ is minimized with

$$L_{OLS} = (y - Xb)'(y - Xb).$$

This loss function is only based on in-sample accuracy, which means that no penalty is included for deviating from prior coefficients β_0. Solving the first-order condition for b gives

$$b_{OLS} = (X'X)^{-1}(X'y). \tag{2.1}$$

These estimated are solely dependent on the data. Researchers typically wish to balance such estimates with prior hypotheses β_0. I will here focus on estimators that penalize deviations from β_0 with an ℓ_2 norm; namely, a standard form of Bayesian Regression, Zellner's g-prior, and Ridge Regression.

Bayesian regression is well-known for allowing researchers to make a gradual tradeoff between prior beliefs and data-optimized OLS solutions. A popular prior specification of the linear regression model is the natural conjugate prior distribution of Raiffa and Schlaifer (1961), whereby $p(\beta|\sigma^2) \sim \mathcal{N}(\beta_0, \sigma^2 B_0)$ and $p(\sigma^2) \sim IG(\alpha_0/2, \delta_0/2)$ has an inverted gamma distribution. Under the current specifications, a closed-form solution of the posterior mean is available and is given by the column vector

$$b_{Bayes} = (X'X + B_0^{-1})^{-1}(X'y + B_0^{-1}\beta_0). \tag{2.2}$$

Although the solution of the k^{th} coefficient $b_{Bayes,k}$ need not lie between $\beta_{0,k}$ and $b_{OLS,k}$ (Chamberlain and Leamer, 1976, pp. 74), it is clear that when $B_0 \to \infty$, there is no penalty for deviating from β_0 and we are back at the OLS solution. In case $B_0 \to 0$, deviations from β_0 are so heavily penalized that they are not allowed.

After scaling each regressor, researchers may still have difficulties in anticipating how $B_0 \in [0, \infty]$ corresponds to a degree of trust in his prior coefficients relative to a data-optimized solution. One response has been to develop 'noninformative' priors so that the influence of β_0 is as small as possible again (Jeffreys, 1946, Gelman et al., 2014). Yet, even when one has little information about the underlying relations between X and y, one might still want to perform subset selection or encourage the grouping of regressors.

An intermediate solution in the Bayesian context was offered by Zellner (1986). His g-prior, $\beta \sim \mathcal{N}(\beta_0, g\sigma^2(X'X)^{-1})$, along with a Jeffrey's prior on $\sigma^2 \propto \frac{1}{\sigma^2}$, leads to a posterior mean of

$$b_{Zellner} = \frac{1}{1+g}\beta_0 + \frac{g}{1+g}b_{OLS}, \tag{2.3}$$

which helps to regulate the degree of shrinkage towards β_0 through $g \in [0, \infty)$. To make this even more clear, one could define $g = \frac{1-u}{u}$ to get

$$b_{Zellner} = u\,\beta_0 + (1-u)b_{OLS},$$

so that the estimator becomes a weighted average between β_0 and b_{OLS} with weights of $u \in [0, 1]$. Observe that a parameter's degree of shrinkage is not related to model fit or to cross-correlations between regressors, so Zellner's g-prior does not perform grouping of correlated regressors or subset selection of relevant regressors.

Frequentist shrinkage methods have been developed as well. In Ridge regression (Hoerl and Kennard, 1970), the sum of squared residuals is supplemented with a term that penalizes deviations from zero,

$$L_{Ridge} = (y - Xb)'(y - Xb) + \lambda b'b.$$

Ridge regression has a tendency to make coefficients equal due to its squared norm and this may be convenient when using multicollinear regressors. By solving the first-order condition for b, the estimator becomes

$$b_{Ridge} = (X'X + \lambda I_K)^{-1}(X'y) \tag{2.4}$$

Post-hoc heuristics have been suggested for choosing λ (*ibid.*), but this tuning parameter is usually selected through cross-validation. Marquardt and Snee (1975) emphasize that 'nonessential ill conditioning' can be removed by standardizing the data when performing Ridge regression (pp. 3). They propose to transform X with Z-scores, $\frac{x_k - \text{mean}(x_k)}{\text{std}(x_k)}$, and to center the dependent variable with $y - \text{mean}(y)$. Parameters can subsequently be rescaled by dividing b_k by $\text{std}(x_k)$, and the intercept can be estimated by taking the average of $y - Xb$.

The sensitivity of Ridge regression to the choice of parametrization does imply that the interpretation of the tuning parameter λ is even more opaque than with Bayesian regression, because the researcher can no longer adjust the scale of the data in some favorable manner (Smith and Campbell, 1980, Leamer, 1981). A comparison between b_{Ridge} and b_{Bayes} makes it clear that the prior distribution of β in Ridge regression is assumed to be $\mathscr{N}(0, \sigma^2 I_K/\lambda)$. In a similar vein, it follows that b_{Ridge} equals $b_{Zellner}$ if $\lambda = \frac{1}{g}$ and $(X'X)^{-1} = I_k$. When regressors are orthostandard (orthogonal and standardized), so that $(X'X)^{-1} = \frac{1}{N-1} I_K$, Ridge regression is the same as $b_{Zellner}$ when λ is defined as $\frac{N-1}{g}$. The implication is that for any $\lambda > 0$, b_{Ridge} is directly proportional to b_{OLS} under these conditions. The degree of shrinkage in b_{Ridge} is based on the singular values of X and is unaffected by the strength of the correlation between a regressor and y.

If one wants to use prior coefficients other than zero, then deviances from β_0 could be penalized in the following manner

$$L_{Ridge} = (y - Xb)'(y - Xb) + \lambda(b - \beta_0)'(b - \beta_0).$$

This loss function was developed by Swindel (1976), and results in

$$b_{Ridge} = (X'X + \lambda I_K)^{-1}(X'y + \lambda\beta_0).$$

For this slightly more general Ridge estimator, the prior specification is given by $\beta|\sigma \sim (b_R, \sigma^2 I_K/\lambda)$. Assuming that data are standardized, this means that Ridge solutions correspond exactly to the posterior mean of the Bayesian estimator defined above when we define $B_0 = I_K/\lambda$.

To sum up, in b_{Bayes} there is a tradeoff between model fit and deviations from β_0, but it could be difficult to influence this tradeoff through B_0; and in Zellner's g-prior, the degree of shrinkage is easily controlled, but it is unaffected by model-fit or cross-correlations. Ridge regression does take cross-correlations into account, but practitioners typically experience difficulties in anticipating how a choice of tuning parameter translates into a degree of shrinkage per parameter. Its tuning parameter is often defined as $\lambda = 10^u$ for a hundred equally distributed values of u (Zou and Hastie, 2005), whereby the range of the grid is altered *a posteriori* per application (Friedman et al., 2010, pp. 17). Such problems will be solved when astimators are used, because these methods make the influence of λ on the degree of shrinkage towards β_0 more predictable. I will assume throughout that data are standardized, so that the K regressors in X do not include an intercept.

2.3. b_2 Astimators

The aim of the first AST is to let the astimated parameters deviate from prior parameters insofar as accuracy sufficiently increases. To determine what a sufficient increase is, it is convenient to define a loss function that balances relative accuracy and relative simplicity. In the most general case, this loss function will be minimized over $j = 1, \ldots, J$ candidate configurations c_j. Accordingly, $\text{Fit}(c_j) \geq 0$ is defined to be high when in-sample accuracy is low. By dividing the fit of configuration c_j by the fit of the prior configuration c_0 one obtains a relative accuracy measure.

Turning to relative simplicity, configuration c_j's deviation from the prior configuration is defined as $d(c_j, c_0)$. The highest permissible deviation from c_0 when λ is at its lowest is given by $c_{\max}$. A measure for relative simplicity is thus obtained by dividing $d(c_j, c_0)$ by the maximum permissible deviation from c_0. To make it clear below when I refer to the *maximum* permissible deviation from c_0 when $\lambda = 0$, I use $q(c_{\max}, c_0)$ with a letter q instead of d. A general formulation of an AST loss function is given by

$$L_{AST}(c_j) = \underbrace{\frac{\text{Fit}(c_j)}{\text{Fit}(c_0)}}_{\text{Relative Accuracy}} + f(\lambda) \underbrace{\frac{d(c_j, c_0)}{q(c_{\max}, c_0)}}_{\text{Relative Simplicity}}, \tag{2.5}$$

where λ strikes the balance between the relative increase of a model's in-sample accuracy and the relative decrease of a model's simplicity.

By monotonically transforming equation (2.5), so that the $L_{AST}(c_j)$ values are ordered in the same way, the loss function can be represented in the form of a penalty,

$$L_{AST}(c_j) \propto \text{Fit}(c_j) + \underbrace{f(\lambda)\frac{d(c_j, c_0)}{q(c_{\max}, c_0)}\text{Fit}(c_0)}_{\text{Penalty}}. \tag{2.6}$$

The scalar λ can thus be seen to penalize deviations from a prior configuration c_0 in optimizing over the fit. In case $c_j = c_{\max}$, it follows that $\frac{d(c_j, c_0)}{q(c_{\max}, c_0)} = 1$ and that configuration j must have a fit that is $f(\lambda)$ times better than the fit of c_0 in order to be preferred to c_0.

2.3.1. b_{2i} *Astimator*

The general recipe of an AST loss function in equation (2.5) can now be applied to the linear regression model by using the following ingredients. The measure of fit for the j^{th} set of configurations $c_j = b_j$ is given by the sum of squared residuals, so $\text{Fit}(b_j) = s_j = e_j' e_j$. It follows that the accuracy of Xb relative to $X\beta_0$ can be defined as

$$\text{Relative Accuracy} = \frac{(y - Xb)'(y - Xb)}{(y - X\beta_0)'(y - X\beta_0)}.$$

The relative accuracy term remains unaltered throughout this chapter.

All of the changes are made with respect to relative simplicity. Since we are dealing with an ℓ_2 norm, the deviance from $\beta_{0,k}$ is defined as $d(b_{j,k}, \beta_{0,k}) = (b_{j,k} - \beta_{0,k})^2$. This deviance is made relative to the index q, which will depend on the maximum deviation from $\beta_{0,k}$ when $\lambda = 0$, so $c_{\max} = b_{OLS,k}$. For now, I will define q as $q_{2i} = (b_{OLS,k} - \beta_{0,k})^2$. The 2 refers to the ℓ_2 norm and the i is added to emphasize that this relative simplicity index is defined in terms of an *individual* deviation from $\beta_{0,k}$, which is independent of the deviations between $b_{OLS,j}$ and $\beta_{0,j}$ of other parameters. The simplicity of b relative to b_{OLS} therefore becomes

$$\text{Relative Simplicity} = \sum_{k=1}^{K} \frac{(b_k - \beta_{0,k})^2}{(b_{OLS,k} - \beta_{0,k})^2}.$$

For reasons that will quickly become apparent, relative accuracy and relative simplicity should be balanced through a function of λ_k that is defined as $f(\lambda_k) =$

$\frac{\lambda_k}{1-\lambda_k}$. Putting these terms together results in the following AST loss function

$$L_{2ASTi} = \frac{(y - Xb)'(y - Xb)}{(y - X\beta_0)'(y - X\beta_0)} + \sum_{k=1}^{K} \frac{\lambda_k}{1-\lambda_k} \frac{(b_k - \beta_{0,k})^2}{(b_{OLS,k} - \beta_{0,k})^2}, \tag{2.7}$$

$$= \frac{(y - Xb)'(y - Xb)}{(y - X\beta_0)'(y - X\beta_0)} + (b - \beta_0)'\Lambda Q_{2i}^{-1}(b - \beta_0), \tag{2.8}$$

where Q_{2i} and Λ are diagonal matrices of size K. The diagonal elements of Q_{2i} are given by q_{2i}. The matrix Λ has diagonal elements $\frac{\lambda_k}{1-\lambda_k}$. In case all λ_k are the same, I will just refer to these values as λ. I will also denote the sum of squared residuals of the prior β_0 as $s_0 = (y - X\beta_0)'(y - X\beta_0)$.

By solving the first-order condition for b, one gets

$$b_{2ASTi} = (X'X + \Lambda Q_{2i}^{-1}s_0)^{-1}(X'y + \Lambda Q_{2i}^{-1}s_0\beta_0), \tag{2.9}$$

which will also be referred to as a 'b_{2i} astimator'. The researcher only has to specify λ and β_0, because the rest are known. The higher $\lambda \in [0, 1]$, the higher the relative importance of simplicity over fit. When $\lambda = 0$, the data-optimized OLS solution is chosen; and when $\lambda = 1$, the prior parameter is chosen.[1] The prior parameters β_0 can for instance be selected based on previous experience. If one has no clue on how to choose a prior coefficient or whether x_k is relevant in forecasting y, then a good choice could be to set $\beta_{0,k}$ equal to zero. When all $\beta_{0,k}$ are zero ('$\beta_0 = \vec{0}$'), the astimator becomes

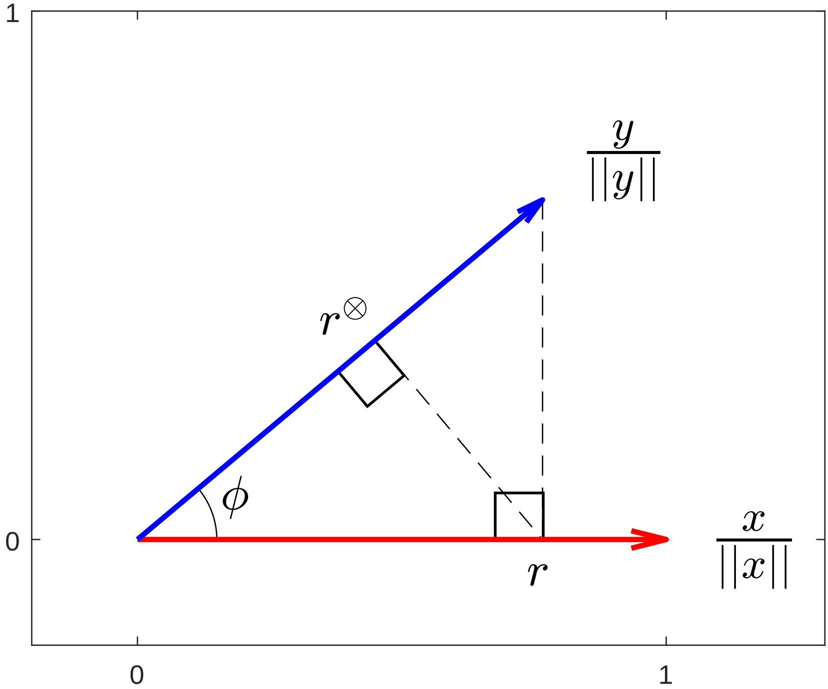
$$b_{2ASTi} = (X'X + \Lambda Q_{2i}^{-1}s_0)^{-1}(X'y), \qquad \beta_0 = \vec{0}. \tag{2.10}$$

To examine the properties of a b_{2i} astimator in closer detail, I will first compare it to the Bayesian and Frequentist estimators above. Subsequently, it will be explained more concretely how λ influences the AST. I will begin with a simple situation whereby $K = 1$ and $\beta_0 = 0$, then study multiple regressors that are uncorrelated while relaxing the assumption that $\beta_0 = 0$, and subsequently analyze what happens in the presence of multicollinearity. For readibility, some aspects will also be relegated to subsections of Appendix 2.A. In subsection **2.A.1**, a derivation of a general ℓ_2 based astimator is provided; and in **2.A.2**, a straightforward Matlab code for b_2 astimators is presented. All of the reformulations of b_{2i} and the other astimators below are derived in **2.A.3**.

[1] I will set $b_{2ASTi,k} = \beta_{0,k}$ when $\lambda_k = 1$. Multiply equation (2.7) by $\sum_{k=1}^{K}(1 - \lambda_k)$, which is a monotonic transformation. For $\lambda = 1$, the relative accuracy measure then contributes 0 to the loss function, so that β_0 is the optimal solution.

The b_{2i} astimator corresponds to a prior specification of $\beta|\sigma \sim \mathcal{N}(\beta_0, \Lambda^{-1}s_0^{-1}Q_{2i}\sigma^2)$. This can be inferred by contrasting b_{2ASTi} to b_{Bayes} in equation (2.2). The b_{2i} astimator is not sensitive to the parameterization of data (see **2.A.4**). Under the exceptional condition that $s_0Q_{2i}^{-1} = I_k$, the astimated solutions are the same as b_{Ridge} with a penalty of $\frac{\lambda}{1-\lambda}$. Zellner's g-prior is obtained when $u = \lambda$ and $X'X = s_RQ_{2i}^{-1}$; and I will now show that the relation between an astimator and the g-prior is particularly interesting.

Figure 2.1: *Geometric Interpretation of* $r^{\otimes}$



Note: the length of column vector x is given by the norm of the inner product $||x|| = \sqrt{x'x}$. A unit vector of length 1 is therefore defined as $\frac{x}{||x||}$. If ϕ is the angle between x and y, then Pearson's correlation coefficient r is the orthogonal projection $\cos\phi$ of unit vector $\frac{y}{||y||}$ onto unit vector $\frac{x}{||x||}$, so that $-1 \leq r \leq 1$. The measure $r^{\otimes} = \cos^2\phi$ is the square of that projection, which implies that $0 \leq r^{\otimes} \leq 1$. The term $r^{\otimes}$ can be represented as a secondary projection onto unit vector y to make the connection with the R^2 measure of in-sample fit apparent. When both x and y are standardized with Z-scores and $K = 1$, the sample correlation r is equal to $b_{OLS} = \frac{x'y}{n-1}$. So, for $0 \leq \phi \leq \frac{\pi}{2}$, the smaller the angle ϕ between x and y, the larger r, the larger $r^{\otimes}$, and the better the fit of xb_{OLS}.

To further study when $X'X = s_RQ_{2i}^{-1}$, let us assume that there is a single regressor ($K = 1$), that the data are standardized, and that $\beta_0 = \vec{0}$. Under these conditions, it can be derived that $b_{Zellner}$ is the same as b_{2ASTi} when xb_{OLS} has a perfect fit in terms of the famous R^2 coefficient of determination. That is,

$$\begin{aligned}(x'x) &\leq s_0 Q_{2i}^{-1},\\ &\leq (y'y)/(b_{OLS}\, b'_{OLS}),\\ &\leq (x'x)/r^{\otimes},\end{aligned}$$

where $r^{\otimes} = \frac{(x'y)(x'y)'}{(x'x)(y'y)} \in [0,1]$ is equal to R^2 for centered data and $\beta_0 = \vec{0}$. I will prove the equivalence between $r^{\otimes}$ and R^2 for the more general case where $K \geq 1$ shortly. The sign $\otimes$ has been added to 'r outer' to stress that an outer product is taken, although for $K = 1$ this is the same as an inner product. A geometric representation of $r^{\otimes}$ is presented in Figure 2.1 and is directly related to the Cauchy-Schwarz inequality.[2]

By substituting $s_0 Q_{2i}^{-1} = (x'x)/r^{\otimes}$ into equation (2.10), the following relation between $r^{\otimes}$ and an ℓ_2 based astimator can be obtained,

$$b_{2AST} = \left(1 + \frac{\lambda}{r^{\otimes}(1-\lambda)}\right)^{-1} b_{OLS}, \qquad K = 1, \beta_0 = \vec{0}, \quad (2.11)$$

where I have dropped the letter 'i' in b_{2ASTi} because there is no difference among ℓ_2 based astimators when $K = 1$. What does this formulation say about the influence of the AST tuning parameter λ? When x and y move in the exact same (or exact opposite) direction, $r^{\otimes} = 1$ and xb_{OLS} will have a perfect fit. The solution of b_{2AST} in equation (2.11) will in that case be equal to $b_{Zellner}$ for all $u = \lambda \in [0,1]$, so that $\lambda \cdot 100\%$ specifies in percentage terms with what degree b_{OLS} is shrunk towards $\beta_0 = 0$. Zellner's estimator always shrinks b_{OLS} by the same amount towards β_0 for a given u, regardless of whether the regressor is relevant to the sampled y or not. Through $r^{\otimes}$, a b_2 astimator sooner approximates 0 for a given λ the more x moves in the orthogonal direction of y.

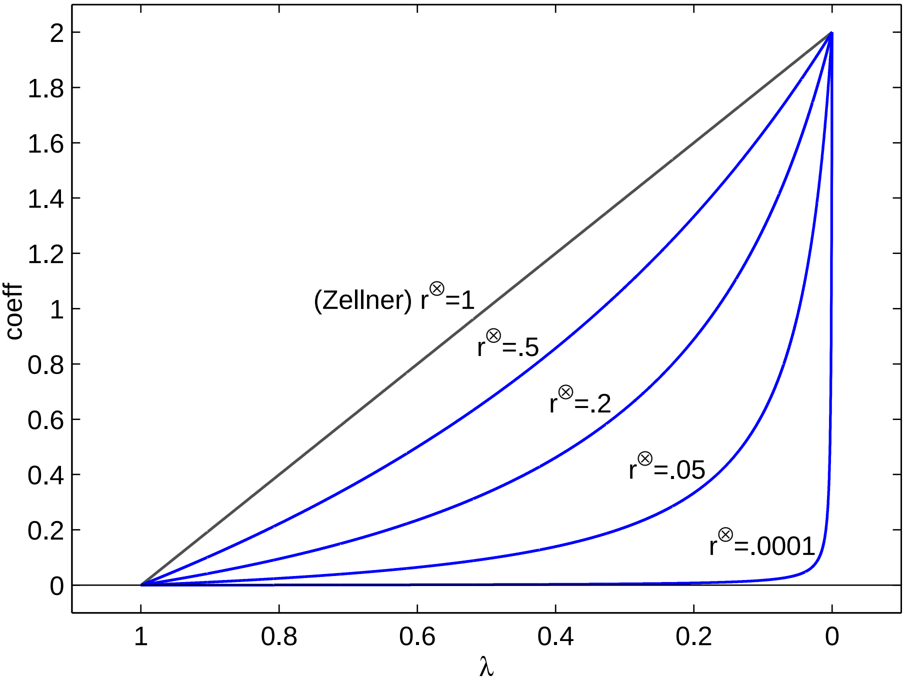
So, the AST tuning parameter specifies the minimum influence of β_0, and this influence increases the worse is the fit of the data-optimized xb. In case $r^{\otimes}$ gets closer to 1, the effect of $r^{\otimes}$ fades away as λ goes to 1 and $(1-\lambda)$ goes to 0. This helps to prevent a shrinkage towards β_0 that is overly stringent for a given λ. For λ values close to 0, b_2 moves in the direction of $(\lambda = r^{\otimes}, b = 0)$.[3]

[2] The Cauchy-Schwarz inequality states that $0 \leq |(x,y)| \leq ||x||\ ||y||$ (Kreyszig, 1999, pp. 361), from which it also follows that $0 \leq (x'y)'(x'y) \leq (x'x)(y'y)$ and $0 \leq r^{\otimes} \leq 1$.

[3] The tangent line (and first-order Taylor approximation) of equation (2.11) at the point $\lambda = 0$ is given by $b_{2AST} = (1 - \lambda/r^{\otimes})b_{OLS}$.

Consequently, subset selection is quickly approximated, since an $r^{\otimes}$ that is nearly equal to zero will ensure that b_{2AST} is close to 0 once $\lambda \approx r^{\otimes}$. Let it here be noted that, in the general case with K orthostandard regressors, centered y, and $\epsilon \sim \mathcal{N}(0, \sigma^2 I_K)$, the expected value of R^2 under the true $\beta = \vec{0}$ is given by $\mathbb{E}(R^2) = \frac{K}{N-1}$.[4] For $K = 1$ and a sample size of $N = 11$, say, the expected value of R^2 is still 10% even when the true $\beta = 0$.

Figure 2.2: *Stylized Solutions of* b_{2AST} *with* $K = 1$, $\beta_0 = 0$, *and* $b_{OLS} = 2$



Under the assumption that $\beta_0 = 0$, this figure shows stylized solutions of $b_{2AST} = (1 + \frac{\lambda}{r^{\otimes}(1-\lambda)})^{-1} b_{OLS}$ with $b_{OLS} = 2$ and varying values of λ and $r^{\otimes}$. The closer $r^{\otimes}$ is to 1, the more similar b_{2AST} is to $b_{Zellner}$.

To illustrate more concretely in which manner the relevance of a regressor influences its degree of shrinkage in b_{2AST}, Figure 2.2 shows stylized solutions of how a single regression coefficient moves from $\beta_0 = 0$ to $b_{OLS} = 2$ as λ decreases from 1 to 0. To generate these results, I varied λ in equation (2.11) for a fixed $r^{\otimes}$ and a prespecified $b_{OLS} = 2$. The upper line shows that λ is the minimum degree of shrinkage of b_{2AST} when $r^{\otimes} = 1$. The regression coefficient is exactly

[4] $R^2 \sim Beta(\frac{K}{2}, \frac{N-K-1}{2})$, see http://davegiles.blogspot.nl/2013/10/more-on-distribution-of-r-squared.html.

halfway between b_{OLS} and β_0 when λ is a half, for example. In the second highest line, the influence of β_0 is further enlarged at a given λ, because an $r^{\otimes}$ of 0.5 is less than perfect. Observe also that when λ is close to zero, each solution path moves towards the point $(\lambda = r^{\otimes}, b = 0)$. A practically irrelevant regressor with $r^{\otimes} = 0.0001$ is approximately zero for most values of λ. I will now show that these stylized solutions are of equal relevance in the multivariate case.

When there are multiple regressors and the prior $\beta_0 = \vec{0}$, a similar expression as equation (2.11) arises if we also assume that regressors are orthostandard. The b_{2ASTi} solutions can then be written as

$$b_{2ASTi,k} = \Big(1 + \frac{\lambda_k}{R^{\otimes}_{kk}(1-\lambda_k)}\Big)^{-1} b_{OLS,k}, \qquad \beta_0 = \vec{0}, X\perp, \tag{2.12}$$

where the $K \times K$ matrix

$$R^{\otimes} = (X'y)(X'y)'(X'X)^{-1}(y'y)^{-1}. \tag{2.13}$$

An $R^{\otimes}_{kk}$ of 1 again implies that $b_{2ASTi,k}$ is equal to $b_{Zellner,k}$, in which case $\lambda_k \cdot 100\%$ becomes a direct measure for the degree of shrinkage towards $\beta_{0,k} = 0$. The smaller $R^{\otimes}_{kk}$, the sooner $b_{2ASTi,k}$ moves to zero for a given λ.

To interpret the diagonal elements of $R^{\otimes}$, the matrix could once more be related to a Cauchy-Schwarz inequality,[5] but it is easiest to remark that for centered data, $\mathrm{tr}(R^{\otimes})$ again equals R-squared, which also implies that $0 \leq \mathrm{tr}(R^{\otimes}) \leq 1$ under these conditions. Note that the trace ('tr') takes the sum of the diagonal elements of a matrix. The identity between $\mathrm{tr}(R^{\otimes})$ and R^2 follows quickly from $R^2 = \frac{b'_{OLS}X'Xb_{OLS}}{y'y} = (X'y)'(X'X)^{-1}(y'y)^{-1}(X'y)$. Just define the $K \times 1$ vectors $(X'y)$ and $(X'X)^{-1}(y'y)^{-1}(X'y)$ and use that an inner product between two vectors is the trace of their outer product. In plain language, R^2 is a scalar that gives an overall measure of fit, while *the diagonal elements of the matrix $R^{\otimes}$ allow us to identify the contribution of each regressor to the fit of the model.*

If Xb_{OLS} has a perfect fit and each orthogonal regressor has an equal contribution to $R^2 = 1$, then $R^{\otimes}_{kk} = \frac{1}{K}$ for each k. When contributions to R^2 vary among regressors, equation (2.12) tells us that these differences will be emphasized quite strongly by a b_{2i} astimator. Assuming orthostandard X, a regressor that is more perpendicular to y will have a smaller $b_{OLS,k}$ solution, which is also shrunk more quickly towards zero because of its small $R^{\otimes}_{\lambda,kk}$.

[5] In the multivariate case, the Cauchy-Schwarz norm can be written as $0 \leq |(X,y)|_F \leq ||(X)||_F \; ||y||_F$, where $||X||_F = \sqrt{\mathrm{tr}(X'X)}$ is the Frobenius norm for matrices (Yang, 2000).

One can also quantify the relevance of individual deviations from prior hypotheses without assuming that $\beta_{0,k} = 0$. For standardized data, R^2 can be defined as a measure that compares the fit of a data-optimized Xb in comparison to the fit of the prior $X\beta_0$ for $\beta_0 = \vec{0}$. In fact, the relative accuracy term in an AST loss function generalizes the optimization over R^2 to situations where β_0 and λ may be different from zero. This more general formulation is

$$R^2_\lambda = 1 - \frac{(y - Xb_\lambda)'(y - Xb_\lambda)}{(y - X\beta_0)'(y - X\beta_0)} = 1 - \text{Relative Accuracy}. \tag{2.14}$$

The larger the data-optimized improvement of the in-sample accuracy of the prior model, the closer R^2_λ is to 1. This quantity is only the same as the original R^2 when $b_R = \vec{0}$ and when the data is standardized (or a constant is included in the model).

Relaxing the assumption that $\beta_0 = \vec{0}$ also implies that

$$R^{\otimes} = (X'\tilde{y}_0)(X'\tilde{y}_0)'(X'X)^{-1}(\tilde{y}_0'\tilde{y}_0)^{-1}, \tag{2.15}$$

where $\tilde{y}_0 = y - X\beta_0$. In **2.A.5** it is proven that tr $R^{\otimes} = R^2_\lambda$ for $\lambda = 0$. Even when β_0 is allowed to be different from zero, that is, the diagonal elements of $R^{\otimes}$ show the contribution of each regressor to the fit of a data-optimized model relative to a prior model.

Armed with these results, one can now let go of the assumption that $\beta_0 = \vec{0}$ in specifying b_{2ASTi} in terms of $R^{\otimes}$. For orthogonal regressors, the result is that

$$b_{2ASTi,k} = \frac{(1-\lambda_k)R^{\otimes}_{kk}}{t_k} b_{OLS,k} + \frac{\lambda_k}{t_k}\beta_{0,k}, \qquad X \perp, \tag{2.16}$$

where the total $t_k = (1-\lambda_k)R^{\otimes}_{kk} + \lambda_k$, see Appendix **2.A.6**. The astimator clearly takes a weighted average between $b_{OLS,k}$ and $\beta_{0,k}$ with weights that sum to 1. If $R^{\otimes}_{kk} = 1$, we obtain Zellner's estimator $(1-\lambda)b_{OLS,k} + \lambda_k\beta_{0,k}$, and a smaller $R^{\otimes}_{kk}$ again causes the influence of $\beta_{0,k}$ to increase.

Finally, when there are multiple correlated regressors and when it is assumed for convenience that $\beta_0 = \vec{0}$, the b_{2i} astimator can be defined as

$$b_{2ASTi} = \left(I_K + \Lambda(X'X)^{-1}Q^{-1}_{2i}s_0\right)^{-1} b_{OLS}, \qquad \beta_0 = \vec{0} \tag{2.17}$$

Using that $Q^{-1}_{2i,kk} = b^{-2}_{OLS,k}$, it can subsequently be derived that the k^{th} diagonal element of $Q_{2i}^{-1} s_0$ is given by

$$Q_{2i,kk}^{-1} s_0 = \left(i_k (X'X)^{-1} (X'y)(X'y)'(X'X)^{-1}(y'y)^{-1} i_k' \right)^{-1},$$
$$= \left(i_k (X'X)^{-1} R^{\otimes} i_k' \right)^{-1},$$

where i_k is a $1 \times K$ vector that is 1 at k and 0 otherwise. The vector i_k is included to select the k^{th} diagonal element of $(X'X)^{-1} R^{\otimes}$.

A smaller $R^{\otimes}_{kk}$ continues to imply that parameter b_k will move more quickly towards $\beta_{0,k} = 0$, and $\text{tr}(R^{\otimes})$ continues to equal R^2. When k and j are correlated, though, the $R^{\otimes}_{kk}$ of regressor k can increase at the cost of a decreasing $R^{\otimes}_{jj}$; and it is not uncommon in my experience to observe that $R^{\otimes}_{jj}$ becomes negative. In more exceptional cases, $R^{\otimes}_{kk}$ can even be larger than one.[6] When $R^{\otimes}$ is used to assess the relevance of regressors, we therefore need to counter the volatility of its diagonal elements by grouping $R^{\otimes}_{kk}$ values of highly correlated regressors together.

The relative simplicity measure of the current b_{2i} astimator already makes its behavior quite predictable under multicollinearity. For $\beta_0 = \vec{0}$, this term is given by $\sum_k \frac{b_k^2}{b^2_{OLS,k}}$. Note that the denominators $b^2_{OLS,k}$ are independent of the deviations $b^2_{OLS,l}$ of other regressors $l \neq k$. Yet, if there is a group of highly correlated regressors, the tendency of the b_{2i} astimator to focus on a single member of that group will be limited. The reason is that the relative simplicities $\left(\frac{b_k}{b_{OLS,k}}\right)^2$ grow with a factor 2, which implies that a given increase in $|b_k|$ is penalized more if $|b_k|$ is already large. Whether regressors are correlated or not, b_{2i} will therefore stimulate parameters to have more similar relative deviations when minimizing the penalty in L_{2ASTi}.

The b_{2i} astimator approximates subset selection in the sense that parameters of irrelevant regressors are equated to approximately β_0 for low λ. Next, I will analyze an astimator that merely stimulates grouping, and subsequently develop the recommended astimator which effectively approximates subset selection and grouping.

2.3.2. b_{2a} *Astimator*

From the forecasting combination literature, we know that giving an equal weight to different regressors often results in hard-to-beat forecasts (Bates and

[6] Think of a simulation study of $y = X\beta + \epsilon$ with only $N = 5$ observations, $K = 4$ equally relevant regressors, $\beta_k = 2$, and standard normal X and ϵ.

Granger, 1969, Smith and Wallis, 2009). In a similar spirit, forecasting accuracy might improve when risks are more diversified across multicollinear regressors. It is unfortunate in this regard that the b_{2i} astimator does not assign more similar weights to highly correlated regressors. On the other hand, a researcher could also have reasons for wanting to ignore a (spurious) regressor that is highly correlated with another. Moreover, when stimulating regressors to receive a similar deviation from β_0 as the others, subset selection may no longer be approximated, because lots of small cross-correlations could have a large effect on the manner in which parameters are estimated. I propose, therefore, that we try to gain more control over how correlated regressors are dealt with.

As a first step, one can define an L_{2ASTa} loss function that uses a matrix Q_{2a} with diagonal elements of $q_{2a} = \frac{1}{K}\sum_l (b_{OLS,l} - \beta_{0,l})^2$. This implies that the *average* OLS deviation from a prior parameter is used to determine a parameter's relative simplicity; whereas, in b_{2i}, an *individual* discrepancy between $b_{OLS,k}$ and $\beta_{0,k}$ was employed. To be clear, the resulting loss function is given by

$$L_{2ASTa} = \frac{(y - Xb)'(y - Xb)}{(y - X\beta_0)'(y - X\beta_0)} + \sum_{k=1}^{K} \frac{\lambda_k}{1 - \lambda_k} \frac{(b_k - \beta_{0,k})^2}{\frac{1}{K}\sum_l (b_{OLS,l} - \beta_{0,l})^2},$$

and the astimator becomes

$$b_{2ASTa} = (X'X + \Lambda Q_{2a}^{-1} s_0)^{-1}(X'y + \Lambda Q_{2a}^{-1} s_0 \beta_0). \qquad (2.18)$$

Nothing has changed with respect to L_{2ASTi} and b_{2ASTi} except that Q_{2a} replaced Q_{2i}. The b_{2ASTa} solutions are a rescaled version of Ridge regression with $\lambda_{Ridge} = Q_{2a}^{-1} s_0 f(\lambda) = \frac{1}{\frac{1}{K}\sum_l (b_{OLS,l} - \beta_{0,k})^2}(y - X\beta_0)'(y - X\beta_0)\frac{\lambda}{(1-\lambda)}$. Standardization of regressors is required, because the scaling of X influences the average deviation between $b_{OLS,k}$ and $\beta_{0,k}$. Provided that λ is intuitively defined, laborious transformations of the data, like the ones advocated in Bayesian regression, are no longer necessary, though.

As aforementioned, all ℓ_2 based astimators result in the same solutions as equation (2.11) when $K = 1$. If regressors are orthogonal, the b_{2a} astimator can be rewritten as

$$b_{2ASTa,k} = \frac{(1 - \lambda_k)\frac{1}{K}R^2}{t_k} b_{OLS,k} + \frac{\lambda_k}{t_k}\beta_{0,k}, \qquad X \perp, \qquad (2.19)$$

for the total $t_k = (1 - \lambda_k)\frac{1}{K}R^2 + \lambda_k$. Under orthogonality, the influence of β_0 is at least λ, and is extended insofar as $\frac{1}{K}R^2$ is small. What this means is that regressors are shrunk based on an overall measure of fit instead of their individual contributions $R^{\otimes}_{kk}$. Consequently, any volatility in $R^{\otimes}_{kk}$ due to cross-correlations has no bearing on b_{2ASTa}. If Xb_{2a} has a perfect fit and regressors are orthogonal, b_{2a} is the same as $b_{Zellner}$ but for the factor $\frac{1}{K}$. When the overall fit of the data-optimized model is poor (low R^2), all parameters are shrunk towards β_0 equally quickly.

For multiple correlated regressors and $\beta_0 = \vec{0}$, we get

$$\begin{aligned} b_{2ASTa} &= \left(I_K + \Lambda (X'X)^{-1} Q_{2a}^{-1} s_0\right)^{-1} b_{OLS}, \\ &= \left(I_K + \Lambda \frac{1}{\frac{1}{K}R^2} \frac{(X'X)^{-1}}{\text{tr } (X'X)^{-1}}\right)^{-1} b_{OLS}. \end{aligned}$$

The influence of $\beta_{0,k}$ is again dictated by R^2. When regressors are correlated, b_{2ASTa} will stimulate their parameters to have a similar 'nominal' deviance from $\beta_{0,k}$. That is to say, instead of the *relative* deviance $\frac{(b_k-\beta_{0,k})^2}{(b_{OLS,k}-\beta_{0,k})^2}$ being the same as another $\frac{(b_l-\beta_{0,l})^2}{(b_{OLS,l}-\beta_{0,l})^2}$, the *nominal* deviance $(b_k - \beta_{0,k})^2$ becomes more similar to $(b_l - \beta_{0,l})^2$.

One can understand why that happens by taking a closer look at the relative simplicity measure again. The denominator $\frac{1}{K}\sum_l (b_{OLS,l} - \beta_{0,l})^2$ can be ignored, because that is the same for all parameters. Turning to the numerator $(b_k - \beta_{0,k})^2$, observe that it grows linearly with a factor 2 for a given increase in b_k. In deciding which regressors should be allowed to deviate more from $\beta_{0,k}$ based on information that is shared among correlated regressors, less relevant parameters are therefore given more leeway to deviate from β_0, because they have a smaller b_k to begin with. As a result, the added effect of small cross-correlations can easily stimulate a barely relevant regressor b_k to have a squared nominal deviation from $\beta_{0,k}$ that greatly exceeds $(b_{OLS,k} - \beta_{0,k})^2$. In the following section, an astimator will be introduced that allows the researcher to specify through a tuning parameter $c_{\min} \in [0, 1]$ how high cross-correlations need to be for parameters to be grouped together.

2.3.3. b_{2c} *astimator*

In comparison to b_{2ASTi}, a disadvantage of b_{2ASTa} (and Ridge regression) is that subset selection is no longer approximated. The b_{2i} astimator does approach subset selection by equating parameters of irrelevant regressors to approximately β_0 for most values of λ. On the other hand, b_{2ASTi} does not encourage the grouping of highly correlated regressors. To play to the strengths of both b_{2ASTi}

and b_{2ASTa}, one should take into account how each regressor is correlated with the other regressors when taking a weighted average of $(b_{OLS,l} - \beta_{0,l})^2$. The third and final astimator that will be presented in this chapter gives control over the influence of cross-correlations, so that grouping and subset selection can both be performed effectively. It is called b_{2ASTc}, where the letter c stands for correlation.

Central to the b_{2c} astimator is $\Theta(X)$, which is a normalized matrix of absolute cross-correlations $|corr(X)|$. The k^{th} column of $\Theta(X)$ is called θ^k. Normalization just means that the rows of each column of absolute correlations are divided by the sum of that column, so that the columns add up to $\sum_{l=1}^{K} \theta_l^k = 1$, where l denotes a row. Using these correlation-based weights, one can define L_{2ASTc} through the diagonal elements $q_{2c} = \sum_l \theta_l^k (b_{OLS,l} - \beta_{0,l})^2$ of Q_{2c}. This results in the following loss function

$$L_{2ASTc} = \frac{(y - Xb)'(y - Xb)}{(y - X\beta_0)'(y - X\beta_0)} + \sum_k \frac{\lambda_k}{1 - \lambda_k} \frac{(b_k - \beta_{0,k})^2}{\sum_l \theta_l^k (b_{OLS,l} - \beta_{0,l})^2}.$$

The first-order condition leads to

$$b_{2ASTc} = (X'X + \Lambda Q_{2c}^{-1} s_0)^{-1} (X'y + \Lambda Q_{2c}^{-1} s_0 \beta_0). \tag{2.20}$$

As aforesaid, derivations of the astimators and straightforward Matlab codes are presented in Appendix 2.A. Let me emphasize once more that regressors should be standardized.

Before rewriting b_{2ASTc} into a more convenient form, it is good to get more acquainted with how Q_{2c} balances between an *individual* Q_{2i} and an *average* Q_{2a}. Assume that $\beta_{0,k} = 0$ for all $K = 4$ parameters and consider the following two examples. First, when $\theta^3 = [0\ 0\ 1\ 0]'$, this means that X_3 is completely uncorrelated with the other regressors, so that $Q_{2c}(3,3) = (b_{OLS,3} - \beta_{0,3})^2 = Q_{2i}(3,3)$. That is, subset selection is approximated just like in the initial b_{2ASTi} of equation (2.9). Second, in case $\theta^3 \approx [.25\ .25\ .25\ .25]'$, the third regressor is almost perfectly correlated with the other regressors ($\theta_l^3 \approx \frac{1}{K}$). Parameters are therefore grouped together through $Q_{2c}(3,3) \approx \frac{1}{K} \sum_l (b_{OLS,l} - \beta_{0,l})^2 = Q_{2a}(3,3)$, which leads to b_{2a} of equation (2.18). So, the correlation vector θ_j^k determines the degree to which parameters have a similar nominal deviation from β_0.

Accordingly, when there are multiple orthostandard regressors, the b_{2c} astimator balances between b_{OLS} and β_0 in

$$b_{2ASTc,k} = \frac{(1 - \lambda_k) \sum_l \theta_l^k R_{ll}^{\otimes}}{t_k} b_{OLS,k} + \frac{\lambda_k}{t_k} \beta_{0,k}, \quad X \perp, \tag{2.21}$$

where the total $t_k = (1-\lambda_k)\sum_l \theta_l^k R_{ll}^{\otimes} + \lambda_k$. For each parameter, the degree of shrinkage towards $\beta_{0,k}$ is influenced by the weights θ_l^k that b_{2ASTc} assigns to the diagonal elements of $R^{\otimes}$ through $\sum_l \theta_l^k R_{ll}^{\otimes}$. Note that, by measuring a regressor's relevance in terms of a weighted average of diagonal $R^{\otimes}$ values, one can counter arbitrary fluctuations in $R^{\otimes}$ caused by cross-correlations. Since regressors are currently assumed to be uncorrelated, θ_l^k is 1 at k and 0 otherwise ($\Theta = I_K$), so that $b_{2ASTc} = b_{2ASTi}$ and subset selection is approximated through the diagonal elements of $R^{\otimes}$. The vector $r_{2c,k}^{\otimes} = \sum_l \theta_l^k R_{ll}^{\otimes}$ could generally be useful in quantifying the in-sample relevance of deviating from each prior hypothesis.

In the presence of multicollinearity, one can assume for ease of display that $\beta_0 = \vec{0}$ to get

$$b_{2ASTc} = \Big(I_K + \Lambda(X'X)^{-1}Q_{2c}^{-1}s_0\Big)^{-1} b_{OLS}, \qquad \beta_0 = \vec{0}, \tag{2.22}$$

where $Q_{2c,kk}^{-1}s_0 = 1/\mathrm{tr}\Big(\mathrm{diag}(\theta^k)(X'X)^{-1}R^{\otimes}\Big)$. Note that $\mathrm{tr}(\mathrm{diag}(\theta^k)R^{\otimes}) = r_{2c,k}^{\otimes}$. Through Θ, two parameters will be stimulated to have a similar nominal deviation from β_0 insofar as their cross-correlation is high. Only in the case that all regressors are (nearly) the same does $\theta_l^k \approx \frac{1}{K}$ and $b_{2c} \approx b_{2a}$.

One of the main advantages of b_{2ASTc} is that the matrix Θ can be adjusted manually. One can, for example, set $\Theta_{i,j}$ and $\Theta_{j,i}$ to zero when prior parameters are different ($\beta_{0,i} \neq \beta_{0,j}$). Another important incentive for altering Θ is that (many) small correlations between regressors could have a large effect. A researcher can specify how large the minimum degree of correlation must be for the deviance of $(b_k - \beta_{0,k})$ to be influenced by some other deviance $(b_j - \beta_{0,j})$. Put differently, one can set $|corr(X)| < c_{\min}$ to zero for a minimum correlation of $c_{\min} = 0.5$, say. It is through $c_{\min}$ that the second AST of grouping parameter together can be controlled, as I will illustrate with a simulation study and an empirical application in the following section.

2.4. Analyzing the Influence of Tuning Parameters

2.4.1. *Simulation Studies*

Having introduced a b_{2i} astimator that focuses on subset selection, a b_{2a} astimator that merely stimulates grouping, and a b_{2c} astimator that does both, I will now further analyze the theoretical claims of the previous sections with a simulation study. The influence of the tuning parameters will be assessed with a simulation exercise whereby there are two relevant and highly correlated regressors and two irrelevant and uncorrelated regressors. That is, twenty data points will be simulated with $y = X\beta + \epsilon$, where $\beta = [2\ 2\ 0\ 0]'$, $\epsilon \sim \mathcal{N}(0, 1)$, $X \sim \mathcal{N}(0, \Sigma)$, and $\Sigma = I$ except for $\Sigma_{\{2,1\},\{1,2\}} = .9$. The priors will be defined as $\beta_0 = \vec{0}$. The $N = 20$ realizations of this simulation study are presented in Appendix **2.A.7**.

Figure 2.3 gives solutions paths for a single simulated data set, whereby the independent variables are standardized with Z-scores and the dependent variable is centered. The main goal of a 'solution path' is to reveal the manner in which b_k move from the prior $\beta_{0,k}$ to the data-optimized $b_{OLS,k}$ as the penalty parameter changes. A reason for preferring one solution path over another could be that the relation between λ and the degree of shrinkage is straightforward, that irrelevant regressors barely deviate from their priors for many values of λ, or that highly correlated regressors are assigned a similar parameter value. The panel in the upper left corner of Figure 2.3, for example, again shows that $b_{Zellner}$ linearly shrinks coefficients from b_{OLS} to β_0 as the tuning parameter $u \in g = \frac{1-u}{u}$ goes from 0 to 1. Observe that the degree of shrinkage is not influenced by a regressor's relevance or by its cross-correlations.

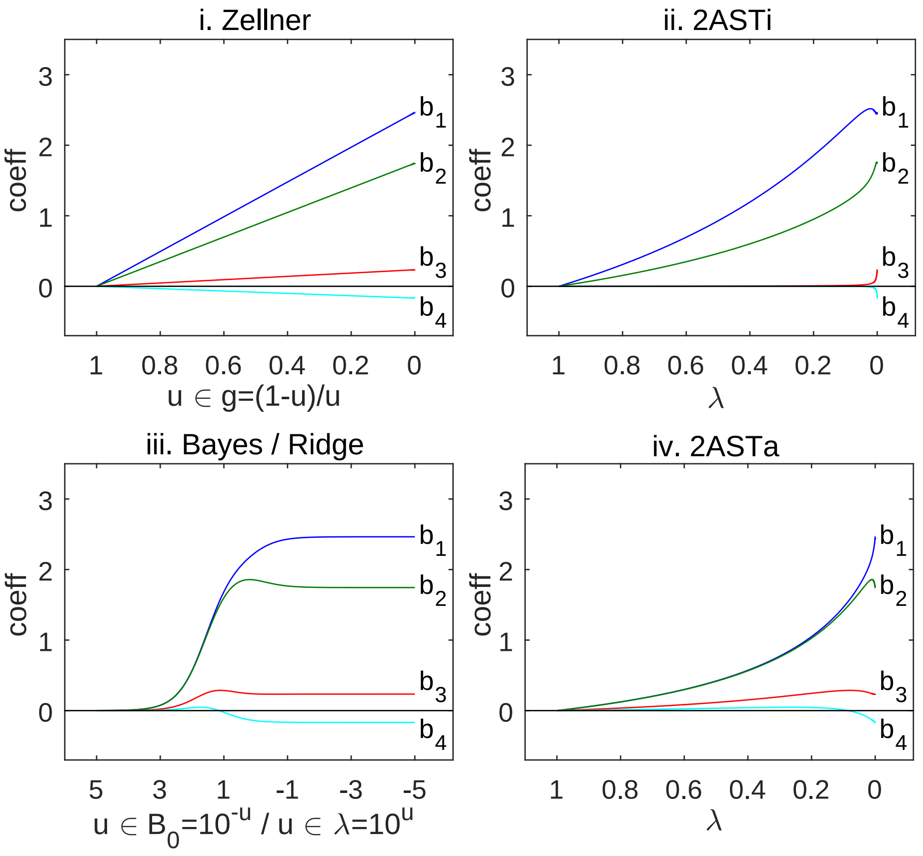
The panel in the upper right corner presents the solutions paths of b_{2ASTi}. I have explained above that the subset selection of b_{2ASTi} is determined by the diagonal elements of $R^{\otimes}$. In the current data set, this matrix is given by

$$R^{\otimes} = \begin{pmatrix} .56 & .40 & .05 & -.04 \\ .56 & .40 & .05 & -.04 \\ .17 & .12 & .02 & -.01 \\ .06 & .04 & .01 & -.00 \end{pmatrix}.$$

The sum of the diagonal elements equals $R^2 = \mathrm{tr} R^{\otimes} = 0.97$. The irrelevant regressors x_3 and x_4 indeed have tiny values of $R^{\otimes}_{3,3} = 0.02$ and $R^{\otimes}_{4,4} = -0.004$, while the contributions of the relevant regressors to R^2 are quite large with 0.56 for x_1 and 0.40 for x_2. As predicted, the tendency of b_{2ASTi} to select a single regressor out of a group of correlated regressors is limited by the ℓ_2 norm. The proportional difference between b_1 and b_2 remains roughly similar. Due to the small $R^{\otimes}$ values, the irrelevant parameters b_3 and b_4 are almost exactly equated to zero for most values of λ.

The lower left panel of Figure 2.3 shows solutions paths for Bayesian and Ridge regression, which are the same for $B_0 = I_K/\lambda$ under the prior specifications presented in Section 2.2. Around $u = 1$, the first two parameters are grouped

Figure 2.3: *Solutions Paths: ℓ_2 Estimators and Their Astimated Analogues*



This figure shows solution paths for estimators and astimators, with the coefficients on the vertical axis and the tuning parameter on the horizontal axis. The tuning parameters are $B_{0,k}$ for b_{Bayes} and g for Zellner's g-prior and are defined in terms of u. Astimators use λ as their tuning parameter. Data ($N = 20$) are simulated with $y = X\beta + \epsilon$, $\beta = [2\ 2\ 0\ 0]'$, $\epsilon \sim \mathcal{N}(0,1)$, $X \sim \mathcal{N}(0,\Sigma)$, and $\Sigma = I$ except for $\Sigma_{\{2,1\},\{1,2\}} = 0.9$. Prediction model: $\hat{y} = Xb$, whereby $\beta_0 = [0\ 0\ 0\ 0]'$.

together while b_3 is also slightly stimulated to deviate from 0. The main difficulty with the Bayesian estimator (and Ridge regression) is to anticipate how a choice of the prior variance (B_0) influences the tradeoff between accuracy and simplicity for each parameter. In the current example, I have defined $B_0 = 10^{-u} I_K$. When $u = 5$, the parameters are all shrunk towards $\beta_{0,k} = 0$. Apparently, the value of $u = 5$ is associated with a high degree of confidence in β_0 here. At around $u = 3$, the parameters suddenly start to alter. We can only infer *after* producing the estimates, that a value of $u = -1$ corresponds to a small degree of confidence in β_0, since the $b_{OLS,k}$ solutions are dominant from this point onwards.

The lower right panel presents b_{2a}, which is the astimated analogue of Ridge regression. Remember that, in computing the relative simplicity measure, this astimator takes a simple average over all the squared deviations from the priors, so $q_{1a} = \sum_l \frac{1}{K}(b_{OLS,l} - \beta_{0,l})^2$. Since R^2 is close to 1, the degree of shrinkage towards zero of the first two parameters roughly corresponds to $r^{\otimes} = 1/K = 0.25$ in the stylized solutions of Figure 2.2. Although λ is nicely defined to be between 0 and 1, subset selection of irrelevant regressors is no longer approximated with this Ridge-type astimator. It also takes a while for the first two parameters to be grouped together.

The recommended b_{2ASTc} with $q_{c,k} = \sum_l \theta_l^k (b_{OLS,l} - \beta_{0,l})^2$ balances between b_{2ASTi} and b_{2ASTa} based on absolute cross-correlations. For the current data set, the absolute correlation matrix is given by

$$|\text{corr}(X)| = \begin{pmatrix} 1 & .94 & .20 & .14 \\ .94 & 1 & .28 & .14 \\ .20 & .28 & 1 & .05 \\ .14 & .14 & .05 & 1 \end{pmatrix}.$$

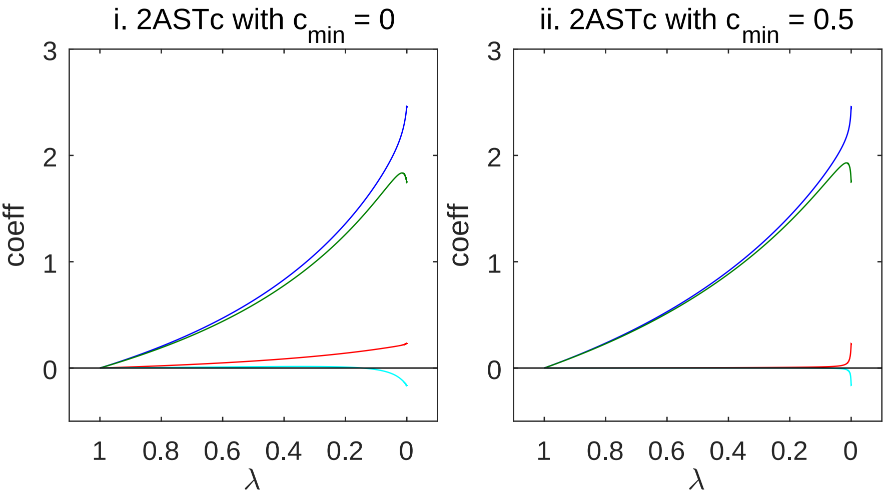
Note that x_1 and x_2 have a high cross-correlation of 0.94. Standardizing this matrix results in

$$\Theta = \begin{pmatrix} .44 & .40 & .13 & .11 \\ .41 & .42 & .18 & .11 \\ .09 & .12 & .65 & .04 \\ .06 & .06 & .03 & .75 \end{pmatrix}.$$

Θ is the same as $|\text{corr}(X)|$, except that the k^{th} column θ^k now sums to 1 (rounding errors aside). Through $c_{\min}$, the researcher can specify how high the minimum amount of correlation must be for parameters to be grouped together.

In the left panel of Figure 2.4, one can see that if the smallest of cross-correlations is allowed to influence the relative simplicity index ($c_{\min} = 0$), the

Figure 2.4: *Solution Paths: 2ASTc*



This figure shows solution paths for the b_{2c} astimator, with the estimated coefficients on the vertical axis and the tuning parameter on the horizontal axis. Data ($N = 20$) are simulated with $y = X\beta + \epsilon$, $\beta = [2\ 2\ 0\ 0]'$, $\epsilon \sim \mathcal{N}(0, 1)$, $X \sim \mathcal{N}(0, \Sigma)$, and $\Sigma = I$ except for $\Sigma_{\{2,1\},\{1,2\}} = 0.9$. Prediction model: $\hat{y} = Xb$, with $\beta_0 = \vec{0}$.

third and fourth parameters are still urged to some degree to have a similar deviance from $\beta_{0,k}$ as the others. Alternatively, one can also specify that all absolute correlations $|\text{corr}(X)|$ below $c_{\min} = 0.5$ are equated to zero. The matrix Θ then becomes

$$\Theta = \begin{pmatrix} .52 & .48 & 0 & 0 \\ .48 & .52 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

This specification of Θ ensures that only the first two parameters are grouped together. The resulting solutions are presented in the right panel of Figure 2.4. Note that the irrelevant regressors are inactivated just as quickly as in b_{2ASTi}, and that the grouping of b_1 and b_2 is performed more effectively than in b_{2ASTa}.

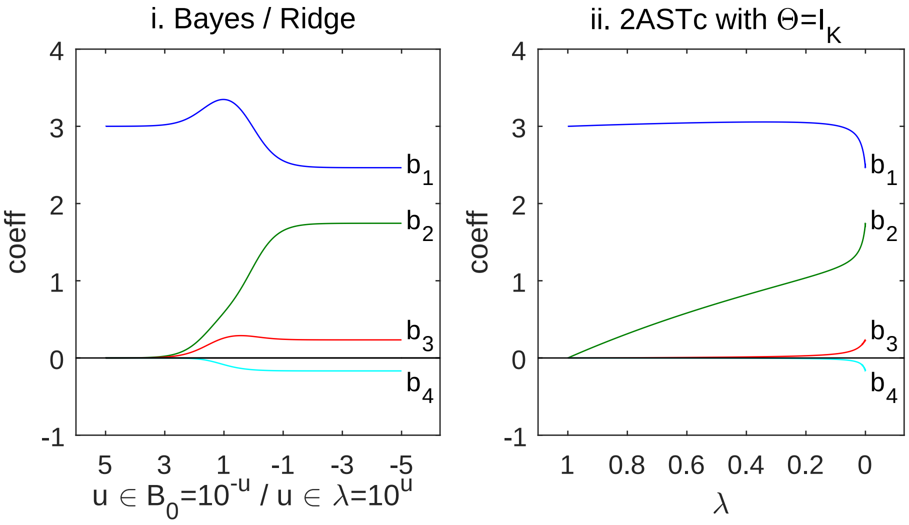
Another implication of setting $c_{\min} = 0.5$ is that $\text{diag}(R^{\otimes}) = [0.56\ \ 0.40\ \ 0.02\ \ -0.00]'$ is changed into the correlation-adjusted $r^{\otimes}_{2c} = [0.48\ \ 0.48\ \ 0.02\ \ -0.00]'$, which gives a better sense of the relevance of each regressor.[7] Observe also that the degree of shrinkage of b_1 and b_2 with an $r^{\otimes}_{2c}$ of

[7] To be clear, the first element of $r^{\otimes}_{2c}$ is computed as $\sum_l \theta^1_l R^{\otimes}_{ll} \approx 0.52 \cdot 0.56 + 0.48 \cdot 0.40 + 0 \cdot 0.02 - 0 \cdot 0.00 \approx 0.48$.

0.48 is similar to the degree of shrinkage of the stylized solutions in Figure 2.2 with $r^{\otimes} = 0.5$.

Continuing with the same simulated date set, I will finally illustrate how the grouping of regressors can be controlled when prior coefficients are $\beta_0 = [3\ 0\ 0\ 0]'$ instead of $\vec{0}$. Stimulating parameters to have a similar deviance from their priors usually makes little sense when the priors are different. Yet, the left panel of Figure 2.5 shows that the estimators of Bayes (and Ridge) move b_1 considerably above 3 to make the deviance of $(b_1 - \beta_{0,1})^2$ more similar to the deviance of the other parameters. With b_{2ASTc}, one can set $\Theta = I_K$, so that the first parameter does not affect the relative simplicity index of the other parameters. The result is that b_1 no longer takes a detour in converging towards $\beta_{0,1} = 3$, as the right panel shows. In this way, b_{2ASTc} allows the researcher to control the grouping of parameters.

Figure 2.5: *Solution Paths:* $\beta_0 = [3\ 0\ 0\ 0]'$



This figure shows solution paths for Bayes/Ridge and 2ASTc, with the estimated coefficients on the vertical axis and the tuning parameter on the horizontal axis. Data $(N = 20)$ are simulated with $y = X\beta + \epsilon$, $\beta = [2\ 2\ 0\ 0]'$, $\epsilon \sim \mathcal{N}(0, 1)$, $X \sim \mathcal{N}(0, \Sigma)$, and $\Sigma = I$ except for $\Sigma_{\{2,1\},\{1,2\}} = 0.9$. Prediction model: $\hat{y} = Xb$, and $\beta_0 = [3\ 0\ 0\ 0]'$.

The main takeaway from this analysis is that b_{2ASTc} makes it possible for a researcher to anticipate and influence *a priori* how data-optimized parameters of correlated regressors are balanced with β_0. I will finally turn to a famous case study.

2.4.2. *Case Study: Prostate Data*

The ℓ_2 based estimators and astimators will now be used to predict the level of prostate specific antigen (PSA) based on a constant and eight clinical measures. The data are obtained from Hastie et al. (2009).

Table 2.1: *Description Regressors Prostate*

	Name	Description	b_{OLS}	$R^{\otimes}_{kk}$	$r^{\otimes}_{2c}$
y	PSA	prostate specific antigen			
x_1	lcavol	log of cancer volume	.67	.42	.21
x_2	svi	seminal vesicle invasion $(0, 1)$	.32	.15	.15
x_3	lweight	log prostate weight	.27	.10	.10
x_4	lbph	log of benign prostatic hyperplasia	.14	.02	.02
x_5	pgg45	percent of Gleason scores 4 or 5	.13	.05	.00
x_6	gleason	Gleason score (categorical)	.04	.01	.00
x_7	lcp	log of capsular penetration	-.15	-.07	.10
x_8	age	a person's age	-.16	-.02	-.02

Regressors are numbered from top to bottom based on the value of OLS applied to standardized data (fourth column). In the last column, $R^{\otimes}$ is adjusted for cross-correlations in the spirit of b_{2c} through $r^{\otimes}_{2c} = \Theta' \mathrm{diag}(R^{\otimes})$ and $c_{\min} = 0.5$.

Table 2.2: *Prostate Correlation Matrix*

	svi	lweight	lbph	pgg45	gleason	lcp	age
1. lcavol	.54	.28	.03	.43	.43	.68	.22
2. svi	1	.16	-.09	.46	.32	.67	.12
3. lweight		1	.44	.11	.06	.16	.35
4. lbph			1	.08	.08	-.01	.35
5. pgg45				1	.75	.63	.28
6. gleason					1	.51	.27
7. lcp						1	.13
8. age							1

Regressors are numbered from top to bottom based on the value of OLS applied to scaled data (see Table 2.1).

Table 2.1 enlists the clinical measures and Table 2.2 gives the cross-correlations between the regressors. Some of these variables may immediately make sense, like the log of cancer volume, the log of prostate weight, or a person's age. Other regressors might be less clear. The variable svi indicates the presence of prostate cancer in the connective tissue around the seminal vesicles and outside the prostate (Potter et al., 2000). Gleason scores ('gleason') are obtained by a microscopic analysis of samples from a prostate biopsy. The variable pgg45 gives the proportion of high-grade carcinoma (Gleason 4 or 5). The benign prostatic hyperplasia (lbph) is a noncancerous enlargement of a prostate. Capsular penetration (lcp), finally, means that prostate cancer cells have invaded through the prostate capsule (Pan, 2012). The question is which regressors are relevant predictors of PSA. I will investigate this matter by setting $\beta_{0,k} = 0$ for all $k = 1, 2, \ldots, K$.

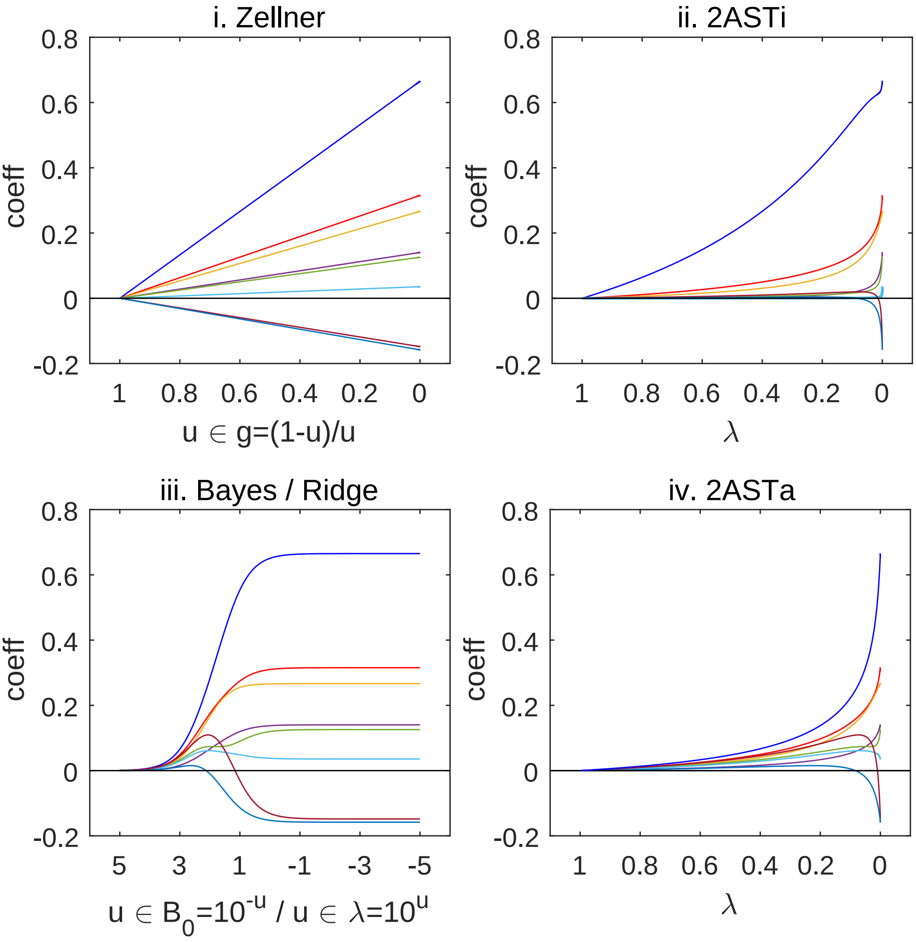
To analyze the relative influence of regressors, solution paths will be presented based on standardized data. I have ordered the regressors in Table 2.1 from the highest OLS estimate to the lowest. The second to last column presents the diagonal elements of $R^{\otimes}$ and it sums to $R^2 = 0.66$.

Figure 2.6 shows that Zellner's g-prior gives the researcher great control over the degree of shrinkage, but without regard for a regressor's contribution to accuracy. Note that $\lambda = 0$ corresponds to the b_{OLS} solutions in Table 2.1, so the highest line is the log of cancer volume, lcavol. With the b_{2i} astimator in the top right corner, the degree of shrinkage of lcavol is quite close to being linear because of its $R^{\otimes}$ value of 0.42. Under the current prior specifications, the Bayesian estimator is equal to Ridge regression for these standardized data. The relationship between a degree of shrinkage and λ is again quite unpredictable. In this case, most of the action happens between $B_0 = 0.001$ to $B_0 = 1$. The analogous b_{2a} astimator in the lower right corner of Figure 2.6 quickly promotes parameters to have similar deviations from $\beta_0 = \vec{0}$.

The b_{2c} astimator only groups parameters together when their associated cross-correlation exceeds $c_{\min} = 0.5$. The correlation-adjusted contributions to R^2 are reported in the last column of Table 2.1 under $r_{2c}^{\otimes} = \Theta' \text{diag}(R^{\otimes})$. The log of cancer volume (lcavol) is again identified as the foremost predictor of the level of prostate specific antigen and four out of eight predictors are revealed to barely contribute to R^2 accuracy. The estimated relevance of lcp increases after it is compensated for its high cross-correlation with lcavol and svi. Figure 2.7 shows that lcp gets a more similar (squared) deviation from β_0 as these regressors when λ increases. Meanwhile, the parameters of the four irrelevant predictors are quickly shrunk towards zero.

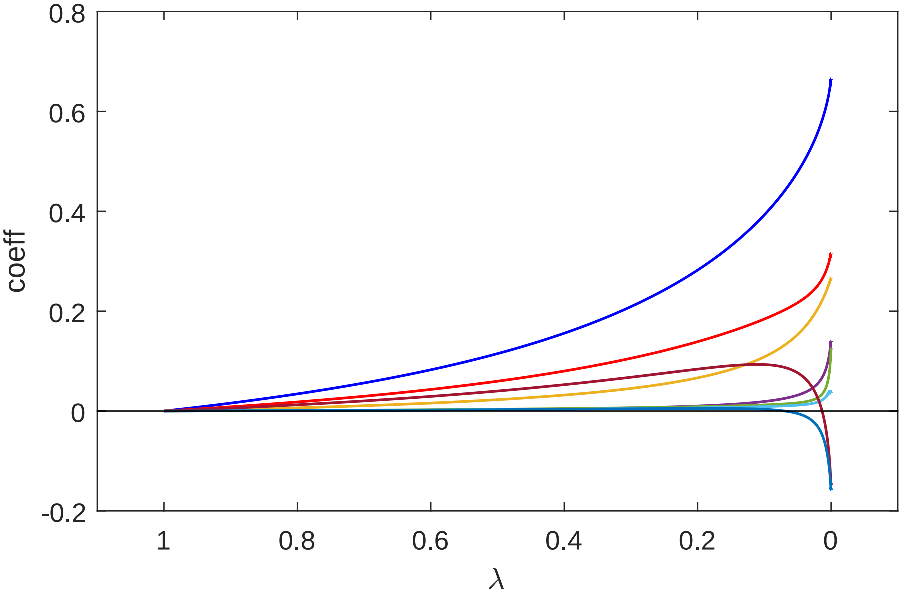
Lastly, it should be pointed out that the choice of $c_{\min}$ is important in determining the relevance that is assigned to a variable. A $c_{\min}$ of 0.4 instead of 0.5 would cause pgg45 to be compensated for its cross-correlations with the first two regressors, for example. The development of a more data-dependent choice of such a tuning parameter is left for future research.

Figure 2.6: *Solutions Paths Prostate Data*



This figure shows solutions paths about the standardized prostate data. From top to bottom at $\lambda = 0$, the regressors are lcavol (1), svi (2), lweight (3), lbph (4), pgg45 (5), gleason (6), lcp (7), and age (8).

Figure 2.7: *Solutions Paths Prostate Data:* b_{2ASTc} *with* $c_{\min} = 0.5$



This figure shows solutions paths of b_{2ASTc} with $c_{\min} = 0.5$ applied to the standardized prostate data. From top to bottom at $\lambda = 0$, the regressors are lcavol (1), svi (2), lweight (3), lbph (4), pgg45 (5), gleason (6), lcp (7), and age (8).

2.5. Discussion

In this chapter, I have attempted to improve upon Bayesian and Frequentist methods by giving the researcher more control over the first AST which penalizes deviations from β_0 and the second AST which promotes grouping at the cost of in-sample accuracy. The b_{2c} astimator allows the researcher to influence the first AST through λ and the second AST through $c_{\min}$.

Relative simplicity was defined with an ℓ_2 norm and relative accuracy was related to an $R^{\otimes}$ matrix, whose diagonal elements indicate the contribution of each regressor to the R^2 measure of fit. I might mention that the off-diagonal elements of $R^{\otimes}$ can be interpreted as well. Define $R_{xx} = \text{corr}(X)$ and a vector of correlations $\vec{r}_{xy} = [r_{x_1y}\ r_{x_2y}\ \dots\ r_{x_Ky}]'$. If $\beta_0 = \vec{0}$, y is centered and X is standardized, it directly follows that $R^{\otimes} = \vec{r}_{xy}\vec{r}_{xy}^{\,\prime}R_{xx}^{-1}$.[8] If we assume that X is orthogonal ($R_{xx} = I_K$), then each element is just given by $R^{\otimes}_{k,l} = r_{x_ky}r_{x_ly}$, the product of the regressors' correlations with y.

In future research, astimators might be developed with an ℓ_1 norm so that exact subset selection can be performed. To facilitate the choice of λ, techniques should be explored that make this specification more dependent on the data. The out-of-sample performances of estimators and astimators ought to be compared as well.

2.A. Appendix: Further Details Regarding b_{2AST}

2.A.1: *Show that b_{2ASTc}, which has b_{2ASTi} and b_{2ASTa} as special cases, corresponds to the global minimum of L_{2ASTc}.*

1. **Solve first-order condition**
   Start with the loss function

$$L_{2ASTc} = \frac{(y - Xb)'(y - Xb)}{(y - X\beta_0)'(y - X\beta_0)} + (b - \beta_0)'\Lambda Q_{2c}^{-1}(b - \beta_0).$$

[8] For centered y and standardized data X, $\vec{r}_{xy} = \frac{(X'y)}{\sqrt{(N-1)(y'y)}}$ and $R_{xx} = \frac{1}{N-1}(X'X)$, so that $R^{\otimes} = \frac{(X'y)}{\sqrt{(N-1)(y'y)}}\frac{(X'y)'}{\sqrt{(N-1)(y'y)}}(N-1)(X'X)^{-1} = (X'y)(X'y)'(X'X)^{-1}(y'y)^{-1}$.

Take partial derivatives with respect to each b_k,

$$\frac{\partial}{\partial b} L_{2ASTc} = -2X'y + 2X'Xb + 2\Lambda Q_{2c}^{-1} s_0 (b - \beta_0),$$

where $s_0 = (y - X\beta_0)'(y - X\beta_0)$. The first-order condition $\frac{\partial}{\partial b} L_{2ASTc} = 0$ results in

$$(X'X + \Lambda Q_{2c}^{-1} s_0) b = X'y + \Lambda Q_{2c}^{-1} s_0 \beta_0,$$

so the solutions are given by

$$b_{2ASTc} = (X'X + \Lambda Q_{2c}^{-1} s_0)^{-1} (X'y + \Lambda Q_{2c}^{-1} s_0 \beta_0).$$

2. **Check that the solution is a minimum**
The solution of a first-order condition is a minimum if the 'Hessian' matrix with second-order partial derivatives is positive definite ('PD'). The Hessian is given by

$$\frac{\partial^2}{\partial b \partial b'} L_{2ASTc} = 2X'X + 2\Lambda Q_{2c}^{-1} s_0. \tag{2.23}$$

i. The sum of a PD and a PSD (positive semi-definite) matrix is PD.

   Consider $K \times K$ matrices A and B whereby A is PD and B is PSD. In that case $x'Ax > 0$, $x'Bx \geq 0$, and $x'(A + B)x = x'Ax + x'Bx > 0$.

ii. $X'X$ is PD if X has rank K.

   For a column vector z of size K that does not consist entirely of zeros, it follows that $X'X$ is PSD, since

$$z'X'Xz = (Xz)'(Xz) = \sum_{n=1}^{N} c_n^2 \geq 0,$$

   where the $N \times 1$ vector $c_n = Xz$. This solution is positive definite if there exists no exact linear relationship between the columns of X, in which case X is full rank.[9]

iii. The second term on the right hand side of equation (2.23) is at least PSD because it is a diagonal matrix with nonnegative diagonal elements.

[9] As explained in Heij et al. (2004, pp. 733), the rank of an $N \times K$ matrix A is equal to the largest number r for which there exists a square submatrix of A of size r that has a non-zero determinant. If A has a rank of $r < K$, then there exists a non-zero $K \times 1$ vector z such that $Az = 0$. Alternatively, one might say that the rank of A corresponds to the number of linearly independent columns of A.

For a nonzero $K \times 1$ vector z and a $K \times K$ diagonal matrix D with diagonal elements d_k, $z'Dz = d_1 z_1^2 + d_2 z_2^2 + \cdots + d_K z_K^2 \geq 0$. If all $d_k > 0$, then D is PD.

iv. Consequently, if X is full rank and/or all $\lambda_k > 0$, the Hessian is PD and b_{2ASTc} is a minimum of L_{2ASTc}.

3. **Ensure that the solution is a global minimum**
   Finally, b_{2ASTc} is a global minimum since L_{2ASTc} is convex. A nonnegative weighted sum of convex functions is itself convex (Boyd and Vandenberghe, 2004, pp. 79), and L_{2ASTc} is the sum of two parabolas which are both convex in b.

2.A.2: *Matlab program for b_{2ASTi}, b_{2ASTa} and b_{2ASTc}:*

```
%Initial definitions:
K = size(X,2); bR = zeros(K,1); lam = 0.5;
LAM = lam/(1-lam)*eye(K);
XX=X'*X; XY=X'*Y; bOLS = XX\XY;
sR = (Y-X*bR)'*(Y-X*bR);

%2ASTi
Qi = diag((bOLS-bR).^-2);
b2ASTi = (XX + LAM*Qi*sR) \ (XY+LAM*Qi*sR*bR);

%2ASTa (assuming standardized data)
Qa = mean((bOLS-bR).^-2)*eye(K);
b2ASTa = (XX + LAM*Qa*sR) \ (XY+LAM*Qa*sR*bR);

%2ASTc (assuming standardized data)
cmin = 0.5;
Theta = abs(corr(X));                %Define absolute correlations.
Theta(Theta<cmin) = 0;               %Equate to zero: abs. corr.<cmin
Theta = Theta*diag(1./sum(Theta));   %Normalize Theta.
Qc = diag((Theta'*(bOLS-bR).^2).^-1);
b2ASTc = (XX + LAM*Qc*sR) \ (XY+LAM*Qc*sR*bR);
```

2.A.3: *Show that a b_{2c} astimator can be written as*

$$b_{2ASTc} = \Big(I_K + \Lambda(X'X)^{-1}Q_{2c}^{-1}s_0\Big)^{-1} b_{OLS}, \qquad \beta_0 = \vec{0}, \tag{2.24}$$

where $Q_{2c,kk}^{-1}s_0 = 1/tr\Big(diag(\theta^k)X'X^{-1}R^{\otimes}\Big)$; and show that b_{2i} and b_{2a} are special cases.

First, for $\beta_0 = \vec{0}$,

$$\begin{aligned} b_{2AST} &= (X'X + \Lambda Q_{2c}^{-1} s_0)^{-1}(X'y), \\ &= (I + (X'X)^{-1}\Lambda Q_{2c}^{-1} s_0)^{-1}(X'X)^{-1}(X'y), \end{aligned}$$

and b_{2ASTc} in equation (2.24) follows from substituting b_{OLS}. Second,

$$\begin{aligned} Q_{2c,kk}^{-1} s_0 &= 1/\Big(\sum_l \theta_l^k (b_{OLS,l})^2\Big)(y'y), \\ &= 1/\mathrm{tr}\Big(\mathrm{diag}(\theta^k)(X'X)^{-1}(X'y)(X'y)'(X'X)^{-1}(y'y)^{-1}\Big), \\ &= 1/\mathrm{tr}\Big(\mathrm{diag}(\theta^k)(X'X)^{-1}R^{\otimes}\Big). \end{aligned}$$

Special cases:

- For orthogonal data (or $K = 1$), $(X'X)^{-1}$ terms cancel.
- For b_{2ASTi}, define $\Theta = I_k$.
- For b_{2ASTa}, define $\theta_l^k = 1/K$ and use that $\mathrm{tr}R^{\otimes} = R^2$.

2.A.4: *Show that b_{2ASTi} does not depend on the choice of parametrization in the sense that Cb is an astimate of Cb when b is an astimate of β for a $K \times K$ diagonal transformation matrix C.*

X is changed into XC^{-1}, b into Cb, and Q_i, with its diagonal elements $\frac{1}{K}(b_{OLS,k} - \beta_{0,k})^2$, is changed into $(C')^{-1}QC^{-1}$. The L_{2ASTi} loss function becomes,

$$\begin{aligned} L_{2ASTi} = \frac{(y - XC^{-1}Cb)'(y - XC^{-1}Cb)}{(y - XC^{-1}C\beta_0)'(y - XC^{-1}C\beta_0)} + \dots \\ \dots \frac{\lambda}{1-\lambda}\frac{1}{K}(Cb - C\beta_0)'(C')^{-1}Q_iC^{-1}(Cb - C\beta_0), \end{aligned}$$

and since $(Cb - C\beta_0)' = (b - \beta_0)'C'$, the scaling matrix C cancels so that the loss function remains the same. Other transformations, such as centering X and y, may influence results.

2.A.5: *It was defined in equation* (2.14) *that*

$$R^2_{\lambda=0} = 1 - \frac{(y - Xb_{OLS})'(y - Xb_{OLS})}{(y - X\beta_0)'(y - X\beta_0)}.$$

Show that $R^2_{\lambda=0} = tr(R^{\otimes})$, *where*

$$R^{\otimes} = (X'\tilde{y}_R)(X'\tilde{y}_R)'(X'X)^{-1}(\tilde{y}_R'\tilde{y}_R)^{-1},$$

for $\tilde{y}_0 = y - X\beta_0$*; see equation* (2.15)*.*

Substituting $b_{OLS} = (X'X)^{-1}X'y$ in the second line below, we can write

$$\begin{aligned} R^2_{\lambda=0} &= 1 - \frac{(y - Xb_{OLS})'(y - Xb_{OLS})}{(y - X\beta_0)'(y - X\beta_0)} \\ &= 1 - \Big(y'y - 2b_{OLS}'X'y + b_{OLS}'X'Xb_{OLS}\Big)(\tilde{y}_0'\tilde{y}_0)^{-1}, \\ &= 1 - y'\Big(I_N - X(X'X)^{-1}X'\Big)y(\tilde{y}_0'\tilde{y}_0)^{-1}. \end{aligned}$$

I will now introduce a matrix W such that $Wy = y - X\beta_0$. This means that $W = I_N - \text{diag}(\gamma X\beta_0)$, where the column vector γ has $\frac{1}{y_n}$ on the n^{th} row. Adding and subtracting the same quantity gives

$$\begin{aligned} R^2_{\lambda=0} &= 1 - y'\Big(WI_NW - WX(X'X)^{-1}X'W\Big)y(\tilde{y}_0'\tilde{y}_0)^{-1}, \\ &= 1 - \Big(\tilde{y}_0'\tilde{y}_0 - (X'\tilde{y}_0)'(X'X)^{-1}(X'\tilde{y}_0)\Big)(\tilde{y}_0'\tilde{y}_0)^{-1}, \\ &= (X'\tilde{y}_0)'(X'X)^{-1}(\tilde{y}_0'\tilde{y}_0)^{-1}(X'\tilde{y}_0), \\ &= \text{tr}(R^{\otimes}). \end{aligned}$$

For the last step, define the $K \times 1$ vectors $(X'\tilde{y}_0)$ and $(X'X)^{-1}(\tilde{y}_0'\tilde{y}_0)^{-1}(X'\tilde{y}_0)$ and use that an inner product between two vectors equals the trace of their outer product.

2.A.6: *Show that it holds under orthogonality that*

$$b_{2ASTc,k} = \frac{(1-\lambda_k)r^{\otimes}_{2c,k}}{(1-\lambda_k)r^{\otimes}_{2c,k} + \lambda_k} b_{OLS,k} + \frac{\lambda_k}{(1-\lambda_k)r^{\otimes}_{2c,k} + \lambda_k}\beta_{0,k}, \qquad X \perp .$$

Using that $\beta_0 = (X'X)^{-1}(X'X)\beta_0$, it follows that $b_{OLS} - \beta_0 = (X'X)^{-1}(X'\tilde{y}_0)$ for $\tilde{y}_0 = y - X\beta_0$. Similar to **2.A.3**, this means that

$$\begin{aligned} Q^{-1}_{2c,kk}s_0 &= 1/\Big(\sum_l \theta^k_l (b_{OLS,l} - \beta_{0,l})^2\Big)(\tilde{y}_0'\tilde{y}_0), \\ &= 1/\text{tr}\Big(\text{diag}(\theta^k)(X'X)^{-1}(X'\tilde{y}_0)(X'\tilde{y}_0)'(X'X)^{-1}(\tilde{y}_0'\tilde{y}_0)^{-1}\Big), \\ &= 1/\text{tr}\Big(\text{diag}(\theta^k)(X'X)^{-1}R^{\otimes}\Big). \end{aligned}$$

Take

$$b_{2ASTc} = (X'X + \frac{\lambda_k}{1-\lambda_k} Q_{2c}^{-1} s_0)^{-1}(X'y + \frac{\lambda_k}{1-\lambda_k} Q_{2c}^{-1} s_0 \beta_0).$$

Use the orthogonality of X to multiply each term inside and outside of the inverse by $(X'X)^{-1} = \frac{1}{N-1} I_K$, let $r^{\otimes}_{2c,k} = \text{tr diag}(\theta^k) R^{\otimes}$, and substitute $(X'X)^{-1} Q^{-1}_{2c,kk} s_0 = 1/r^{\otimes}_{2c,k}$ to get

$$b_{2ASTc,k} = (r^{\otimes}_{2c,k} + \frac{\lambda}{1-\lambda})^{-1}(r^{\otimes}_{2c,k} b_{OLS,k} + \frac{\lambda}{1-\lambda} \beta_{0,k}).$$

Multiply each term inside and outside of the inverse by $(1-\lambda_k)$ to obtain the desired result.

2.A.7: Table 2.3 presents the realizations of the simulated data in Section 2.4. The data ($N = 20$) are simulated with $y = X\beta + \epsilon$, $\beta = [2\ 2\ 0\ 0]'$, $\epsilon \sim \mathcal{N}(0,1)$, $X \sim \mathcal{N}(0, \Sigma)$, and $\Sigma = I$ except for $\Sigma_{\{2,1\},\{1,2\}} = 0.9$.

Table 2.3: *Untransformed Realizations of Simulated Data*

n	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
y	-4.47	-7.55	0.75	-5.61	-4.60	0.98	3.05	-1.02	1.65	-1.03	6.53	4.48	-4.20	0.18	-0.44	-2.87	0.72	3.07	9.92	0.57
x_1	-0.99	-1.57	0.54	-1.52	-1.54	0.33	0.63	-0.19	0.03	0.01	1.33	1.14	-0.61	0.00	0.23	-0.87	0.15	0.83	2.10	0.36
x_2	-0.62	-1.56	0.20	-1.13	-1.62	-0.01	0.77	-0.54	0.38	0.17	2.25	1.17	-1.18	-0.13	-0.30	-0.69	0.16	0.48	3.14	0.33
x_3	-0.34	-0.81	0.16	-0.26	0.01	-1.08	-0.86	0.34	0.64	-0.14	0.14	-0.04	-2.06	-0.96	-0.51	1.84	-1.04	1.39	0.68	-0.03
x_4	-1.99	2.20	0.16	-0.73	-1.13	-1.42	0.43	0.06	-0.41	-0.28	-0.11	1.91	-0.24	-0.77	0.28	0.88	1.88	-1.47	0.43	1.04

This table present realization of simulated data used in Section 2.4. Note that y has *not* been centered and X has *not* been transformed into Z-scores in this table. In the simulation exercise, the $N = 20$ data points are simulated with $y = X\beta + \epsilon$, $\beta = [2\ 2\ 0\ 0]'$, $\epsilon \sim \mathcal{N}(0,1)$, $X \sim \mathcal{N}(0,\Sigma)$, and $\Sigma = I$ except for $\Sigma_{\{2,1\},\{1,2\}} = .9$.

3

Accuracy-Simplicity Tradeoffs and the Linear Regression Model: b_{12} Astimators

3.1. Introduction

When estimating linear relations β between the dependent variable y and the independent variables X in the linear regression model, a researcher might want to balance his prior opinion β_0 with data-optimized solutions b_{OLS}. His incentive for doing so will be particularly strong when his prior is well-grounded in previous research, or when the data at hand is small or possibly heterogeneous. In Bayesian regression, a researcher can get some inkling of how his prior will be balanced with a data-optimized solution by appropriately transforming each variable. In Frequentist shrinkage methods, even this cumbersome strategy is generally unavailable.

A linear regression astimator allows a researcher to control the Accuracy Simplicity Tradeoff ('AST') between a data-optimized coefficient and a prior through λ. The tuning parameter λ determines the minimum amount of shrinkage towards β_0 if regressors are uncorrelated. A variable's contribution to R^2 accuracy determines to what extent parameters are further shrunk towards zero. Subset selection is approximated in this way, because irrelevant regressors with a low contribution to R^2 are barely allowed to deviate from β_0.

The flexibility of choosing regression parameters can be further restricted by a second AST which promotes the simplicity of grouping parameters together. By using a single regression coefficient (except for the sign) for multiple highly correlated regressors, one can hedge against the risk of wrongfully setting a relevant parameter to $\beta_0 = 0$. In case no distinction is made between high and low cross-correlations, irrelevant regressors will be compensated too liberally and subset selection will no longer be approximated. Through a tuning parameter called $c_{\min} \in [0,1]$, the researcher can specify how high cross-correlations need to be for parameters to be grouped together.

In the previous chapter I have developed astimators with an ℓ_2 norm that either focus on subset selection, grouping, or both. The latter was called a b_{2c} astimator, where the '2' refers to an ℓ_2 norm being used and the 'c' indicates that *correlations* are being accounted for. In this chapter, the same three variants will be investigated but now with an ℓ_1 norm. Methods with an ℓ_1 norm perform exact subset selection, which means that parameters will be inactivated (exactly equated to β_0) even before λ has reached it maximum value. Famous examples are the Adaptive Lasso and the Lasso.

One goal of the current chapter is to come to grips with the fickle behavior of the ℓ_1 based estimators. The Adaptive Lasso and the Lasso make it virtually impossible to anticipate before the data is analyzed at which λ values parameters will be added to the active set. I will present their astimated versions and thereby solve this issue. It will be shown that the moment at which a parameter is allowed to deviate from β_0 is directly related to its contribution to R^2 accuracy when regressors are uncorrelated. I will also analyze the influence of cross-correlations on subset selection and grouping. This discussion will lead to the introduction of a b_{1c} astimator.

To further promote the grouping of parameters while performing exact subset selection, one can also combine ℓ_1 and ℓ_2 norms. This is the basic idea behind the well-known Elastic Net. Like the other benchmark methods, the Elastic Net does not differentiate between high and low cross-correlations. As a result, it takes too long for irrelevant regressors to be inactivated and for highly correlated and relevant regressors to be grouped together. That is why a b_{12c} astimator (with an ℓ_1 and an ℓ_2 norm) is introduced, which does control both ASTs effectively.

Regarding the structure of this chapter, the next section introduces the ℓ_1 based benchmarks. The properties of their astimated analogues are discussed for uncorrelated regressors in Section 3.3 and correlated regressors in Section 3.4. Astimators that combine ℓ_1 and ℓ_2 norms are introduced in Section 3.5. The theoretical conjectures are examined with a simulation study and an empirical application in Section 3.6.

3.2. Lasso, Adaptive Lasso, and the Elastic Net

The linear regression model is given by

$$y = X\beta + \epsilon,$$

for an $N \times 1$ dependent variable y, and $N \times K$ matrix of regressors X, a $K \times 1$ vector of coefficients β and an $N \times 1$ vector of disturbances ϵ. The index $n = 1, 2, \ldots, N$ is used to refer to individual observations and the index $k = 1, 2, \ldots, K$ marks individual regressors.

Tibshirani (1996) introduced an estimator of β called the 'Lasso' (Least Absolute Shrinkage/Selection Operator), that is closely related to Ridge regression. The difference is that absolute rather than squared deviations from the prior $\beta_0 = \vec{0}$ are used in penalizing large b. The Lagrangian of the Lasso is

$$L_{Lasso} = \frac{1}{2}(y - Xb)'(y - Xb) + \lambda \sum_{k=1}^{K} |b_k|, \tag{3.1}$$

whereby $\frac{1}{2}$ is added for computational convenience. The Lasso is an interesting technique because it automatically performs subset selection by setting certain parameters exactly equal to zero even when λ has not reached its maximum value. Active parameters that are allowed to deviate from β_0 are denoted as b_A, where the active set is defined as $A = \{k : b_k \neq \beta_0, k = 1, 2, \ldots, K\}$. Following a suggestion by Tibsharini, the Bayesian Lasso uses a conditional Laplace prior. Data must be standardized by centering y and by taking Z-scores of X.

Among the many algorithms for solving the ℓ_1 based loss function are the pathwise coordinate descent algorithm and the Least Angle Regression ('LARS') (Osborne et al., 2000, Efron et al., 2004). The LARS procedure helps to interpret Lasso regression, because it shows that only those parameters are allowed to deviate from $\beta_0 = \vec{0}$ whose regressors have the largest correlation to the current residual $\tilde{y} = (y - X_A b_A)$. A LARS type algorithm has been developed whereby the λ values of the turning points are identified, see Zou et al. (2007) and Tibshirani jr (2011). Other than that a large $\lambda \in [0, \infty)$ will shrink solutions more strongly towards $\beta_0 = \vec{0}$ than a small λ, the interpretation of this tuning parameter has remained unclear. I have not come across procedures that allow the researcher's prior β_0 to differ from zero.

Two well-known extensions to the Lasso are the Adaptive Lasso and the

Elastic Net. Zou's (2006) Adaptive Lasso adds a vector of weights $\hat{w}$ to the penalty term,

$$L_{Adaptive\ Lasso} = \frac{1}{2}(y - Xb)'(y - Xb) + \lambda \sum_{k=1}^{K} \hat{w}_k |b_k|,$$

where $\hat{w}$ can be given by $\hat{w}_k = \frac{1}{|b_{OLS,k}|^\gamma}$ for some $\gamma > 0$. I will use $\gamma = 1$ for reasons that will become apparent later. The estimator has 'oracle properties', meaning that it asymptotically performs as well as if the true underlying submodel were given in advance (Fan and Li, 2001). It has been remarked that Lasso type estimators may have issues with highly correlated data. 'In practice,' write (Wang et al., 2011, pp. 471), 'Adaptive Lasso suffers (sometimes more severely than Lasso) from the multicollinearity caused by large correlations among covariates because OLS estimates are very unstable in this situation.'

To handle correlated regressors, many variations of the Lasso have been applied, like the Random Lasso (Wang et al., 2011), the Group Lasso (Yuan and Lin, 2006), and the 1d Fused Lasso (Tibshirani et al., 2005). Here, I will focus on the well-known Elastic Net, which Hastie developed together with Zou in 2005. In this variation, the penalty is a combination of ℓ_1 and ℓ_2 norms,

$$L_{Elastic\ Net} = (y - Xb)'(y - Xb) + \lambda \sum_{k=1}^{K} \Big(\frac{\alpha}{2} b_k^2 + (1 - \alpha)|b_k| \Big).$$

The formulation is based on Hastie et al. (2009, pp. 73). I have defined parameter α such that the Lasso is used when $\alpha = 0$ and that Ridge regression is employed when $\alpha = 1$. By choosing some intermediate value of α, one can perform subset selection and stimulate the grouping of correlated regressors.

Lastly, I will also refer to Zellner's g-prior (1986), where the prior specifications $\beta \sim \mathcal{N}(\beta_0, g\sigma^2(X'X)^{-1})$ and $\sigma^2 \propto \frac{1}{\sigma^2}$ result in a posterior mean of

$$b_{Zellner} = \frac{1}{1+g}\beta_0 + \frac{g}{1+g} b_{OLS}, \tag{3.2}$$

with $g \in [0, \infty)$. Defining $g = \frac{1-u}{u}$, one gets

$$b_{Zellner} = u\,\beta_0 + (1 - u) b_{OLS},$$

which makes it clear that Zellner's estimator amounts to taking a weighted average between β_0 and b_{OLS} with weights of $u \in [0, 1]$.

In the following, I will further analyze the behavior of estimators that make use of an ℓ_1 norm by rescaling the Lasso, the Adaptive Lasso, and the Elastic Net; and I will propose a b_{12c} astimator which combines an ℓ_1 and an ℓ_2 norm. I will argue, for example, that the fraction of a $\frac{1}{2}$ in ℓ_1 based estimators is necessary for interpretative purposes and that the Adaptive Lasso always suffers more from multicollinearity than the Lasso when it comes to discarding relevant but correlated regressors. I will also show that a regressor's contribution to the R-squared measure of fit plays a pivotal role, just like it did in the previous chapter.

3.3. b_1 Astimators: Uncorrelated Data

A generic formulation of an ℓ_1 based Lagrangian is given by

$$L_{1AST} = \frac{(y - Xb)'(y - Xb)}{(y - X\beta_0)'(y - X\beta_0)} + \sum_k 2\lambda_k \frac{|b_k - \beta_{0,k}|}{q_{1,k}}, \tag{3.3}$$

for $\lambda > 0$. A b_{1i} astimator, whereby the penalty of a regressor is barely affected by those of others, results from defining $q_{1i,k} = |b_{OLS,k} - \beta_{0,k}|$ in terms of *individual* ('i') deviances. Conditional on $\beta_0 = \vec{0}$, the associated loss function is a rescaled version of an Adaptive Lasso with weights $\hat{w}_k = \frac{1}{|b_{OLS,k}|}$. The loss function of b_{1a} is obtained by taking an equal-weighted *average* ('a') to define the index $q_{1a,k} = \frac{1}{K}\sum_l |b_{OLS,l} - \beta_{0,l}|$; and this is just a rescaled version of a more general b_{Lasso}.

The third ℓ_1 based astimator that I will examine here is called b_{1c} and is designed to perform subset selection and grouping more effectively by controlling the effects of *correlations* ('c'). The matrix Θ standardizes absolute cross-correlations $|\text{corr}(X)|$ by rescaling the columns to sum to one. If we define θ^k as the k^{th} column of Θ, then $q_{1c,k}$ becomes $\sum_l \theta_l^k |b_{OLS,l} - \beta_{0,l}|$. As I will illustrate below, the tuning parameter $c_{\min}$ equates cross-correlations lower than $c_{\min}$ to 0 in Θ, so that only highly correlated regressors are grouped together.

Analytic solutions are not available for ℓ_1 based loss functions, which is why I first followed Friedman et al. (2007) in formulating a coordinate descent algorithm, see Appendix 3.A.1. Based on this exercise, I subsequently developed Algorithm 3.1 in Appendix 3.A.2, which gives the entire solution path of ℓ_1 based astimators for known λ while prior coefficients β_0 may deviate from zero.

To explain how astimators perform subset selection, I have defined $R^2 \in [0, 1]$ in the previous chapter as the relative improvement of the data-optimized results over the prior model. So,

$$R^2 = 1 - \frac{(y - Xb)'(y - Xb)}{(y - X\beta_0)'(y - X\beta_0)},$$

which equals 0 if there is no improvement over the prior model and 1 if the data-optimized solutions have a perfect fit. In case the data is standardized and $\beta_0 = \vec{0}$, this measure is equal to the original R^2 'coefficient of determination'. The 'R-outer' matrix is given by

$$R^{\otimes} = (X'\tilde{y}_0)(X'\tilde{y}_0)'(X'X)^{-1}(\tilde{y}_0'\tilde{y}_0)^{-1},$$

with $\tilde{y}_0 = y - X\beta_0$. The matrix $R^{\otimes}$ helps to explain how astimators perform subset selection by only allowing certain parameters to deviate from their prior parameter.

More particularly, the diagonal elements of $R^{\otimes}$ were shown to add up to R^2. In this way, we can use $R^{\otimes}_{kk}$ to quantify how much an individual deviation from a prior hypothesis contributes to the overall improvement of the prior model. In case $\beta_0 = \vec{0}$, $R^{\otimes}_{kk}$ gives us a sense of how important each regressor is to R^2, although we do need to correct for cross-correlations to interpret the relevance of individual regressors. Due to cross-correlations, the diagonal elements might not lie between 0 and 1. For standardized data and $\beta_0 = \vec{0}$, I defined the matrix $R_{xx} = \text{corr}(X)$ and a vector of correlations $\vec{r}_{xy} = (r_{x_1y}\ r_{x_2y}\ \ldots\ r_{x_Ky})'$ to find that $R^{\otimes} = \vec{r}_{xy}\vec{r}_{xy}^{\,\prime}R_{xx}^{-1}$. This allowed me to show that, when regressors are orthostandard, the elements of $R^{\otimes}_{k,l}$ can simply be written as $r_{x_ky}r_{x_ly}$, which is the product between the regressors' correlations with y.

The same situations will be studied as in the previous chapter to describe how accuracy and simplicity terms influence solution paths of ℓ_1 based astimators. The current section expresses the astimators in terms of $R^{\otimes}$ in case regressors are uncorrelated. I will start with a single regressor and $\beta_0 = 0$ and I will subsequently relax these assumptions. In the section that follows, b_1 astimators are compared under multiple correlated regressors. I will finish by indicating how a LARS-type algorithm must be adjusted in case β_0 is allowed to vary from zero. Any remaining aspects are discussed in Appendix 3.A.3. It is assumed throughout that y is centered and X is standardized with Z-scores.

For a single regressor and $\beta_0 = 0$, one can use that the sign of b_{1AST} is equal to the sign of b_{OLS} to get rid of the absolute signs in L_{1AST} of equation (3.3).

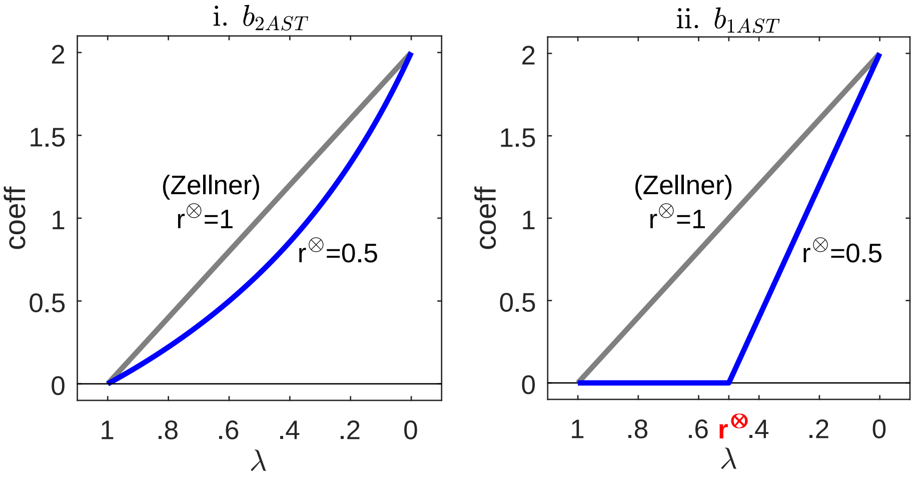
The solutions for $\lambda \in [0,1]$ are then given by

$$b_{1AST} = \begin{cases} (1-\lambda/r^{\otimes})b_{OLS} & \text{if } \lambda \leq r^{\otimes}, \\ 0 & \text{if } \lambda > r^{\otimes}, \end{cases} \tag{3.4}$$

where $r^{\otimes} \in [0,1]$ measures the degree to which x moves in the same (or opposite) direction as y; see **3.A.3.1**.

From equation (3.4) it becomes evident once more that $r^{\otimes}$ affects the proportion with which b_{OLS} is shrunk towards β_0 as λ increases. When $r^{\otimes} = 1$, the solutions of b_{1AST} are the same as Zellner's g-prior with $u = \lambda$. It follows that $\lambda \cdot 100\%$ specifies in percentage terms what the minimal influence of β_0 is. The moment that the regressor is activated is marked by $\lambda = r^{\otimes}$. It should be mentioned that if the true $\beta_0 = 0$ and $\epsilon \sim \mathcal{N}(0, \sigma^2 I_K)$, one may expect that b_{1AST} activates the parameter on average at $\lambda = \frac{1}{N-1}$. Under these conditions, $\mathbb{E}(R^2) = \frac{K}{N-1}$ with $\text{var}(R^2) = \frac{2K(N-K+1)}{N(N-1)^2}$ for standardized data.[1]

Figure 3.1: *Stylized Solutions of b_{2AST} and b_{1AST} with $r^{\otimes} = 0.5$ and $b_{OLS} = 2$*



This figure shows stylized solution paths for $b_{2AST} = (1+\frac{\lambda}{r^{\otimes}(1-\lambda)})^{-1}b_{OLS}$ and $b_{1AST} = (1-\lambda/r^{\otimes})b_{OLS}\mathbb{1}_{\lambda \leq r^{\otimes}}$ with $\beta_0 = 0$, $b_{OLS} = 2$ and $r^{\otimes} = 0.5$ or $r^{\otimes} = 1$. The upper line in each panel with $r^{\otimes} = 1$ corresponds to $b_{Zellner}$.

To compare how the relevance of a regressor influences the degree of shrinkage

[1] See http://davegiles.blogspot.nl/2013/10/more-on-distribution-of-r-squared.html for a small derivation.

towards $\beta_0 = 0$ for ℓ_2 and ℓ_1 based astimators, Figure 3.1 is presented. Remember that $b_{2AST} = (1 + \frac{\lambda}{r^{\otimes}(1-\lambda)})^{-1} b_{OLS}$. I have specified that $b_{OLS} = 2$ and that $r^{\otimes} = 0.5$ or 1. It can be observed that $b_{2AST} \approx b_{1AST}$ when λ is close to zero, and this occurs because the tangent line (and first-order Taylor expansion) of b_{2AST} at the point $\lambda = 0$ is given by $(1 - \lambda/r^{\otimes})b_{OLS} = b_{1AST}$. As a result, an irrelevant regressor with a low $r^{\otimes}$ is (approximately) equated to $\beta_0 = 0$ at $\lambda \approx r^{\otimes}$ for b_{2AST}. The effect of $r^{\otimes}$ in b_{2AST} abates as λ increases, as the current example with $r^{\otimes} = 0.5$ shows.

For b_{1AST} in the right panel, a given increase in λ leads to a decrease in b_{1AST} that remains directly proportional to $r^{\otimes}$. As indicated on the horizontal axis, b_{1AST} is activated precisely when $\lambda = r^{\otimes}$. A smaller angle between x and y would have brought the astimated solutions closer to $b_{Zellner}$ through a higher $r^{\otimes}$.

In the multivariate case, a decrease in λ causes an active $b_{1AST,k}$ parameter to move linearly in the direction of its restricted $b_{AOLS} = (X_A' X_A)^{-1}(X_A' y)$ solution, where the subscript A selects rows and/or columns of active parameters. This follows from minimizing L_{1AST} for the active regressors. To get rid of the absolute signs in L_{1AST}, one can replace $|b_k - \beta_{0,k}|$ by $z_k(b_k - \beta_{0,k})$; provided that z_k is -1 or $+1$ depending on whether $(b_k - \beta_0)$ is negative or positive. The first-order conditions then lead to

$$b_{1AST,A}(\lambda) = b_{AOLS} - \lambda (X_A' X_A)^{-1} Q_{1,A}^{-1} (y'y) z_A, \qquad \beta_0 = \vec{0}, \tag{3.5}$$

where the diagonal matrix Q_1 has diagonal elements of q_1. If we wish to use $b_{1AST,A}(\lambda)$, we only need to check whether the sign of z_k indeed corresponds to that of $(b_{1AST,k} - \beta_{0,k})$ for a given λ, and this will determine whether a regressor is active or not.

For our current discussion about the role of accuracy and simplicity in relation to b_{1i}, b_{1a}, and b_{1c}, it is sufficient to focus on the moment that inactive regressors become active. As a first step, we can further simplify the situation by assuming that the standardized regressors are *orthogonal* and that $\beta_0 = \vec{0}$. Starting with a current $\lambda_{\text{cur}} = \infty$, so that all regressors are inactive, it can be shown that regressor k is added to the active set at $\hat{\lambda} = \max\limits_{k,\ 0 \leq \tilde{\lambda}_k \leq \lambda_{\text{cur}}} \tilde{\lambda}_k$, where

$$\tilde{\lambda}_k(z_k) = \frac{z_k (x_k' y)(y'y)^{-1}}{Q_{1,kk}^{-1}}, \qquad k \in A^c, \beta_0 = \vec{0}, \perp X. \tag{3.6}$$

In words, inactive regressor k is activated at $\hat{\lambda}$ once its $\tilde{\lambda}_k$ is higher than the others but below the current λ_{cur} value.

To begin with the Adaptive Lasso type b_{1i} astimator, it can be remarked that the individual simplicity measure $Q_{1i,kk} = |b_{OLS,k} - \beta_{0,k}|$ is barely affected by other regressors. Even when variables are correlated and $\beta_0 \neq \vec{0}$, the first moment that a regressor (x_k) is activated can be rewritten as $\tilde{\lambda}_k = R^{\otimes}_{kk}$. In case $\beta_0 = \vec{0}$, its (initial) path is

$$b_{1ASTi,k} = \begin{cases} (1 - \lambda / R^{\otimes}_{kk}) b_{kOLS} & \text{if } \lambda \leq R^{\otimes}_{kk} \text{ and } \beta_0 = \vec{0}, \\ 0 & \text{if } \lambda > R^{\otimes}_{kk} \text{ and } \beta_0 = \vec{0}, \end{cases}$$

where $b_{kOLS} = (x'_k x_k)^{-1} x'_k y$, see **3.A.3.2** in the Appendix. If a regressor is completely uncorrelated with the active regressors, it will also be activated at $R^{\otimes}_{kk}$ and linearly move towards b_{kOLS}. As is shown in **3.A.3.1**, this holds true for $\beta_0 \neq \vec{0}$ as well. So, λ again determines the minimum influence of β_0 when regressors are uncorrelated; and that influence increases to the extent that β_0 is competitive to a data-optimized solution, as measured by the regressor's contribution to R^2 accuracy.

Next, consider the astimated version of the Lasso, which is the b_{1a} astimator with $Q_{1a,kk} = \frac{1}{K} \sum_l |b_{OLS,l} - \beta_{0,l}|$. Under orthogonality (we need not assume that $\beta_0 = \vec{0}$), each moment of activation occurs at

$$\tilde{\lambda}_k = \frac{1}{K} \sqrt{r^{\otimes}_k} \sum_{l=1}^{K} \sqrt{r^{\otimes}_l}, \qquad k \in A^c, \perp X, \tag{3.7}$$

see **3.A.3.1**. If we suppose that $\beta_0 = \vec{0}$, that the model has a perfect fit, and that each regressor k contributes equally ($r^{\otimes}_k = 1/K$), then all parameters are activated at $\frac{1}{K}$. In case only a subset of K_B parameters are equally responsible for the perfect fit, then this subset too will be activated at $\frac{1}{K}$, since $\tilde{\lambda}_k$ of a relevant regressor then equals $\frac{1}{K} \frac{1}{\sqrt{K_B}} K_B \frac{1}{\sqrt{K_B}} = \frac{1}{K}$. In general, equation (3.7) shows that the addition of an irrelevant regressor will always have a negative impact on the timing of when a relevant regressor is activated.

For the highlighted b_{1c} astimator with $Q_{1c,kk} = \sum_l \theta^k_l |b_{OLS,l} - \beta_{0,l}| = q_{1c,k}$, orthogonal regressors are added to the active set at the moment that

$$\tilde{\lambda}_k = \sqrt{r^{\otimes}_k} \sum_{l=1}^{K} \theta^k_l \sqrt{r^{\otimes}_l}, \qquad k \in A^c, \perp X, \tag{3.8}$$

where β_0 need not be a zero vector, see **3.A.3.1**. Since θ^k_l will be 1 at k and 0 otherwise ($\Theta = I_K$) when regressors are uncorrelated, $\tilde{\lambda}_k = r_k^{\otimes}$ and $b_{1c} = b_{1i}$.

To anticipate the behavior of b_{1c} in the case of correlated regressors, let us assume that there are $K = 4$ uncorrelated regressors; and that we want to stimulate the first two parameters to have a similar deviation from $\beta_0 = \vec{0}$ via

$$\Theta = \begin{pmatrix} 1/2 & 1/2 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

The first column $\theta^1 = [\frac{1}{2}\ \frac{1}{2}\ 0\ 0]'$ implies that the relative simplicity index of x_1 becomes $q_{1c,1} = \frac{1}{2}|b_{OLS,1}| + \frac{1}{2}|b_{OLS,2}|$; and since the second column is the same, $q_{1c,2} = q_{1c,1}$. Equation (3.8) subsequently shows that b_1 and b_2 will be activated at a $\tilde{\lambda}_k$ value that is closer to their average contribution to R^2, since $\tilde{\lambda}_1$ will then equal $\frac{1}{2}r_1^{\otimes} + \frac{1}{2}\sqrt{r_1^{\otimes} r_2^{\otimes}}$ instead of $r_1^{\otimes}$.

3.4. b_1 Astimators: Correlated Data

Next, I will study what happens when we relax the assumption that regressors are uncorrelated, so that $x_k' X_A (X_A' X_A)^{-1}$ can be different from zero. The expression for $\tilde{\lambda}_k$, which dictates when a regressor is activated, is then given by

$$\tilde{\lambda}_k(z_k) = \frac{z_k x_k' \Big(y - X_A b_{AOLS} \Big) (y'y)^{-1}}{Q_{1,kk}^{-1} - z_k x_k' X_A (X_A' X_A)^{-1} Q_{1,A}^{-1} z_A}, \qquad k \in A^c, \beta_0 = \vec{0}. \tag{3.9}$$

The numerator of $\tilde{\lambda}_k$ tells us how much x_k moves in the same (or opposite) direction as y once the explanatory potential of the active regressors $X_A b_{AOLS}$ has been deducted. The smaller the absolute correlation between x_k and this unexplained part of y, the larger $\tilde{\lambda}_k$, and the sooner x_k is activated. A conspicuous problem is that an inactive regressor that is highly correlated with active regressors will have little to add to the unexplained potential of y even though it will be nearly as relevant as these active regressors.

The denominator of $\tilde{\lambda}_k$ helps to deal with this issue, because $\tilde{\lambda}_k$ increases the more x_k is correlated with X_A. More specifically, the denominator of $\tilde{\lambda}_k$ becomes smaller when $Q_{1,kk}^{-1}$ is reduced by the diagonal elements of Q_{1A}^{-1}, and this occurs to the extent that x_k is correlated with the active regressors through $x_k' X_A (X_A' X_A)^{-1}$. So, a small denominator with high correlations between x_k and X_A brings the moment $\hat{\lambda}_k$ closer that regressor k is activated.

With these correlations in mind, the differences between an average and an individual simplicity measure can be further examined through equation (3.9). The diagonal elements of Q_{1a}^{-1} all have the same value by definition, so that $Q_{1a,kk}^{-1}$ will be compensated greatly by those $Q_{1a,A}^{-1}$ with which it is correlated.

When an individual simplicity index of b_{1i} is used, on the other hand, a given deviation from $\beta_{0,k} = 0$ is penalized less for a relevant regressor with a large $|b_{OLS,k}|$, because its $Q_{1,kk}^{-1} = 1/|b_{OLS,k}|$ is smaller. As a result, the diagonal elements of the active $Q_{1i,A}^{-1}$ will generally be smaller than those of the inactive $Q_{1i,kk}^{-1}$. The amount with which $Q_{1i,kk}^{-1}$ is lessened by $Q_{1i,A}^{-1}$ thereby decreases, so that regressor k will be added at a much lower $\tilde{\lambda}_k$ even though it might correlate just a bit less with y than the active regressors.

This tendency of the b_{1i} astimator to ignore relevant regressors might be a disadvantage when regressors are highly correlated, but b_{1i} also ensures that irrelevant regressors are inactivated at a low λ already. The b_{1a} astimator, by contrast, compensates irrelevant regressors too liberally, which in turn delays the grouping effect of relevant parameters.

To gain more control over the influence of cross-correlations, one can use a b_{1c} astimator, where $Q_{1c,kk}^{-1} = \left(\sum_l \theta_l^k |b_{OLS,l} - \beta_{0,l}| \right)^{-1}$. Remember that absolute cross-correlations lower than $c_{\min} \in [0,1]$ can be set to zero when computing the standardized correlation matrix Θ. In case $c_{\min} = 0$, the smallest of cross-correlations are of influence in estimating β_k. When all of the cross-correlations are below $c_{\min}$, b_{2c} is the same as b_{1i}. In the intermediate case, $Q_{1c,kk}$ of a relevant regressor is unaffected by an irrelevant regressor with which it is barely correlated.

To derive asymptotic correctness for the Lasso it is often assumed that $\max_{k \in A^c} ||x_k' X_A (X_A' X_A)^{-1}||_1$ are small so that irrelevant regressors do not influence the relevant ones too much (Hastie et al., 2009, pp. 91). With b_{1c}, I impose that small cross-correlations scarcely influence $\tilde{\lambda}_k$, because poor regressors will have a much smaller $Q_{1c,kk}$.

In Appendix 3.A.2, a computationally efficient solution procedure is derived for when regressors are activated and inactivated, whereby $\beta_{0,k}$ is also allowed to deviate from zero. In relation to LARS, I show under these more general conditions that the regressor is added to the active set which has the largest correlation to the current residual of $(y - X_A b_A - X_{\neg A} \beta_{0,\neg A})$; and that the solutions move in the direction of $\tilde{b}_{AOLS,k} = (X_A' X_A)^{-1} (X_A' (y - X_{\neg A} \beta_{0,\neg A}))$. To speed up LARS type algorithms that have been proposed before in the literature, the appendix also discusses in which cases and in which manner the moment should be computed that active regressors become inactive.

Finally, I will specify the range over which λ is defined. As aforesaid, the diagonal elements of $R^{\otimes}$ can easily be smaller than 0 when regressors are correlated; and in more unusual circumstances, they can be larger than 1. The only restriction that always holds is that the trace of $R^{\otimes}$ equals R^2. Having just shown that the first regressor is activated at the largest $R^{\otimes}_{kk}$ for b_{1i}, this means that λ need not lie below 1 for an ℓ_1 based astimator. Combining an ℓ_1 norm with an ℓ_2 norm forces $\lambda \in [0, 1]$.

3.5. b_{12} Astimators

The Elastic Net combines Ridge regression and the Lasso, because Ridge regression has a [stronger] tendency to assign a similar coefficient to highly correlated regressors and the Lasso has the advantage of performing exact subset selection. The astimated versions of the Ridge and Lasso loss functions are L_{2ASTa} and L_{1ASTa}, respectively. I will now take a weighted average between the latter two loss functions to obtain

$$L_{12ASTa} = \frac{(y - Xb)'(y - Xb)}{(y - X\beta_0)'(y - X\beta_0)} + \dots$$

$$\dots \lambda \sum_{k=1}^{K} \Big(\frac{\alpha}{1-\lambda} \frac{(b_k - \beta_{0,k})^2}{\frac{1}{K}\sum_l (b_{OLS,l} - \beta_{0,l})^2} + 2(1-\alpha) \frac{|b_k - \beta_{0,k}|}{\frac{1}{K}\sum_l |b_{OLS,l} - \beta_{0,l}|} \Big),$$

where $\alpha \in [0, 1]$ determines the weight that is assigned to the ℓ_2 norm ($\alpha \to 1$) relative to the ℓ_1 norm ($\alpha \to 0$). For $\alpha > 0$, b_k will move towards $\beta_{0,k}$ as $\lambda \to 1$, so that $0 \leq \lambda \leq 1$ once more.

There are two important differences between the b_{12a} astimator and the b_{2c} astimator of the previous chapter. First, the former combines two loss functions that are both defined in terms of an average simplicity measure ($L_{2AST\mathbf{a}}$ and $L_{1AST\mathbf{a}}$), while the latter combines $L_{2AST\mathbf{a}}$ with $L_{2AST\mathbf{i}}$. As I have just explained, loss functions with an average simplicity measure will not stimulate subset selection as effectively as an individual simplicity measure does.

Second, it remains obscure in b_{12a} how the tuning parameter α affects the manner in which correlated and irrelevant regressors are dealt with. An α of 0.75, say, does *not* even mean that alterations in b will for 75% be due to an ℓ_2 norm and for 25% to an ℓ_1 norm. What does $\alpha = 0.75$ imply in terms of the degree to which regressors receive a similar parameter value? How do cross-correlations between regressors affect such an inclination? Would it not be preferable to inactive irrelevant regressors as soon as possible and to only group highly correlated regressors?

In short, I expect that the b_{2c} astimator gives a better control over subset selection and grouping than b_{12a}. The exact subset selection of the latter astimator is a noteworthy aspect, but it comes with a price of solutions having to be approximated with a coordinate descent algorithm. The b_{2c} astimator has the distinct advantage of offering straightforward analytic solutions.

If a researcher is keen on performing exact subset selection, then he can use

$$L_{12ASTc} = \frac{(y - Xb)'(y - Xb)}{(y - X\beta_0)'(y - X\beta_0)} + \ldots$$
$$\ldots \lambda \sum_{k=1}^{K} \Big(\frac{\alpha}{1-\lambda} \frac{(b_k - \beta_{0,k})^2}{\sum_l \theta_l^k (b_{OLS,l} - \beta_{0,l})^2} + 2(1-\alpha) \frac{|b_k - \beta_{0,k}|}{\sum_l \theta_l^k |b_{OLS,l} - \beta_{0,l}|} \Big), \tag{3.10}$$

so that the b_{2c} astimator is combined with a b_{1c} astimator. The b_{2c} astimator will emphasize grouping more strongly (even when Θ is defined as I_K), whereas the b_{1c} astimator makes it easier to identify and avoid irrelevant deviations from β_0. So, when cross-correlations are high, an α close to 1 may be more optimal, and when cross-correlations are low, an α near 0 could be preferable.

Having studied differences between ℓ_1 based astimators theoretically, I will now continue to illustrate their behavior with a simulated data set and an empirical application.

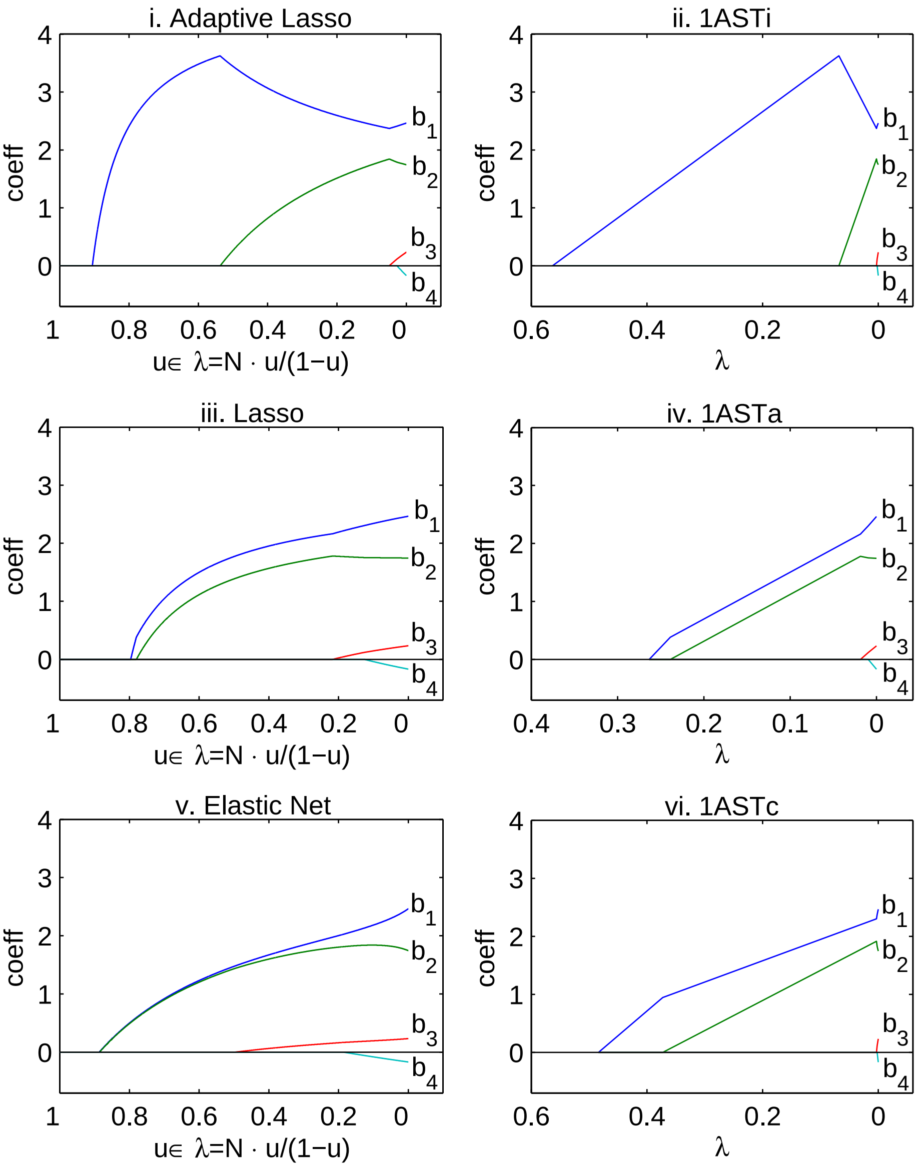
3.6. Analyzing the Influence of Tuning Parameters

3.6.1. *Simulation Studies*

To examine the postulated influence of tuning parameters on the astimated solutions, I will now turn to a simulation exercise whereby there are two relevant and highly correlated regressors and two irrelevant and uncorrelated regressors. A sample of twenty data points are simulated with $y = X\beta + \epsilon$, where $\beta = [2\ 2\ 0\ 0]'$, $\epsilon \sim \mathcal{N}(0,1)$, $X \sim \mathcal{N}(0,\Sigma)$, and $\Sigma = I$ except for $\Sigma_{\{2,1\},\{1,2\}} = 0.9$. The priors are defined as $\beta_0 = \vec{0}$. In the resulting data set, which was also used in the previous chapter, the diagonal elements of $R^{\otimes}$ equal $[0.56 \quad 0.40 \quad 0.02 \quad -0.00]'$ and the cross-correlations between x_1 and x_2 are 0.94.

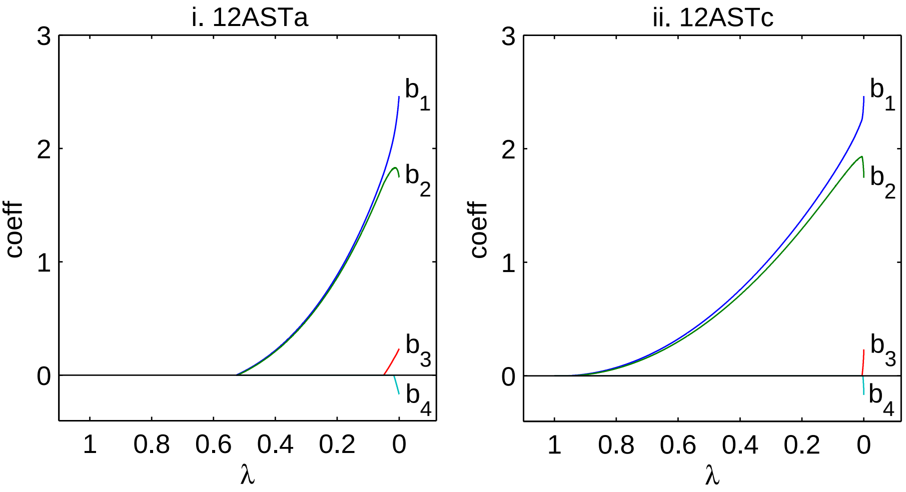
Figure 3.2 gives solutions paths for this data, whereby the independent variables are standardized with Z-scores and the dependent variable is centered.

Figure 3.2: *Solutions Paths* b_1



This figure shows solution paths for the benchmark estimators on the left hand side and ℓ_1 based astimators on the right hand side, with the selected coefficients on the vertical axis and the tuning parameter on the horizontal axis. The λ tuning parameter is occasionally defined in terms of a function of u for ease of display. Data ($N = 20$) are simulated with $y = X\beta + \epsilon$, $\beta = [2\ 2\ 0\ 0]'$, $\epsilon \sim \mathcal{N}(0,1)$, $X \sim \mathcal{N}(0,\Sigma)$, and $\Sigma = I$ except for $\Sigma_{\{2,1\},\{1,2\}} = .9$. Prediction model: $\hat{y} = Xb$, whereby $\beta_0 = [0\ 0\ 0\ 0]'$.

Figure 3.3: *Solutions Paths* b_{12}



This figure shows solution paths of astimators that combine ℓ_1 and ℓ_2 norms. Data ($N = 20$) are simulated with $y = X\beta + \epsilon$, $\beta = [2\ 2\ 0\ 0]'$, $\epsilon \sim \mathcal{N}(0, 1)$, $X \sim \mathcal{N}(0, \Sigma)$, and $\Sigma = I$ except for $\Sigma_{\{2,1\},\{1,2\}} = .9$. Prediction model: $\hat{y} = Xb$, whereby $\beta_0 = [0\ 0\ 0\ 0]'$.

The top panels present the Adaptive Lasso and its astimated analogue b_{1ASTi}. The only difference between these methods is the tuning parameter λ. Focusing on the astimated version, parameter b_1 is activated precisely at $\lambda = R^{\otimes}_{1,1} = 0.56$. Even though x_2 makes a small angle with y, the variable will have little to add to the explanatory potential of y once $x_1 b_1$ has been deducted, due to the large cross-correlation of 0.94. For this reason, b_{1ASTi} only allows the second regressor to deviate from zero once $\lambda = 0.07$.

The middle two panels of Figure 2.3 show the Lasso and its astimated version b_{1ASTa}. In line with the theory above, it takes longer for the Lasso type astimator to ignore irrelevant regressors than for the Adaptive Lasso type astimator. Although the Lasso type astimator does not isolate a single member from a group of correlated regressors, it does delay the moment that the relevant regressors are activated considerably when irrelevant regressors are included. Since $R^2 = 0.97$ is near perfect, the first two parameters are both activated at around $\lambda \approx 1/K = 0.25$.

The Elastic Net in the lower left panel of Figure 3.3 has the nice property that the first two parameters are joint together with an ℓ_2 norm while exact subset selection is performed with an ℓ_1 norm. I have used $\alpha = 0.5$. In comparing the

Elastic Net to the (Adaptive) Lasso, one can observe that it takes longer for the irrelevant regressors to be inactivated, which shows that subset selection is unnecessarily affected by a general tendency to shrink parameters to a similar value. It also takes quite a while for b_1 and b_2 to receive a similar value.

The b_{1ASTc} solutions in the lower right panel delete irrelevant regressors from the active set as quickly as b_{1ASTi}. Unlike the latter astimator, b_{1ASTc} does not merely focus on a single regressor out of a group of correlated regressors. Parameters b_1 and b_2 are activated at 0.48 and 0.36, which is close to their average contribution to the fit of the model. To make the grouping of the first two parameters as strong as in b_{1ASTa}, one could equate absolute correlations exceeding $c_{\min}$ to 1.

Next, Figure 3.3 presents the solutions of b_{12ASTa} and b_{12ASTc} with $\alpha = 0.5$. The former is closely associated with the Elastic Net. Unlike the Elastic Net, b_{12ASTc} deactivates irrelevant regressors at a low λ while the relevant regressors are slowly shrunk towards 0. In this example, the b_{12ASTc} solutions are highly similar to b_{2ASTc} of the previous chapter. The difference is that exact rather than approximate subset selection is now performed.

3.6.2. *Case Study: Diabetes*

Finally, I will discuss a well known empirical case study about $N = 442$ diabetes patients (Efron et al., 2004). In this application, a measure of how the disease progressed one year after a baseline is regressed on $K = 10$ baseline variables. The explanatory variables are age, sex, body mass index ('BMI'), average blood pressure ('BP') and six blood serum measurements ('S_1', 'S_2',...). This case study is typically used to illustrate that the Lasso can deactivate formerly active parameters as λ decreases.

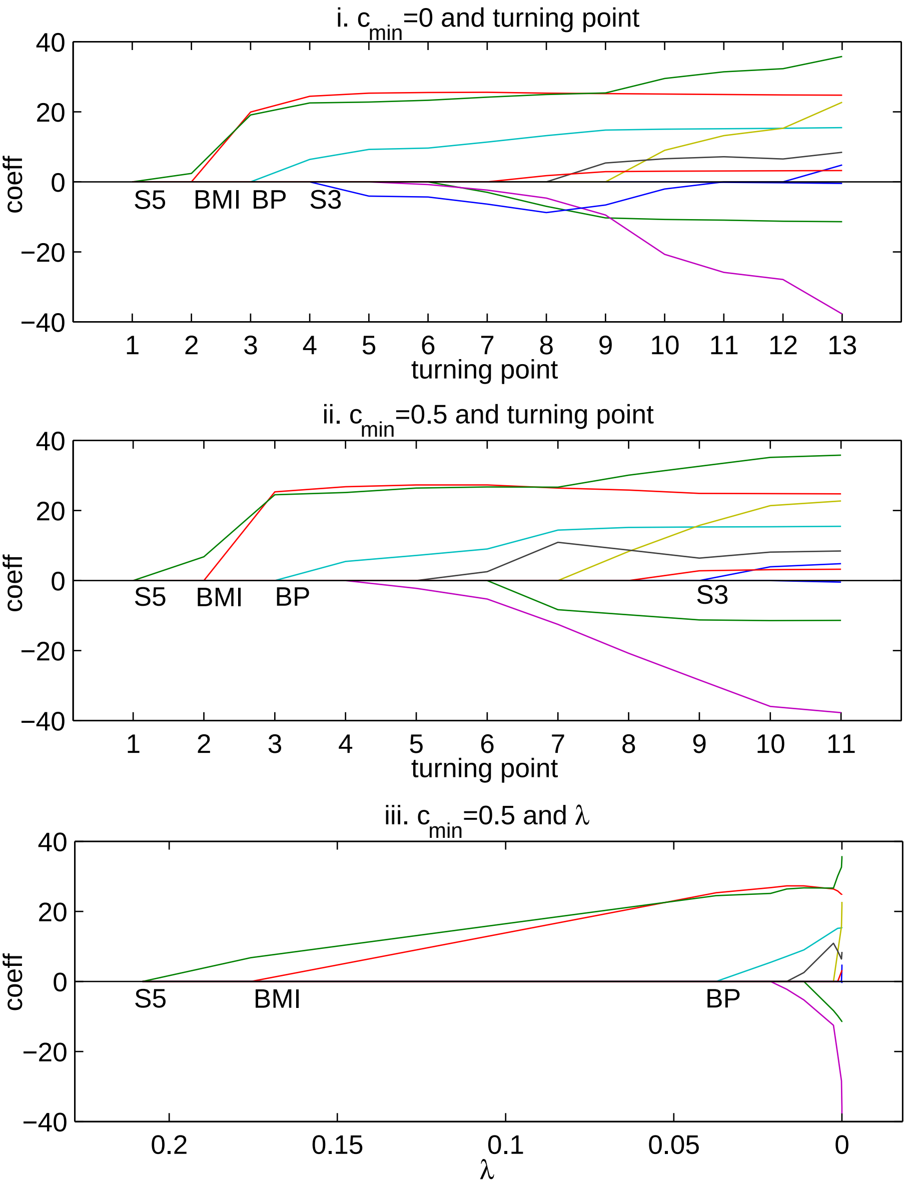
Table 3.1: *Description Regressors Diabetes*

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
	S_5	BMI	BP	S_3	S_1	sex	S_6	S_4	S_2	age
S_3	-.40	-.37	-.18	1	.05	-.38	-.27	-.74	-.20	-.08
$R^{\otimes}_{kk}$	.26	.19	.09	-.02	-.10	-.01	.02	.05	.05	-.00

This table presents the diagonal $R^{\otimes}$ values of the regressors of the diabetes data as well as the cross-correlations with the third blood serum measurement S_3. Regressors are enlisted in the order of when they are first added to the active set by b_{1c} with $c_{\min} = 0$ in the top panel of Figure 3.4.

Such behavior can also be observed in the top panel of Figure 3.4, where b_{1c} in equation (3.3) with $c_{\min} = 0$ is plotted against the turning points of when regressors are (de-)activated. The parameter of the third blood serum

Figure 3.4: b_{1c} *Astimation with Diabetes*



This figure illustrates how b_{1c} selects parameters of the Diabetes case study. Panel i uses $c_{\min} = 0$ and plots the coefficients against the turning points of when regressors are activated or deactivated. Regressor 'S3' is compensated here for having correlations (which are below 0.5) with the first three parameters that were activated. Panel ii is different from panel i in that $c_{\min} = 0.5$, so that 'S3' is no longer compensated for its correlations with the first three regressors. Panel iii uses $c_{\min} = 0.5$ like panel ii, but plots the coefficients against λ, which enables one to infer the contribution of a regressor to the fit of the model.

measurement S_3 is added to the active set at the fourth turning point, later deactivated at the eleventh turning point, and activated once more at the twelfth turning point. To explain why that happens, Table 3.1 presents cross-correlations with S_3, whereby the regressors are sorted in the order of when they are first added to the active set by b_{1c} with $c_{\min} = 0$. S_3 is activated quite early, because it is compensated for having correlations of -0.40 and -0.37 with the first two regressors. The subsequent change in sign results from the fact that regressors are later added with which S_3 is highly correlated.

The b_{1c} astimator enables the researcher to control such influences of cross-correlations on the grouping of parameters. By setting the minimum cross-correlation to be $c_{\min} = 0.5$, the variable S_3 receives no compensation for its cross-correlations with the first two regressors and is only activated at the ninth turning point. This is shown in the middle panel of Figure 3.4. Note that this panel does not indicate how the moment of activation of a regressor is related to its contribution to in-sample accuracy.

Astimators also make it easier to anticipate and interpret at which values of λ regressors are activated. This is illustrated in the bottom panel of Figure 3.4, where b_{1c} with $c_{\min} = 0.5$ is plotted against λ rather than the turning points. The moment of activation is closely related to the diagonal elements of $R^{\otimes}$. These values are presented in Table 3.1 and add up to $R^2 = 0.52$. The first regressor is activated at $\lambda = 0.21$ instead of 0.26, for example, due to its correlations with S_1 and S_4. After compensating for high-cross correlations with an ℓ_1 norm, it becomes clear that the fifth blood serum and the body mass index account for the lion's share of the model's in-sample accuracy.

3.7. Discussion

For Bayesian and Frequentist estimators it is nearly impossible to anticipate and influence how prior hypotheses are balanced with data-optimized solutions. Through the tuning parameter λ, astimators enable a straightforward interpretation of the moment that coefficients deviate from β_0. The benchmark estimators also have difficulties in controlling the second AST, where the simplicity of grouping parameters is promoted at the expense of in-sample accuracy. The reason is that these methods do not differentiate between high and low cross-correlations. The b_{1c} and b_{12c} astimators do allow a researcher to exercise control over the second AST through $c_{\min}$.

Now that the interpretation of λ has become clear, it ought to be investigated how a researcher's estimate of λ can be made more dependent on the data. The out-of-sample performances of the different estimators and astimators should also be compared.

3.A. Appendix: Further Details Regarding b_{1AST} and b_{12AST}

3.A.1. *Coordinate Descent Algorithm for ℓ_1 Based Astimators*

In this appendix I will describe how a coordinate descent algorithm can be used to find solutions of b_1 and b_{12} astimators. A general loss function of b_1 astimators is given by

$$L_{1AST} = \frac{(y - Xb)'(y - Xb)}{(y - X\beta_0)'(y - X\beta_0)} + 2\sum_{k=1}^{K} \lambda_k \frac{|b_k - \beta_{0,k}|}{q_1}. \tag{3.11}$$

To define q_1, one can think of $q_{1c,k} = \sum_l \theta_l^k |b_{OLS,l} - \beta_{0,l}|$, which has q_{1i} and q_{1a} as special cases. Following Friedman et al. (2007), I will concentrate on parameter b_k while keeping the other $b_{j \neq k}$ values fixed at $\tilde{b}_j$. The loss function can be then written as

$$L_{1AST} = \frac{(y - X_{\neg k}\tilde{b}_{\neg k} - x_k b_k)'(y - X_{\neg k}\tilde{b}_{\neg k} - x_k b_k)}{s_0} + \dots$$
$$\dots 2\Big(\lambda_k \frac{|b_k - \beta_{0,k}|}{q_{1,k}} + \sum_{j\neq k}^{K} \lambda_j \frac{|\tilde{b}_j - \beta_{0,j}|}{q_{1,j}}\Big),$$

We can now introduce a vector z_k that is -1 or $+1$ depending on whether the sign of $(b_k - \beta_{0,k})$ is negative or positive, so that we can drop the absolute signs and replace $|b_k - \beta_{0,k}|$ by $z_k(b_k - \beta_{0,k})$. The first-order condition $\frac{\partial L_{1AST}}{\partial b_k} = 0$ then leads to

$$\tilde{b}_k = (x_k' x_k)^{-1}\Big(x_k'(y - X_{\neg k}\tilde{b}_{\neg k}) - \lambda_k \frac{z_k}{q_{1,k}}\Big). \tag{3.12}$$

Before b_k can be equated to $\tilde{b}_k$, we need to check whether the sign of z_k indeed corresponds to the sign of $(\tilde{b}_k - \beta_{0,k})$. From equation (3.12), it is clear that $\tilde{b}_k$ moves linearly in the direction of $\tilde{b}_{kOLS} = (x_k' x_k)^{-1}(x_k'\tilde{y})$ as λ decreases to 0. If the signs between z_k and $(\tilde{b}_k - \beta_{0,k})$ do not match, all solutions to the $\tilde{b}_k$ side of $\beta_{0,k}$ (like $\tilde{b}_{kOLS}$) are illegitimate. Of the possible values of $|b_k - \beta_{0,k}|$, the choice of $b_k = \beta_{0,k}$ will then be closest to the 'optimal' but illegitimate $\tilde{b}_{kOLS}$. For a convex least squares problem, this means that b_k should be equated to $\beta_{0,k}$ if the sign of z_k is different from that of $(\tilde{b}_k - \beta_{0,k})$.

To check the sign of $(\tilde{b}_k - \beta_{0,k})$, we can deduct $\beta_{0,k}$ on both sides of equation (3.12) to get

$$\tilde{b}_k - \beta_{0,k} = (x_k' x_k q_{1,k}/s_0)^{-1}\Big(z_k \bar{\lambda}_k - \lambda_k\Big), \tag{3.13}$$

where

$$\bar{\lambda}_k = x_k'(y - X_{\neg k}\tilde{b}_{\neg k} - x_k\beta_{0,k})\frac{q_{1,k}}{s_0}. \tag{3.14}$$

The denominator is always positive in equation (3.13), because it only contains squared terms. A positive z_k therefore corresponds to a positive $(\tilde{b}_k - \beta_{0,k})$ when $\lambda_k \leq +\bar{\lambda}_k$. In that case, b_k can be equated to $\tilde{b}_k$ and otherwise $b_k = \beta_{0,k}$. From a negative z_k one gets $b_k = \tilde{b}_k$ as long as $\lambda_k \leq -\bar{\lambda}_k$; and $b_k = \beta_{0,k}$ otherwise. Hence the solution of L_{1AST} can be approximated by

$$\tilde{b}_{1AST,k}(\lambda_k) \leftarrow S(\tilde{b}_k, \beta_0, \lambda_k) \tag{3.15}$$

where

$$S(\tilde{b}_k, \beta_{0,k}, \lambda_k, \bar{\lambda}_k) = \begin{cases} \tilde{b}_k & \text{if } \lambda_k \leq |\bar{\lambda}_k| \\ \beta_{0,k} & \text{otherwise,} \end{cases} \tag{3.16}$$

where z_k is replaced by $\text{sign}(\tilde{b}_k - \beta_{0,k})$ in equation (3.12).

The following steps can now be used to astimate b_{1AST}. Select a parameter and update that parameter according to equation (3.15). Update the next parameter in line until convergence. Go from $\lambda = \lambda_{\max}$ to $\lambda_{\min}$ with incremental steps and use each solution as a warm start for the next; also when a single value of λ is of interest. Friedman et al. (2007) make use of a Tseng (1988). Here it is shown that coordinate descent algorithms converge to the minimizer of loss functions like

$$f(b) = g(b) + \sum_{k=1}^{K} h_k(b_k), \tag{3.17}$$

where $g(b)$ is convex and differentiable and the convex penalty term $\sum_{k=1}^{K} h_k(b_k)$ is a sum of functions of each separate parameter. As a convergence criterion, the program stops when there are no changes in the standardized parameters up to the second decimal place. The speed of the algorithm can be increased by computing terms like $x_k' x_k$ outside the loop.

The same steps can be used to derive the coordinate descent algorithm for a general L_{12AST} loss function with some q_k. To obtain the b_{12c} astimator, one can define $q_{2c,k} = \frac{1}{K}\sum_l \theta_l |b_{OLS,l} - \beta_{0,l}|$ and $q_{1c,k} = \frac{1}{K}\sum_l \theta_l (b_{OLS,l} - \beta_{0,l})^2$, for example. The first-order conditions of

$$L_{12AST} = \frac{(y - Xb)'(y - Xb)}{(y - X\beta_0)'(y - X\beta_0)} + \dots$$
$$\dots \frac{\lambda}{K} \sum_{k=1}^{K} \Big(\frac{\alpha}{1-\lambda} \frac{(b_k - \beta_{0,k})^2}{q_{2,k}} + 2(1-\alpha) \frac{|b_k - \beta_{0,k}|}{q_{1,k}} \Big),$$

lead to

$$\tilde{b}_{12AST,k} = \Big(x_k' x_k + \frac{\alpha\lambda}{1-\lambda} \frac{s_0}{q_{2,k}} \Big)^{-1} \Big(x_k'(y - X_{\neg k}\tilde{b}_{\neg k} - x_k \beta_{0,k}) + \dots$$
$$\dots \lambda s_0 \Big(\frac{\alpha}{1-\lambda} \frac{s_0}{q_{2,k}} \beta_{0,k} - \frac{(1-\alpha) z_k}{q_{1,k}} \Big) \Big). \tag{3.18}$$

Since the L_2 terms cancel in the numerator when $\tilde{b}_{12ASTi,k} - \beta_{0,k}$ is computed, one obtains

$$\bar{\lambda}_{12AST,k} = \frac{1}{1-\alpha} x_k'(y - X_{\neg k}\tilde{b}_{\neg k} - x_k \beta_{0,k}) \frac{q_{1,k}}{s_0}, \tag{3.19}$$

which is the same as equation (3.14) above except for $1/(1-\alpha)$. One can now use $S(\tilde{b}_{12ASTi,k}, \beta_0, \lambda_k, \bar{\lambda}_{12ASTi,k})$ again to cycle through parameters until convergence.

3.A.2. *Algorithm for Obtaining the Entire Solution Path of b_{1AST}*

Next, it will be explained how a complete solution path can be obtained for a loss function that uses an ℓ_1 norm in the simplicity measure. To summarize the notation that I will be using, let A select all active parameter rows and/or active parameter columns and let $\neg A$ select the inactive ones. Let the column vector z indicate by -1 or +1 the sign of $b_{1AST_k}(\lambda) - \beta_{0,k}$. Let $\neg k$ denote that parameter k is not included in a set, and let k refer to a row and/or column associated with the k^{th} regressor. Finally, let $b_{AOLS,\neg k}$ performs OLS with $X_{A,\neg k}$ instead of X_A. With this notation, Algorithm 3.1 gives an overview of how the entire solution path can be computed for a given λ. I will now explain the few steps that are required for its derivation. Any remaining details (**3.A.3.1**, **3.A.3.2**, ...) are presented in the next section.

Algorithm 3.1 Entire Solution Path of b_1 Astimators

```
Input: $\beta_0, X, Y, Q$
Output: $b_{1AST}(\lambda)$
Let $\tilde{b}_{AOLS} = (X'_A X_A)^{-1}\Big(X'_A(y - X_{\neg A}\beta_{0,\neg A})\Big)$,
    $b_A(\lambda) = \tilde{b}_{AOLS} - \lambda(X'_A X_A)^{-1} Q^{-1}_{1,A} s_0 z_A,$    (3.20)
the initial $\lambda_{\text{cur}} = \infty$, and let
    $\tilde{\lambda}_k(z_k) = \dfrac{z_k x'_k\Big(y - (X_A\tilde{b}_{AOLS} + X_{\neg A}\beta_{0,\neg A})_{\neg k} - x_k\beta_{0,k}\Big)s_0^{-1}}{Q^{-1}_{1,kk} - z_k x'_k\Big(X_A(X'_A X_A)^{-1}Q^{-1}_{1,A}z_A\Big)_{\neg k}}.$    (3.21)
  while $\hat{\lambda}_k \geq 0$ do
    1. Compute $\tilde{\lambda}_k$ for currently inactive regressors with
       $\max\{\tilde{\lambda}_k(-1), \tilde{\lambda}_k(+1)\}$.
    2. Only calculate $\tilde{\lambda}_k(z_{A,k})$ for a currently active regressor if the sign of a
       new $(\tilde{b}_{AOLS,k} - \beta_0)$ is different from the sign $z_{A,k}$ of $(b_{A,k}(\hat{\lambda}) - \beta_0)$.
    3. Switch activity status of $k^*$ at $\hat{\lambda} = \max_{k,\ 0 \leq \tilde{\lambda}_k \leq \lambda_{\text{cur}}} \tilde{\lambda}_k$.
    4. Compute $b_{1AST}(\hat{\lambda})$ with $b_A(\hat{\lambda})$ and $\beta_{0,\neg A}$.
  end while
```

Notes: Q is a K diagonal matrix, like $Q^{-1}_{1i,kk} = 1/|b_{OLS,k} - \beta 0, k|$; $s_0 = (y - X\beta_0)'(y - X\beta_0)$, A selects all active parameters rows and/or active parameter columns; z_A gives the sign (-1 or +1) for each active regressor based on the current sign of $(b_{A,k}(\hat{\lambda}) - \beta_0)$; $\neg k$ means that element k was excluded from a set (if applicable); $\tilde{b}_{AOLS,\neg k}$ performs $\tilde{b}_{AOLS}$ after k has been removed from X_A or $X_{\neg A}\beta_{0,\neg A}$. Premultiply $Q^{-1}_{1,kk}$ by a factor if you wish to penalize some regressors harder than others.

By minimizing L_{1AST} in equation (3.11) for active regressors, the solutions of $b_A(\lambda)$ are given by equation (3.20). From this expression it is clear that as λ moves towards zero, $b_{A,k}$ moves linearly in the direction of $\tilde{b}_{AOLS,k}$. To solve a b_1 astimator we just need to figure out at which λ values the set of active regressors is altered and connect the dots.

In the derivation of the coordinate descent algorithm in Appendix 3.A.1 above, it was shown that the set of active regressors can be defined as $A = \{k : \lambda < z_k\bar{\lambda}_k\}$, where

$$\bar{\lambda}_k = \frac{x'_k\Big(y - X_{\neg k}b_{\neg k}(\lambda) - x_k\beta_{0,k}\Big)s_0^{-1}}{Q^{-1}_{1,kk}}. \tag{3.22}$$

At a given λ, solutions of parameters other than k are represented by $b_{\neg k}(\lambda)$; and this vector is made up of the active $b_A(\hat{\lambda})$ and inactive β_0 parameters. Since the relevant $b_A(\lambda)$ are given by equation (3.20) and the values of β_0 are specified by the researcher, we can just substitute these expressions in $\bar{\lambda}_k$ and solve for λ in $\lambda < z_k\bar{\lambda}_k$. The solutions are given by $\tilde{\lambda}_k$ in equation (3.21), see **3.A.3.3** for a derivation. The next alteration of an activity status occurs at the highest $\tilde{\lambda}_k$ below the current λ_{cur}.

For presently inactive regressors, $\tilde{\lambda}_k$ can be computed by $\max\{\tilde{\lambda}_k(z_k = -1), \tilde{\lambda}_k(z_k = +1)\}$. Once parameter b_k has become active, equation (3.20) shows that it will move towards its $\tilde{b}_{AOLS,k}$ solution, so the sign in z_A corresponds to the sign of $\tilde{b}_{AOLS,k}$ of when k was most recently added to A. Active parameters can become inactive when $\lambda \geq z_k\tilde{\lambda}_k$, but this will only occur after the sign of $(\tilde{b}_{AOLS,k} - \beta_{0,k})$ has become different from the sign of $(b_{A,k} - \beta_{0,k})$ as a result of some new regressor being included or excluded. Merely in that case does one need to compute $\tilde{\lambda}_k(z_k = z_{A,k})$ to locate when active regressors become inactive. By altering the active set at the largest permissible λ_k in this way, the entire solution path of L_1 loss functions is quickly obtained.

In the current literature, the moment that a regressor is activated or inactivated is usually solved separately to find solutions akin to equation (3.21) (Zou et al., 2007, Tibshirani, 2011). If separate strategies are employed, then I suggest to make use of a similarity between triangles to show that an active b_k will hit β_0 at

$$\tilde{\lambda}_k = \lambda_{\text{cur}} \frac{|\tilde{b}_{AOLS,k} - \beta_{0,k})|}{|\tilde{b}_{AOLS,k} - \beta_{0,k}| + |b_{A,k}(\lambda_{\text{cur}}) - \beta_{0,k}|}, \tag{3.23}$$

which will be faster to compute than an equation like (3.21), because it merely contains scalars that are already available.

About the relationship between the LARS algorithm and the Lasso type b_{1a} astimator with $\beta_{0,k} = 0$, it is clear from condition $\lambda < z_k\bar{\lambda}_k$ with $\bar{\lambda}_k$ defined in equation (3.22), that regressor k is activated once its angle to the current residual, $x_k'(y - X_Ab_A)$, equals $Q^{-1}_{1a,kk}s_0z_k\lambda$. It follows that $Q^{-1}_{1a,kk}s_0z_A\lambda = X_A'(y - X_Ab_A)$ from substituting b_A by equation (3.20),

$$\begin{aligned} X_A'(y - X_Ab_A) &= X_A'y - (X_A'X_A)\Big((X_A'X_A)^{-1}X_A'y - \lambda(X_A'X_A)^{-1}Q^{-1}_{1,A}s_0z_A\Big), \\ &= Q^{-1}_{1a,kk}s_0z_A\lambda; \end{aligned}$$

and from using that $Q^{-1}_{1a,kk}$ is the same for all k.

The result is that a b_{1a} astimator adds a new regressor once its angle to the current residual is the same as those of the active regressors. What is more, the angles between X_A and $y - X_Ab_A$ are equal and they decrease monotonically as λ goes to zero. Add the fact that active regressors are inactivated once the sign of $b_A(\lambda) - \beta_0$ changes, and it becomes clear that the algorithm is equivalent to LARS for $Q_{1,a}$ and $\beta_{0,k} = 0$. The same logic can be applied in the more general case where β_0 is allowed to deviate from zero. The 'current residual' then becomes $(y - X_A b_A - X_{\neg A}\beta_{0,\neg A})$ and the solutions move in the direction of $\tilde{b}_{AOLS,k}$. Forward selection can be performed by ignoring step 2; and backward selection could be executed as well.

3.A.3. *Remaining Claims*

3.A.3.1: *Assume $\beta_0 = \vec{0}$ and orthogonal X and show that each moment of activation of a b_{2ASTc} parameter can be written as*

$$\tilde{\lambda}_k = \sqrt{r_k^{\otimes}} \sum_{l=1}^{K} \theta_l^k \sqrt{r_l^{\otimes}} \tag{3.24}$$

for $r_l^{\otimes} = (x_l'y)(x_l'y)'(x_l'x_l)^{-1}(y'y)^{-1}$, where $Q_{1c,k} = \sum_l \theta_l^k |b_{OLS,k}|$. Also prove that the results of $K = 1$, b_{2ASTi}, and b_{2ASTa} are special cases. Finally, show that the relation in (3.24) holds in case $\beta_0 \neq \vec{0}$, whereby $r_l^{\otimes} = (x_l'\tilde{y}_0)(x_l'\tilde{y}_0)'(x_l'x_l)^{-1}(\tilde{y}_0'\tilde{y}_0)^{-1}$ with $\tilde{y}_0 = y - X\beta_0$.

1. Use that $|a| = \sqrt{a^2}$ for some scalar $a \in \mathbb{R}$ and that $(X'X)^{-1} = \frac{1}{N-1} I_K$ for orthostandard data to rewrite

$$\begin{aligned}
\tilde{\lambda}_k &= z_k(x_k'y)(y'y)^{-1}Q_{1c,kk}, \\
&= (y'y)^{-1}|x_k'y| \sum_{l=1}^{K} \theta_l^k |(x_l'x_l)^{-1}(x_l'y)|, \\
&= \sqrt{(x_k'y)(x_k'y)'(x_k'x_k)^{-1}(y'y)^{-1}} \sum_{l=1}^{K} \theta_l^k \sqrt{(x_l'y)(x_l'y)'(x_l'x_l)^{-1}(y'y)^{-1}}, \\
&= \sqrt{r_k^{\otimes}} \sum_{l=1}^{k} \theta_l^k \sqrt{r_l^{\otimes}}.
\end{aligned}$$

2. Special cases:

- For $K = 1$ or b_{2ASTi}, define $\Theta = I_K$ to get $\tilde{\lambda}_k = \sqrt{r_k^{\otimes}}\sqrt{r_k^{\otimes}} = R_{kk}^{\otimes}$.
- For b_{2ASTa}, define $\theta_l^k = 1/K$ to get $\tilde{\lambda}_k = \frac{1}{K}\sqrt{r_k^{\otimes}} \sum_{l=1}^{K} \sqrt{r_l^{\otimes}}$.

3. In case $\beta_0 \neq \vec{0}$ (and X is orthogonal), we get

$$\tilde{\lambda}_k = z_k(x_k'\tilde{y}_0)(\tilde{y}_0'\tilde{y}_0)^{-1}\sum_{l=1}^{K}\theta_l^k|b_{OLS,l} - \beta_{0,l}|.$$

The derivation is the same as in point **1.** if it is recognized that

$$b_{OLS,l} - \beta_{0,l} = (x_l'x_l)^{-1}(x_l'\tilde{y}_0),$$

so that every y can be replaced by $\tilde{y}_0$.

3.A.3.2: *Given that the prior coefficients β_0 are all zero and that $Q_{1i,kk} = |b_{OLS,k} - \beta_{0,k}|$ is used; prove that the first activation occurs at the largest diagonal elements of*

$$R^{\otimes} = (X'y)(X'y)'(X'X)^{-1}(y'y)^{-1}.$$

Second, show that the (initial) path of the first activated regressor (x_k) is given by

$$b_{1ASTi,k} = \begin{cases} (1 - \lambda/R^{\otimes}_{kk})b_{kOLS} & \text{if } \lambda \leq R^{\otimes}_{kk}, \\ 0 & \text{if } \lambda > R^{\otimes}_{kk}, \end{cases}$$

for active $b_{kOLS} = (x_k'x_k)^{-1}(x_k'y)$ and a $1 \times K$ vector i that is 1 at k and zero otherwise. Third, if $\beta_0 \neq \vec{0}$, it holds that

$$R^{\otimes} = (X'\tilde{y}_0)(X'\tilde{y}_0)'(X'X)^{-1}(\tilde{y}_0'\tilde{y}_0)^{-1}$$

for $\tilde{y}_0 = y - X\beta_0$. Prove in this more general setting that the first activation also occurs at $\lambda = \max_k R^{\otimes}_{kk}$ when b_{1ASTi} is applied.

1. Parameter b_k is activated once $\tilde{\lambda}_k = (x_k'y)(y'y)^{-1}Q_{1i,kk} = R^{\otimes}_{kk}$.

$$\begin{aligned}\tilde{\lambda}_k &= (x_k'y)(y'y)^{-1}|b_{OLS,k}| \\ &= \sqrt{i(X'y)(X'y)'(X'X)^{-1}(y'y)^{-1}(X'y)(X'y)'(X'X)^{-1}(y'y)^{-1}i'}, \\ &= \sqrt{iR^{\otimes}R^{\otimes}i'} \\ &= R^{\otimes}_{kk}.\end{aligned}$$

whereby I have used that $|a| = \sqrt{a^2}$ for some scalar $a \in \mathbb{R}$.

2. The solution path is therefore given by

$$\begin{aligned} b_{1ASTi,k} &= (x_k'x_k)^{-1}\Big(x_k'y - \lambda Q^{-1}_{1i,kk}(y'y)\Big), \\ &= b_{kOLS} - \lambda(x_k'x_k)^{-1}(x_k'y)\frac{1}{(x_k'y)(y'y)^{-1}|b_{OLS,k}|}, \\ &= (1 - \lambda/R^{\otimes}_{kk})b_{kOLS}. \end{aligned}$$

In the second line, I multiplied the second term by $\frac{x_k'y}{x_k'y} = 1$. In the third line, I used that $\tilde{\lambda}_k = R^{\otimes}_{kk}$.

3. For $\beta_0 \neq \vec{0}$, see point **3.** in **3.A.3.1**.

3.A.3.3: *Equate* $\lambda = z_k\bar{\lambda}_k$ *and solve for* λ *to get*

$$\tilde{\lambda}_k(z_k) = \frac{z_k x_k'\Big(y - (X_A\tilde{b}_{AOLS} + X_{\neg A}\beta_{0,\neg A})_{\neg k} - x_k\beta_{0,k}\Big)s_0^{-1}}{Q^{-1}_{1,kk} - z_k x_k'\Big(X_A(X_A'X_A)^{-1}Q^{-1}_{1,A}z_A\Big)_{\neg k}}.$$

Using that $b_A(\lambda) = \tilde{b}_{AOLS} - \lambda(X_A'X_A)^{-1}Q^{-1}_{1,A}s_0z_A$, we can replace $X_{\neg k}b_{\neg k}$ by $X_Ab_A(\lambda) + X_{\neg A}\beta_{0,\neg A}$ in

$$\begin{aligned} \lambda &= z_kx_k'\Big(y - X_{\neg k}b_{\neg k} - x_k\beta_{0,k}\Big)Q_{1,kk}s_0^{-1}, \\ &= z_kx_k'\Big(y - (X_A\tilde{b}_{AOLS} + X_{\neg A}\beta_{0,\neg A})_{\neg k} - x_k\beta_{0,k}\Big)Q_{1,kk}s_0^{-1} + \ldots \\ &\ldots z_kx_kX_A(X_A'X_A)^{-1}Q^{-1}_{1,A}z_A)Q_{1,kk}\lambda. \end{aligned}$$

Bring the latter term to the left hand side and the result quickly follows.

4

Accuracy-Simplicity Tradeoffs and the Selection of Tuning Parameters

4.1. Introduction

In the preceding two chapters, I have argued that the choice of linear regression coefficients entails two instances of an Accuracy Simplicity Tradeoff. Unrestricted data-optimization is penalized in the first AST with the simplicity of the prior β_0 and in the second AST with the simplicity of grouping parameters together. The solutions of b_{12ASTc} in equation (3.10) of Chapter 3 allow researchers to control both ASTs through λ, $c_{\min}$, and α. The first tuning parameter λ determines the degree of shrinkage towards β_0, the second specifies with $c_{\min}$ how high cross-correlations need to be for coefficients to be grouped together, and the third parameter α indicates to what extent an ℓ_1 or an ℓ_2 norm is used in measuring deviations from β_0.

In this chapter I will use simulation studies to compare the out-of-sample forecasting performances of the different estimators and astimators. Before doing so, I will need to discuss how the tuning parameter λ can be selected. Astimators already make it easier to anticipate how λ affects a coefficient's degree of shrinkage towards β_0, because λ represents the minimum degree of shrinkage towards β_0 if regressors are uncorrelated.

The choice of λ could still be complicated because it can be influenced by many factors. On the one hand, the researcher might want to pick a high λ value to express his confidence in a prior that is well grounded in previous research. If the sample size is large, on the other hand, the researcher might trust the data-optimized solution more and opt for a lower λ. If his goal is mostly to group highly correlated regressors together and to delete variables that are irrelevant to the data generating process, then a value of λ close to zero might do.

To make the choice of λ more dependent on the data, one can make use of cross-validation and information criteria. For both techniques the researcher has to define a set of candidate λ values from which the optimal one is chosen. When ℓ_2 estimators are used, this set often has to be manually adjusted *a posteriori* (Friedman et al., 2010, pp. 17). The astimated versions solve this problem because they make the effect of λ easier to anticipate. When information criteria are applied, researchers also need to specify the effective number of parameters $\mathscr{K}$. The main procedures for doing so are to count the number of included regressors or to compute the effective degrees of freedom. As I will explain below, they have both become disputed. As an alternative, I will propose to use the relative simplicity measure of an astimator as its measure for $\mathscr{K}$.

The remainder of this chapter is set out as follows. In Section 4.2 it is discussed how λ can be astimated through cross-validation. In the context of information criteria, a new approach to measuring the effective number of parameters is introduced in Section 4.3. Hypotheses about the performance of the different estimators and astimators are presented in Section 4.4, and they are evaluated with simulation studies in Section 4.5.

4.2. Astimating λ Through Cross-Validation

Let the linear regression model be given by

$$y = X\beta + \epsilon,$$

where the dependent variable y and the residuals ϵ are $N \times 1$ vectors, where X is an $N \times K$ matrix of regressors, and where β is a $K \times 1$ vector of unknown coefficients to be estimated by b. I will also refer to the individual observation $n = 1, 2, \ldots, N$ and regressor $k = 1, 2, \ldots, K$.

Let me remind the reader that although many techniques have been introduced in the previous chapters, they only vary in two dimensions. The first dimension is controlled by α and determines the extent to which exact or approximate subset selection is performed through an ℓ_1 norm $(\alpha = 0)$ or an ℓ_2 norm $(\alpha = 1)$. The second dimension specifies whether a parameter is influenced by the deviance from β_0 of other parameters through an average (a), individual (i), or a correlation-based (c) simplicity measure.

Pivotal to the second dimension is Θ, which is a matrix of absolute cross-correlations where the columns θ_l^k are made to sum to 1. The researcher can specify cross-correlations lower than $c_{\min}$ to be 0 in Θ, so that parameters are only grouped together if their cross-correlations are high. One special case is $\Theta = I_k$, which implies that the penalty of each parameter is independent of the other parameters (b_{12i}). Another special case is that all Θ values equal $\frac{1}{K}$, so that a given deviation from β_0 is penalized by the same amount for each parameter (b_{12a}). The former leads to astimated versions of the Adaptive Lasso (b_{1i}) and the latter to new formulations of the Lasso (b_{1a}) and Ridge regression (b_{2a}). The Elastic Net is closely associated to b_{12a}.

The most general loss function of an astimator is given by

$$\begin{aligned} L_{12ASTc} = {} & \frac{(y - Xb)'(y - Xb)}{(y - X\beta_0)'(y - X\beta_0)} + \dots \\ & \dots \lambda \sum_{k=1}^{K} \Big(\frac{\alpha}{1-\lambda} \frac{(b_k - \beta_{0,k})^2}{\sum_l \theta_l^k (b_{OLS,l} - \beta_{0,l})^2} + 2(1-\alpha) \frac{|b_k - \beta_{0,k}|}{\sum_l \theta_l^k |b_{OLS,l} - \beta_{0,l}|} \Big), \end{aligned}$$

whereby λ strikes the balance between relative accuracy (first term right hand side) and relative simplicity (second term right hand side).

Cross-validation is a well known technique for choosing configurations like λ, α, and $c_{\min}$. In cross-validation, the sample is split into a training sample and a validation sample. The model is estimated with a training sample and these estimates are used to 'predict' the outcomes of the validation sample. By varying the choice of a tuning parameters, one can select the set of configurations that leads to the best pseudo-out-of-sample forecasts. In 10-fold cross-validation, the sample is randomly assigned to 10 folds and fold f is predicted using the other folds for each $f = 1, 2, \dots, 10$, so that each observation is predicted one time.[1]

One disadvantage of cross-validation is that there could still be quite some variability (uncertainty) in the data-optimized choice of the tuning parameter, particularly when the sample size is small. Based on previous experience or theoretical arguments, a researcher might also have good grounds for being more confident in his prior estimate of λ_0 than in the cross-validated alternative. To come to grips with the uncertainty of cross-validation, a tuning parameter γ can be introduced to determine how large a relative increase in accuracy must be to justify a relative deviation from a researcher's hyperparameter λ_0. The loss

[1] This strategy can also be applied in time series forecasting. Alternatively, one could split the sample in a training sample $[1, N - v]$ and a validation sample ($[N - v + 1, N]$) for $v = 15$, say, and produce pseudo predictions with the training sample regarding the validation sample for varying choices of λ. The best λ can then be selected.

function for astimating λ is

$$L_{\lambda_i} = \frac{\text{MSFE}(b_{\lambda_i})}{\text{MSFE}(b_{\lambda_0})} + \frac{\gamma}{1-\gamma}\frac{(\lambda_i - \lambda_0)^2}{\max\{(1-\lambda_0)^2, \lambda_0^2\}}. \tag{4.1}$$

I will give two examples of how λ can be astimated. First, one can think of the stylized fact that an equally weighted combination of forecasts is hard to beat (Bates and Granger, 1969, Smith and Wallis, 2009). This means that $\beta_{0,k} = \frac{1}{K}$ is a strong prior in a forecasting combinations exercise. L_{λ_i} could then be configured with $\lambda_0 = 1$ and $\gamma = 0.5$, so that cross-validated accuracy only has an influence of 50% relative to the simplicity of $\lambda_0 = 1$.

As a second example, it might be expected that the cross-validated choice of λ becomes more volatile when the sample size is as small as $N = 10$, say. With such a small sample size, leaving out a fold of observations can have large effects on the simplicity index $q_{1c} = \sum_l \theta_l^k |b_{OLS,l} - \beta_{0,l}|^1$ (and q_{2c}), which depends on the estimated b_{OLS} and the estimated cross-correlations. Similar problems may occur for the Adaptive Lasso, incidentally, where $|b_k|$ in the penalty term is divided by $|b_{OLS,k}|$. Due to the variations caused be excluding a fold of observations, deviations from β_0 can be penalized too severely (Hastie et al., 2009, pp. 243). Using that irrelevant parameters will be inactivated at a low λ, one can restrain the size of cross-validated λ values by defining $\lambda_0 = 0$. The tuning parameter can then be astimated with a great degree of confidence of $\gamma = 0.9$ or 0.99.

To apply cross-validation, one needs to define a candidate set of λ values from which the optimal one is selected, and this has particularly caused trouble for ℓ_2 based estimators (Ridge regression), where readjustments are frequently required. Astimators help to solve this issue, because they make it is easier to anticipate at which values of λ irrelevant regressors will be inactivated. It should be noted that an ℓ_1 norm may activate the first parameter at a $\lambda_{\max}$ that is smaller (or bigger) than 1.[2] The moment that irrelevant parameters are (approximately) equated to $\beta_0 = \vec{0}$ has been shown to be determined by the diagonal elements of $R^{\otimes} = (X'y)(X'y)'(X'X)^{-1}(y'y)^{-1}$. Since an irrelevant regressor with a low $R^{\otimes}$ will be inactivated at a low λ, the candidate configurations had best grow exponentially from 0 to $\lambda_{\max}$.

Accordingly, I will define $P = 101$ candidate λ values with

$$\lambda_{cand} = mg^p - m, \tag{4.2}$$

[2] In case $\alpha > 0$, $\bar{\lambda}_{12AST,k}$ in equation (3.19) can only be used if it is less than or equal to 1.

where $m = 10^{-4}$, $g = (\frac{\lambda_{max}+m}{m})^{1/(P-1)}$, and $p = 0, 1, \ldots, P-1$. Note that $mg^0 - m = 0$ and that the choice of g follows from solving $mg^{P-1} - m = \lambda_{\max}$. The smaller m, the smaller the initial increase in candidate λ values. In case moments of (de-)activation are exactly calculated for ℓ_1 based astimators, then I will include these in the candidate set of λ values and adjust P in equation (4.2), so that the set still contains 101 members.

For the original ℓ_1 based estimators it will also be exploited that $\lambda_{\max}$ can be computed. Following the Matlab Lasso (2016) package, I will define

$$\lambda_{cand} = \exp\Big(\log(\lambda_{\min} : \mathrm{st} : \log(\lambda_{\max})\Big),$$

with $\lambda_{\min} = m\lambda_{\max}$ and $\mathrm{st} = (\log(\lambda_{\max}) - \log(\lambda_{\min}))/(N_\lambda - 1)$. The notation $a : st : b$ means from a to b with steps of st. Once λ_{cand} is computed, I will replace the first element of λ_{cand} by 0 to ensure that b_{OLS} is one of the candidate solutions as well. The advantage of equation (4.2) is that it leads to a more gradual increase from $\beta_0 = 0$. The optimal λ value for Ridge regression is selected from $\lambda = 10^u$ with $u \in [-5, 5]$. For all techniques I will set $b = \beta_0$ when $\lambda = 1$.

Lastly, the candidate set of $c_{\min}$ values will be defined as $\{0, 0.1, 0.2, \ldots, 1\}$. The actual number of unique evaluations for $c_{\min}$ is typically smaller, because Θ might be exactly the same when $c_{\min} = 0.4$ or 0.7, say. The α of b_{12a} and the Elastic Net is selected from $\{0, 0.1, 0.2, \ldots, 1\}$. The b_{12c} astimator requires one to specify three configurations (λ, α, and $c_{\min}$). To reduce computational costs, I will choose α from $\{0, 0.25, 0.5, 0.75, 1\}$. Note that $c_{\min}$ and α could also be astimated through equation (4.1) by specifying $c_{\min,0} = 0.5$ and $\alpha_0 = 0.5$. One can also increase or decrease α_0 depending on whether cross-correlations are high or not.

4.3. Information Criteria

A faster method for choosing λ than cross-validation is to make use of information criteria. A popular Bayesian IC was introduced by Schwarz (1978),

$$BIC = \log\left[\frac{1}{N}(y - Xb)'(y - Xb)\right] + \frac{\log N}{N}\mathscr{K}.$$

Observe that an IC balances the fit of a model with the number of effective parameters $\mathscr{K}$ based on the sample size N. To penalize the inclusion of additional parameters, one selects b with the lowest IC value. The main difficulty is that one has to specify what the effective number of parameters $\mathscr{K}$ is. I will now discuss the critique that has been raised against the two available methods for measuring $\mathscr{K}$ and propose a convenient alternative.

4.3.1. *Counting the Number of Included Parameters*

In discrete subset selection procedures, a regressor is either included unrestrictedly or excluded altogether. When an IC is applied to determine the best subset of regressors, $\mathscr{K}$ has typically been measured by simply counting the number of active parameters K_A in the candidate models. $K_A = 2$ if there are two regressors in the subset, for example. Note that this measure of the effective number of parameters is only influenced by whether a regressor deviates from β_0 (the first AST), while no corrections are made for parameters of highly correlated regressors being grouped together (second AST).

To compare different ways of measuring $\mathscr{K}$, it will be convenient to have a gradual version of K_A that can be applied to shrinkage methods. To this end, one can calculate what the parameter estimate $b_{\lambda,k}$ would have been if it were possible to freely optimize over the data conditional on the other parameters $(\neg k)$ being equal to $b_{\lambda,\neg k}$ for a given value of λ. Treating $X_{\neg k} b_{\lambda,\neg k}$ as fixed, we can regress $y - X_{\neg k} b_{\lambda,\neg k}$ on x_k to get

$$\tilde{b}_{rOLS,k} = (x_k' x_k)^{-1}\Big(x_k'(y - X_{\neg k} b_{\lambda,\neg k})\Big).$$

It follows that parameter k is free conditional on $X_{\neg k} b_{\lambda,\neg k}$ if $b_k = \tilde{b}_{rOLS,k}$. The extent to which a given parameter is data-optimized is then given by $\frac{|b_k - \beta_{0,k}|}{|\tilde{b}_{rOLS,k} - \beta_{0,k}|}$, so that a tentative measure for $\mathscr{K}$ becomes

$$\hat{\mathscr{K}}_{1ir} = \sum_{k=1}^{K} \frac{|b_k - \beta_{0,k}|}{|\tilde{b}_{rOLS,k} - \beta_{0,k}|}. \tag{4.3}$$

The subscript '$1ir$' says that an ℓ_1 norm is used to measure *individual* deviances between $\beta_{0,k}$ and the *restricted* $\tilde{b}_{rOLS,k}$ solutions.

For many estimation procedures, the solution b_k will always lie between $\beta_{0,k}$ and $\tilde{b}_{rOLS,k}$. The conditions are a special case of those delineating whether a coordinate descent algorithm will optimize over a loss function, see equation (3.17) in Appendix 3.A.1. The first condition is that the accuracy measure must be defined in terms of the sum of squared residuals. The second condition is that the (convex) penalty term must be separable into a sum of functions of each individual parameter. Of all the estimators and astimators that have been presented in the previous chapters, only Zellner's g-prior does not meet the latter requirement. For the other techniques, the contribution of parameter k to $\hat{\mathscr{K}}_{1ir}$ always lies between 0 and 1; and $0 \leq \hat{\mathscr{K}}_{1ir} \leq K$.[3] If, for example, 2 out of 4 parameters are equal to their prior and the other two are unpenalized, then $\hat{\mathscr{K}}_{1ir} = 2$ again.

4.3.2. *Degrees of Freedom*

The $\hat{\mathscr{K}}_{1ir}$ measure generalizes a common usage of equating $\mathscr{K}$ to the discrete number of included parameters K_A, by allowing parameters to be partially included and β_0 to be different from zero. Nevertheless, it goes against decades of research which emphasizes that the number of 'degrees of freedom' of a regression does not equal the number of included parameters when a data-optimized selection procedure is performed. Although the chosen model of a discrete subset selection procedure has K_A parameters, Hastie et al. (2009, pp. 77) feel that 'in some sense' we have used up more than K_A degrees of freedom here.

Accordingly, Mallows (1973) and Stein (1981) already developed a covariance penalty as part of their information criteria with the idea that the covariance between the estimated $\hat{y}$ and the observed y gets larger the harder we fit the data. Ye (1998) went on to define degrees of freedom as

$$\mathscr{K}_{DF} = \frac{1}{\sigma^2} \sum_{n=1}^{N} \text{cov}(\hat{y}_n, y_n), \tag{4.4}$$

which equals $\frac{1}{\sigma^2}$ tr $\text{cov}(\hat{y}_n, y_n)$ for i.i.d. errors with finite variance σ^2 (Efron, 1986).

In OLS, for instance, $\hat{y} = Hy$, whereby the hat-matrix $H = X(X'X)^{-1}X'$ is known from outlier-detection. The OLS degrees of freedom is therefore given by $K = \text{tr } H$ and represents the sum of the sensitivities of the fitted values of $\hat{y}_n$ with respect to the observed y_n. Ye (1998) showed that the covariance penalty can be computed for other methods as well through a computationally intensive parametric bootstrap procedure that measures how sensitive the fitted $\hat{y}_n$ is to adding a random value to the observed y_n.

For some methods, closed-form approximations have been derived. Similar to b_{OLS}, $\mathscr{K}_{DF}$ for Ridge regression is known to be $\hat{\mathscr{K}}_{DF,Ridge} = \text{tr } H_{Ridge}$ with $H_{Ridge} = X(X'X + \lambda I_K)^{-1}X'$. For a b_{2c} astimator, $\hat{\mathscr{K}}_{DF,2c}$ equals

[3] For the most general of astimators b_{12ASTc}, equation (3.18) shows that $\tilde{b}_k$ moves from $\beta_{0,k}$ to $\tilde{b}_{OLS,k}$ as λ decreases to 0.

tr $X(X'X + \Lambda Q_{2c}^{-1} s_0)X'$, where Q_{2c} is a diagonal matrix with diagonal elements of $\sum_{l=1}^{K} \theta_l^k (b_{OLS,l} - \beta_{0,l})^2$ and $s_0 = (y - X\beta_0)'(y - X\beta_0)$. Zou et al. (2007) proved that an asymptotically ('∞') unbiased estimate of the Lasso's degrees of freedom $\hat{\mathscr{K}}^{\infty}_{DF,Lasso}$ for a fixed λ and X is simply given by the number of active parameters that deviate from $\beta_0 = 0$. They also showed that the IC only has to be evaluated at the transition points of when the set of active regressors alters.

When $K = 1$ and $\beta_0 = 0$, for example, Ridge regression and the Lasso will result in the same set of candidate regression coefficients from which an IC chooses the optimal solution, but $\hat{\mathscr{K}}_{DF,Ridge}$ will gradually increase from 0 to 1, while $\hat{\mathscr{K}}^{\infty}_{DF,Lasso}$ already equals 1 after the slightest deviation of b_{Lasso} from 0. The degrees of freedom is higher for b_{1a} (the Lasso), because its solutions are directly proportional to R^2 accuracy, while for b_{2a} (Ridge regression) the influence of R^2 dies out as λ increases. So, once a parameter is activated, the fitted values of the Lasso are far more sensitive to changes in y at a fixed λ. The implication of setting $\hat{\mathscr{K}}^{\infty}_{DF,Lasso} = 1$ for the active Lasso solutions is that the choice will be between β_0 and b_{OLS}, since in-sample fit is monotone nondecreasing for λ. One curious aspect of this result is that it makes it seem as if *gradual* transitions from β_0 to b_{OLS} are irrelevant when the *shrinkage* estimator is applied to K equally relevant regressors.

Kaufman and Rosset (2014) and Janson et al. (2015) recently questioned the validity of $\mathscr{K}_{DF}$. These researchers were disconcerted by the discrepancy between $\mathscr{K}_{DF}$ and measures of model complexity like those of Lasso and Ridge regression. In case an active regressor is inactivated while λ is decreased, for instance, $\hat{\mathscr{K}}^{\infty}_{DF,Lasso}$ (the number of active regressors) decreases while the Lasso's model complexity (defined in terms of λ) increases.[4] For a best subset selection that iteratively deletes the worst regressor based on an IC, $\hat{\mathscr{K}}_{DF}$ can become larger than K for a given subset, which is another clear breach of monotonicity. The title of Janson et al. (2015) summarizes their verdict: 'Effective Degrees of Freedom: a Flawed Metaphor.' It thus appears that we have reached a dead end in the use of ICs, now that the two customary methods for estimating $\mathscr{K}$ are considered to be inappropriate.

To find a way out of the current predicament concerning the measurement of $\mathscr{K}$, let us first try to formulate more clearly in *what* sense subset selection can take up more than K_A parameters. As an example, think of the famous F-test being used to examine whether some x_{new} should be added to the model

[4] To ensure such a decrease in $\hat{\mathscr{K}}_{DF}$, the authors independently showed how a data set can be constructed for which a variable is consistently removed around the same λ value.

or not. To specify $\mathscr{K}$ in this procedure, the mainstream solution is to count the number of active parameters $\hat{\mathscr{K}}_A$ under $H_0 : \beta_{new} = 0$ and under some alternative hypothesis. A problem is that a relevant regressor may not have a 'significant' contribution to the accuracy of a model when it is moderately correlated to many other regressors, or highly correlated to a few others. In these situations, the negligible increase in R^2 as a result of adding the new regressor will easily be outweighed by the penalty for increasing $\hat{\mathscr{K}}_A$ by 1.

Failing to include such relevant regressors not only deteriorates forecasting performance, but will also affect one's ability to understand and influence how outcomes are truly generated. Selecting a single parameter from a group of highly correlated regressors may for that reason well be penalized harder than a 'large' model that groups these parameters together. Where the focus has often been on the first AST about deviations from β_0, it is the second AST of grouping parameters together that can explain why subsets need to be penalized more heavily. As I have mentioned before, the grouping of coefficients can even be seen as a form of dimension reduction, because it amounts to using a single parameter for multiple regressors (apart from the sign). Although the critique of Kaufman and Rosset (2014) and Janson et al. (2015) continues to ignore the second AST, it does underline that $\mathscr{K}_{DF}$ makes it difficult to anticipate and influence how model complexity is penalized.

In relation to K_A of the previous section, it can be remarked that the degrees of freedom of Ridge regression equals the naive $\hat{\mathscr{K}}_{1ir}$ under the assumption that regressors are orthostandard. When y is centered and the orthogonal X are standardized with Z-scores, that is,

$$\begin{aligned}\hat{\mathscr{K}}_{1ir,Ridge} &= \sum_{k=1}^{K} \frac{|b_{Ridge,k}|}{|b_{OLS,k}|},\\ &= K\frac{N-1}{\lambda+N-1},\\ &= \hat{\mathscr{K}}_{DF,Ridge},\end{aligned}$$

see **4.A.1**. Since b_{Ridge} equals $b_{Zellner}$ when $\lambda = \frac{N-1}{g}$ and when regressors are orthostandard (section 2.2), b_{Ridge} just shrinks parameters to zero without regard for the data, see also Tibshirani jr. (2015).[5] No additional penalty for the uncertainty of a data-optimized subset selection is therefore required, so that an estimate of $\hat{\mathscr{K}}_{DF,Ridge}$ may well be equated to the naive $\hat{\mathscr{K}}_{1ir,Ridge}$, which

[5] Using a different framework, Tibshirani jr. (2015) makes the more general claim that Ridge regression has zero 'search degrees of freedom'.

just counts to what extent each parameter is included in accordance with the first AST. Now I will examine how the second AST can be incorporated when counting the effective number of parameters.

4.3.3. *Relative Simplicity Measures*

In the previous section I have argued that the important challenge of how to deal with the uncertainty of subset selection can be conceptualized in an alternative way than through the covariance between $\hat{y}_n$ and y_n. The second AST explains that the risk of ignoring a relevant regressor is reduced if parameters of (highly correlated) regressors are grouped together. Of all the shrinkage estimators, Ridge regression (b_{2ASTa}) captures the essence of simplicity in terms of the second AST, because it merely groups parameters together without performing subset selection. It can also be remarked from the previous paragraph that an astimator's relative simplicity term is used to measure the effective number of parameters. Particularly, when regressors are uncorrelated and $\beta_0 = \vec{0}$, $\hat{\mathscr{K}}_{1ir,Ridge}$ equals $\sum_{k=1}^{K} \frac{|b_{Ridge,k}|}{|b_{OLS,k}|}$, which corresponds to the relative simplicity measure of a b_{1i} astimator applied to the solutions of Ridge regression.

Having reached the topic of simplicity terms, is it not a bit strange to start worrying about the uncertainty of subset selection only after an estimator has been selected? The second AST is what mainly sets different estimators and their simplicity terms apart, after all. Indeed, if we rescale the simplicity term of an estimator, a straightforward measure for the effective number of parameters can be defined which will also be monotonic with respect to an estimator's simplicity term.

In Ridge regression, for example, simplicity is measured through $\hat{\mathscr{K}}_{Ridge} = \sum_{k=1}^{K}(b_k - \beta_{0,k})^2$, and this expression can just be premultiplied by $\frac{K}{\hat{\mathscr{K}}_{\lambda=0}}$ so that it goes from 0 to K as λ goes from ∞ to 0. This amounts to computing

$$\hat{\mathscr{K}}_{2a} = \sum_{k=1}^{K} \frac{(b_k - \beta_{0,k})^2}{\frac{1}{K}\sum_l (b_{OLS,l} - \beta_{0,l})^2}$$

and that exactly equals the relative simplicity measure of b_{2a}; the astimated analogue of Ridge regression. What happens when the relative simplicity term of an astimator is used as its measure for the effective number of parameters?

As a small thought experiment, imagine a situation where y is regressed on two variables that are almost perfectly alike (extreme cross-correlation). Take $y = X\beta + \epsilon$ with $\beta = [1\ 1]'$, $\beta_0 = [0\ 0]'$, standard normal X, $\epsilon \sim \mathscr{N}(0, \Sigma)$, a matrix Σ which has elements that are (close to) 1, and a sample of $N = 1000$, say.

In that case, the fit of $b = [1\ 1]'$ with $K_A = 2$ will almost be exactly the same as $b = [2\ 0]'$ with $K_A = 1$. A discrete subset selection procedure therefore favors the risky alternative of using a single regressor. Conversely, $\hat{\mathscr{K}}_{2a}$ equals 2 in the former and 4 in the latter, which means that Ridge regression gives preference to the risk-diversified alternative of including both correlated regressors. A disadvantage of $\hat{\mathscr{K}}_{2a}$ is that it is too lenient for irrelevant regressors (in which case $\frac{1}{K}\sum_l b^2_{OLS,l}$ is comparatively large).

In terms of subset selection, reasons for preferring one measure of the effective number of parameters over another are the same as those for preferring one astimator over another. Relative simplicity in a b_{2c} astimator is given by

$$\hat{\mathscr{K}}_{2c} = \sum_{k=1}^{K} \frac{(b_k - \beta_{0,k})^2}{\sum_l \theta_l^k (b_{OLS,l} - \beta_{0,l})^2}.$$

When $\hat{\mathscr{K}}_{2c}$ with $c_{\min} = 0.5$ is used as a measure of $\mathscr{K}$ in the experiment above, it gives an equally large penalty as Ridge regression to the risky alternative of $b = [2\ 0]'$. Unlike Ridge regression, it is quite strict on the inclusion of irrelevant regressors, because b_{2c} uses a $2i$ type penalty when regressors are not highly correlated. I should note that $\hat{\mathscr{K}}_{2c}$ can easily lead to extreme values when applied to b_{Ridge} for that reason.[6] When $\hat{\mathscr{K}}_{2c}$ is used with b_{2c}, such nonmonotonicities will not occur. In general, $\hat{\mathscr{K}}_{12c}$ values might differ from K when $\lambda = 0$ and $\alpha \in [0, 1]$, which is why I will premultiply these values by $\frac{K}{\mathscr{K}_{\lambda=0}}$. The data should also be standardized when a relative simplicity term is used to measure $\mathscr{K}$.

Of all the relative simplicity measures, $\hat{\mathscr{K}}_{1i}$ of the Adaptive Lasso corresponds most closely to the gradual version of K_A called $\hat{\mathscr{K}}_{1ir}$, although it does compensate for cross-correlations to some extent. $\hat{\mathscr{K}}_{2i}$ differs from $\hat{\mathscr{K}}_{1i}$ in that it promotes parameters of (un)correlated regressors to have a similar degree of shrinkage. In the example with two nearly identical regressors, the Lasso type $\hat{\mathscr{K}}_{1a}$ measure clearly compensates for cross-correlations because it equals 2 for both $b = [2\ 0]'$ and $b = [1\ 1]'$. Since $\hat{\mathscr{K}}_{1a}$ makes no distinction between high and low cross-correlations in stimulating grouping, it can also be too strict for relevant regressors and too lenient for the irrelevant ones. This problem is alleviated when the relative simplicity term $\hat{\mathscr{K}}_{1c}$ is applied on b_{1c}.

So, perhaps if we have found an astimator that enables a researcher to handle the two ASTs, we will also have found a measure for the effective number of parameters that does the same.

[6] Ridge regression can move $b_{Ridge,k}$ outside of $[\beta_{0,k}, b_{OLS,k}]$ to make the *nominal* deviance from $\beta_{0,k}$ more similar to those of others. This can create extremely high $\hat{\mathscr{K}}_{1c}$ when x_k is irrelevant and $b_{OLS,k}$ is close to zero.

4.4. Hypotheses

Table 4.1: *Hypotheses*

	Subset	Correlations	AST alternative
Zellner	0	0	
Bayes/Ridge	0	+	2ASTa
	+	-	2ASTi
	+	+	2ASTc
Lasso	+	0	1ASTa
Adaptive Lasso	+	-	1ASTi
Subset Selection	+	-	
	+	0	1ASTc
Elastic Net	+	+	12ASTa
	+	+	12ASTc

Scores: better (+), equal (0), or worse (-) w.r.t. OLS. Subset: selecting a relevant subset of regressors. Correlations: dealing with highly correlated regressors.

Now that I have discussed how tuning parameters can be selected, Table 4.4 gives an overview of the hypothesized performance of the different estimators and astimators. It shows whether the techniques are expected to improve the forecasting performance of OLS when there is a subset of relevant regressors ('Subset') or when there are high cross-correlations among regressors ('Correlations'). A method's accuracy will be measured in terms of Mean Squared Forecasting Errors,

$$MSFE_t = \frac{1}{V}\sum_{v=1}^{V}(\hat{y}_v - y_v)^2,$$

for $v \in [1, V]$ out-of-sample observations.

Under the natural conjugate prior distribution of Raiffa and Schlaifer (1961) with $p(\beta|\sigma^2) \sim \mathcal{N}(\beta_0, \sigma^2 B_0)$ and $p(\sigma^2) \sim IG(\alpha_0/2, \delta_0/2)$, it is to be expected that the posterior mean does well in the case of correlated regressors. Being equal to Ridge regression for $B_0 = I_k/\lambda$ and standardized data, the Bayesian estimator will minimize the joint squared norm and this stimulates (highly correlated) regressors to receive the same value. Continuing with Zellner's g-prior,

$$b_{Zellner} = \frac{1}{1+g}\beta_0 + \frac{g}{1+g}b_{OLS},$$

it is not evident that it will help to improve upon OLS in the stipulated situations, because it just shrinks OLS solutions to β_0 with no regard for a regressor's relevance, nor for its correlation with other regressors.

Since the Lasso and the Adaptive Lasso are famous for performing exact subset selection, I presume that they outperform OLS in a subset selection exercise. The Lasso can be redefined as an astimator with an average simplicity measure ('1ASTa'), and I have argued in the previous chapter that it can be expected to have the edge on the Adaptive Lasso in case of correlated regressors. Measuring the effective number of parameters with the simplicity term of the Adaptive Lasso comes closest to counting the number of parameters that are included in the model. The Subset Selection technique minimizes BIC with K_A by iteratively deleting the worst regressor from the model (an intercept is always included). It is expected to perform well in a subset selection exercise and poorly when regressors are highly correlated. The Elastic Net combines Ridge regression and the Lasso and is designed to have accurate forecasts in all situations.

Since none of the Bayesian and Frequentist estimators differentiate between high and low cross-correlations, I expect that subset selection and grouping can be performed more effectively. Of the astimators that have been developed, I will be particularly interested in the performances of b_{2c}, b_{1c}, and b_{12c}, because these astimators were designed to give accurate estimates of the underlying process in each of the situations described in Table 4.4. The optimal choice of α could also depend on cross-correlations. In case they are low, an α closer to 0 can be used to quickly get rid of irrelevant regressors with b_{1c}. An α around 1 might group parameters together more effectively when regressors are highly correlated.

With regard to the selection of tuning parameters, and λ in particular, the current best practice is to apply cross-validation. Consequently, the H_0 hypothesis is that this procedure outperforms the use of information criteria. Before evaluating forecasting accuracies of each method through simulation studies, I will first illustrate the differences between various measures of $\mathscr{K}$.

4.5. Simulation Studies

4.5.1. *Analyzing the Influence of $\mathscr{K}$*

To illustrate the behavior of relative simplicity measures that are used to estimate $\mathscr{K}$, I will return to a simulated data set of the previous chapters with two relevant and highly correlated regressors and two irrelevant and uncorrelated regressors.

In this example, $N = 20$ data points are simulated with $y = X\beta + \epsilon$, where $\beta = [2\ 2\ 0\ 0]'$, $\epsilon \sim \mathscr{N}(0,1)$, $X \sim \mathscr{N}(0,\Sigma)$, and $\Sigma = I$ except for $\Sigma_{\{2,1\},\{1,2\}} = .9$. In terms of in-sample accuracy, it makes little difference whether $b_1 = 4$ and $b_2 = 0$ or whether both parameters are equal to 2, because the two regressors are so highly correlated.

The upper panel of Figure 4.1 shows how the b_{1ASTi} solutions change as λ decreases towards 0. As I have demonstrated in the previous chapter, this Adaptive Lasso type astimator focuses on a single regressor out of a group of correlated regressors. In the current data set, b_1 is activated at $\lambda = R^{\otimes}_{1,1} = 0.56$ and moves in the direction of $b_{1OLS} \approx 4$. The almost equally relevant parameter b_2 is only added to the active set at $\lambda = 0.07$, after which the two parameters move closer together. The irrelevant regressors b_3 and b_4 are activated at the very end.

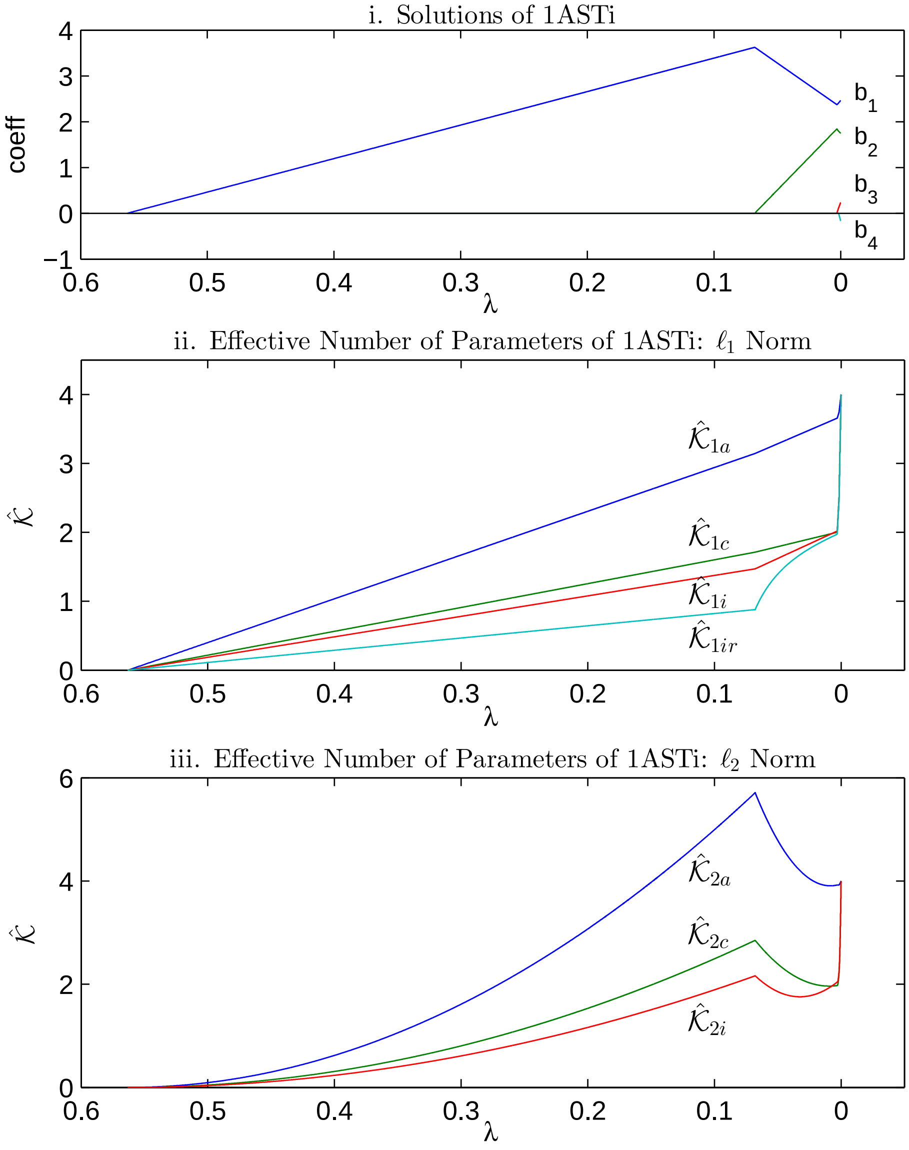
The middle panel of Figure 4.1 presents four measures of the effective number of parameters $\mathscr{K}$ with an ℓ_1 norm. The lowest line represents $\hat{\mathscr{K}}_{1ir}$, which merely counts the degree to which each parameter has been included in terms of the first AST. $\hat{\mathscr{K}}_{1ir}$ moves towards 1 as long as only the first parameter is allowed to deviate from β_0; and heavily penalizes the introduction of the second parameter by quickly increasing in the direction of 2 once it is activated. The Adaptive Lasso type $\hat{\mathscr{K}}_{1i}$ does promote grouping to some degree, since it already exceeds 1 before b_2 is added.

$\hat{\mathscr{K}}_{1c}$ is almost indifferent about whether b_2 is added to the active set or not. Even before the second parameter is added, $\hat{\mathscr{K}}_{1c}$ moves towards 2 in a near linear fashion. No additional penalty is given for the exclusion of the irrelevant regressors, because they are barely correlated with x_1 and x_2. By contrast, $\hat{\mathscr{K}}_{1a}$ does not discriminate between degrees of cross-correlations at all. To penalize the Adaptive Lasso for singling out the first parameter while ignoring the others, $\hat{\mathscr{K}}_{1a}$ already moves towards $K = 4$ in case only the first parameter is activated.

Next, I will turn to the lower panel of Figure 4.1, where estimates of $\mathscr{K}$ with an ℓ_2 norm are depicted. In comparison to $\hat{\mathscr{K}}_{1i}$, $\hat{\mathscr{K}}_{2i}$ promotes grouping more vehemently, because it decreases once b_2 is activated. From the moment that the degree of shrinkage of b_1 and b_2 becomes more similar, $\hat{\mathscr{K}}_{2i}$ quickly lowers in the direction of 2. The grouping effect is stronger for $\hat{\mathscr{K}}_{2c}$, which has already reached 2.8 before b_2 has been added to the active set. The relative simplicity measure of b_{2a} excessively penalizes the lack of grouping of the Adaptive Lasso type astimator, acting as if all regressors in X are perfectly correlated.

One thing to take away from Figure 4.1 is that risky solutions that only focus on a single correlated regressor can be discouraged by applying a measure of

Figure 4.1: *Measures of $\mathscr{K}$ for b_{1ASTi}*


i. Solutions of 1ASTi

ii. Effective Number of Parameters of 1ASTi: ℓ_1 Norm

iii. Effective Number of Parameters of 1ASTi: ℓ_2 Norm

Panel i gives solutions of the Adaptive Lasso type b_{1i} astimator. Panel ii presents the effective number of parameters of b_{1i} when $\hat{\mathscr{K}}$ is computed with the relative simplicity measures of astimators with an ℓ_1 norm. When only b_1 is activated, for example, $\hat{\mathscr{K}}_{1ir}$ in equation (4.3) exactly matches the degree to which b_1 has been included unrestrictedly, while $\hat{\mathscr{K}}_{1a} = \frac{|b_k - \beta_{0,k}|}{\frac{1}{K}\sum_l |b_{OLS,l} - \beta_{0,l}|}$ heavily penalizes the exclusion of the other regressors. Panel iii applies $\hat{\mathscr{K}}$ with an ℓ_2 norm to the solutions of b_{1i}. Regarding the data generating process, $N = 20$ data points are simulated with $y = X\beta + \epsilon$, $\beta = [2\ 2\ 0\ 0]'$, $\epsilon \sim \mathscr{N}(0, 1)$, $X \sim \mathscr{N}(0, \Sigma)$, and $\Sigma = I$ except for $\Sigma_{\{2,1\},\{1,2\}} = .9$. I use $\beta_0 = [0\ 0\ 0\ 0]'$ and $c_{\min} = .5$.

$\mathscr{K}$ that does *not* increase monotonically as λ decreases. A still more sensible strategy is to adjust the astimator along with $\mathscr{K}$, so that they both perform subset selection and grouping effectively.

4.5.2. *Out of Sample Performance*

To asses the out-of-sample performance of estimators and astimators, I will generate data with $y = X\beta + \epsilon$ for $X \sim \mathscr{N}(0, \Sigma)$, $\epsilon \sim \mathscr{N}(0, 1)$, and evaluate sample sizes of $N = 20$, 30, and 60. The intercept is always simulated to be zero. I will standardize the data and retransform the solutions for all techniques. The result is that the parameters of K regressors and an intercept are estimated. Prior coefficients are equated to $\beta_{0,k} = 0$ for all $k = 1, \ldots, K$.

Table 4.2: *Overview Data Generating Processes*

Name	$k = 1, 2, 3, 4$	$k = 5, 6, 7, 8$	Σ (Correlation matrix)
subset	$\beta_k = 2$	$\beta_k = 0$	I_K
corr.	$\beta_k = 2$	-	$c = 0.9$ for $\forall k$
subset & corr.	$\beta_k = 2$	$\beta_k = 0$	$c = 0.9$ for $k \leq 4$, 0 otherwise

This Table gives an overview of three simulation exercises called 'subset', 'corr', and 'subset & corr.' The expression '$c = 0.9$ for $k \leq 4$, 0 otherwise' means that cross-correlations among the first four regressors are 0.9 and that all other cross-correlations are 0.

As Table 4.2 shows, the first simulation exercise is about selecting four out of eight relevant regressors. Cross-correlations in Σ are simulated to be zero. As I have detailed in the previous chapters, a small sample size relative to the number of regressors can cause sample correlations to be quite high even when one simulates that $\Sigma = I_K$. The second task focuses on estimating a model containing four regressors with cross-correlations of 0.9. The third exercise combines subset selection and grouping by including four relevant and highly correlated regressors and four irrelevant and uncorrelated regressors.

Each simulation study is repeated 10,000 times. In each iteration, five thousand out-of-sample predictions are used to compute the MSFEs. I will report the MSFE of the estimators and astimators relative to the MSFE of OLS. Even though b_{1ASTa} is a rescaled version of the Lasso, I will present this and other benchmark methods separately. The primary reason is that the scaling of b_{1ASTa} can affect the quality of cross-validation, because the scaling changes with each fold that is excluded. For the information criteria, solutions of the estimator and its astimated analogue will be exactly the same, so they will only be presented for the latter. In the reference setup of b_{12ASTc} (referred to as '12ASTc') I apply 10-fold cross validation to select the tuning parameters. '($\hat{\mathscr{K}}_{12c}$) 12ASTc' means that λ is selected with BIC using $\mathscr{K}_{12c}$; and that $c_{\min}$

and α are manually specified to be 0.5.

Table 4.3 presents the MSFEs relative to those of b_{OLS} for the three simulation studies. Zellner's g-prior has scores of 1.00 or higher, which means that it is unable to beat b_{OLS} in any of the cases. Turning to methods with an ℓ_2 norm, it can be remarked that b_{2ASTa} (Ridge/Bayes) is good at dealing with highly correlated regressors ('corr.'). When the sample size is twenty, b_{2ASTa} has a score of 0.92, for example, which means that it outperforms OLS by 8%. As expected, b_{2ASTi} has good results in the subset selection exercise with small cross-correlations. The b_{2ASTc} technique performs grouping as well b_{2ASTa}, subset selection as well as b_{2ASTi}, and has the best scores in general when it comes to a situation in which both subset selection and grouping are required.

Regarding the ℓ_1 based techniques, it is remarkable that b_{1ASTa} (Lasso), which is well known for performing subset selection, improves OLS only by a small margin in the first exercise. The Adaptive Lasso type b_{1ASTi} solutions result in quite similar scores as b_{2ASTi} and b_{2ASTc} when selecting four out of eight uncorrelated regressors. The technique starts to get into trouble once cross-correlations increase in the second and third task. The same goes for the Subset Selection method, which gives no compensation for risk-diversified solutions in deciding whether correlated regressors are incorporated or not. In the group of astimators with an ℓ_1 norm, the solutions of b_{1ASTc} with $\hat{\mathscr{K}}_{1c}$ are the most optimal for each task.

Third, I will turn to computationally intensive methods that combine ℓ_1 and ℓ_2 norms. The Elastic Net procedure is ineffective when dealing with the two tasks where subset selection is required, scoring notably worse than the closed-form solutions of b_{2ASTc}. An α lower than 1 in b_{12ASTc} only appears to be preferable in terms of forecasting accuracy when regressors are uncorrelated, in which case b_{1ASTc} slightly outperforms b_{2ASTc}.

On the selection of tuning parameters, it should be remarked that a $c_{\min} \in [0.4, 0.8]$ produces highly similar results in these simulation studies, since cross-correlations are either simulated to be 0 or 0.9. Nevertheless, the cross-validated choice of $c_{\min}$ can vary substantially. In the third exercise with $N = 20$, for example, the selected $c_{\min}$ for b_{2ASTc} is 0 in five percent of the cases and 0.90 in thirty percent of the cases. This explains why forecasts are occasionally improved by setting '$c_{\min} = 0.5$', where cross-validation is applied to select λ while $c_{\min}$ is fixed to be 0.5. To reduce uncertainty in the cross-validated choice of $c_{\min}$, the parameter can be astimated with a prior $c_{\min,0}$ of around 0.5.

Turning to the selection of λ, it is interesting to observe that the more time-consuming method of cross-validation rarely beats BIC. What is more, the

Table 4.3: *Relative MSFEs of Three Simulation Studies*

	subset			corr.			subset & corr.		
N	20	30	60	20	30	60	20	30	60
Zellner	1.02	1.00	1.00	1.01	1.00	1.00	1.01	1.00	1.00
Ridge/Bayes	1.01	1.00	1.00	.92	.95	.98	.87	.93	.98
2ASTa (CV)	1.02	1.01	1.00	.93	.96	.98	.88	.93	.98
$\hat{\mathscr{K}}_{2a}$	1.02	1.02	1.01	.93	.96	.99	.86	.93	.98
DF	.99	1.00	1.00	.92	.95	.99	.86	.93	.98
2ASTi	.89	.92	.96	1.01	1.00	1.00	.97	.97	.97
$\hat{\mathscr{K}}_{2i}$	.86	.90	.95	1.09	1.06	1.02	.92	.94	.96
DF	.87	.91	.95	1.17	1.13	1.07	.94	.98	1.00
2ASTc	.91	.93	.96	.94	.96	.99	.76	.85	.93
c_{min}=0	1.02	1.00	1.00	.93	.96	.98	.83	.90	.96
$c_{min}=\frac{1}{2}$	.91	.93	.96	.93	.96	.98	.74	.83	.92
$\hat{\mathscr{K}}_{2c}$	.88	.91	.95	.93	.96	.99	.78	.85	.93
DF	.89	.91	.95	.92	.95	.99	.74	.83	.92
Lasso	.94	.97	.99	1.00	1.00	1.00	.86	.92	.97
1ASTa	.95	.97	.99	1.00	1.00	1.00	.87	.93	.97
$\hat{\mathscr{K}}_{1a}$	.93	.96	.98	1.00	1.00	1.00	.91	.95	.98
DF	.95	.97	.99	1.00	1.00	1.00	.87	.92	.96
Adap. Lasso	.87	.91	.95	1.02	1.00	1.00	.99	.97	.97
1ASTi	.87	.91	.95	1.02	1.01	1.00	1.00	.98	.97
$\hat{\mathscr{K}}_{1i}$	.85	.89	.94	1.10	1.05	1.01	.92	.93	.95
$\hat{\mathscr{K}}_{1c}$	.95	.97	.99	1.00	1.00	1.00	.96	.98	.99
$\hat{\mathscr{K}}_{1ir}$	.88	.91	.96	1.09	1.03	1.01	.94	.94	.96
Subset Select.	.89	.91	.95	1.14	1.05	1.00	1.01	.96	.95
1ASTc	.87	.91	.95	1.03	1.01	1.00	.84	.90	.95
$\hat{\mathscr{K}}_{1c}$	.85	.89	.94	1.00	1.00	1.00	.84	.89	.94
Elastic Net	.95	.97	.99	.93	.96	.98	.83	.91	.96
12ASTa	.96	.97	.99	.93	.96	.98	.84	.91	.97
$\hat{\mathscr{K}}_{12a}$	.96	.98	.99	.98	.99	1.00	.90	.95	.99
12ASTc	.88	.92	.95	.95	.97	.99	.78	.86	.93
$\hat{\mathscr{K}}_{12c}$	.86	.90	.94	.97	.98	.99	.81	.87	.94

- This table reports the MSFE of estimators and astimators relative to the MSFE of OLS for samples sizes of $N = 20, 30,$ and 60. Simulations are repeated $10,000$ times, with 'subset': 4 out of 8 regressors relevant, no cross-correlations, 'corr.': 4 relevant and highly correlated regressors, 'subset & corr.': 4 relevant and highly correlated regressors and 4 irrelevant and uncorrelated regressors. See Table 4.2 for specifications of data generating processes.
- All of the shrinkage methods use 10-fold cross-validation to select the tuning parameters, unless the indented specification below a bold-faced method states otherwise. The indented '$\hat{\mathscr{K}}_{12c}$' below 12ASTc, for example, means that λ is selected with BIC using $\hat{\mathscr{K}}_{12c}$; while $c_{\text{min}} = \alpha = 0.5$. Lastly, the indented '$c_{min}=\frac{1}{2}$' below 2ASTc means that cross-validation is applied to select λ while $c_{\text{min}} = 0.5$ has been prespecified.

Table 4.4: *Relative MSFEs of Simulation Studies with $N = 10$*

	s.	c.	s.&c.		s.	c.	s.&c.
Ridge/Bayes	.93	.81	.56	**Lasso**	.93	.99	.52
2ASTa (CV)	1.11	.83	.86	**1ASTa**	.97	1.00	.62
γ=.9	.84	.82	.56	γ=.9	.80	1.00	.54
γ=.99	.79	.80	.54	γ=.99	.76	.98	.55
$\hat{\mathscr{K}}_{2a}$	.84	.83	.72	$\hat{\mathscr{K}}_{1a}$	.85	.96	.77
DF	.92	.82	.84	DF	.92	.97	.84
2ASTi	1.47	1.09	.96	**Adaptive Lasso**	1.27	1.10	.77
γ=.9	.78	1.08	.60	**1ASTi**	1.28	1.14	.82
γ=.99	.73	1.05	.59	γ=.9	.83	1.13	.68
$\hat{\mathscr{K}}_{2i}$	.85	1.08	.79	γ=.99	.76	1.10	.66
DF	.89	1.15	.85	$\hat{\mathscr{K}}_{1i}$	.86	1.13	.82
2ASTc	1.10	.88	.65	**Subset Select.**	.98	1.23	.97
γ=.9	.79	.87	.45	**1ASTc**	.97	1.11	.58
γ=.99	.76	.84	.46	γ=.9	.76	1.10	.52
$\hat{\mathscr{K}}_{2c}$	.84	.84	.70	γ=.99	.73	1.06	.53
DF	.90	.82	.80	$\hat{\mathscr{K}}_{1c}$	.86	.96	.76
Elastic Net	.91	.83	.51	**Zellner**	2.26	1.07	5.60
12ASTa	.97	.85	.62				
γ=.9	.80	.84	.52				
γ=.99	.77	.82	.54				
12ASTc	.97	.91	.55				
γ=.9	.76	1.07	.52				
γ=.99	.73	1.03	.53				
$\hat{\mathscr{K}}_{12c}$	.85	.90	.73				

- This table reports the MSFE of estimators and astimators relative to the MSFE of OLS for a samples size of $N = 10$. Simulations are repeated $10,000$ times, with 'subset': 4 out of 8 regressors relevant, no cross-correlations, 'corr.': 4 relevant and highly correlated regressors, 'subset & corr.': 4 relevant and highly correlated regressors and 4 irrelevant and uncorrelated regressors. See Table 4.2 for specifications of data generating processes.
- All of the shrinkage methods use 10-fold cross-validation to select the tuning parameters, unless the indented specification below a bold-faced method states otherwise. The indented '$\hat{\mathscr{K}}_{12c}$' below 12ASTc, for example, means that λ is selected with BIC using $\hat{\mathscr{K}}_{12c}$; while $c_{\min} = \alpha = 0.5$. The indented 'γ=0.9' means that λ is astimated through cross-validation with $\lambda_0 = 0$.

straightforward measure of the effective number of parameters $\mathscr{K}$ in terms of an astimator's relative simplicity term is highly competitive to the more opaque degrees of freedom approach, for which closed-form expressions are only available in a few instances. With respect to the risky solutions of b_{1ASTi} (which tends to focus on a single correlated regressor instead of spreading risks among the entire group of correlated regressors), Table 4.3 can be used to compare $\hat{\mathscr{K}}_{1i}$ to $\hat{\mathscr{K}}_{1ir}$ and $\hat{\mathscr{K}}_{1c}$. The latter leads to more conservative choices among candidate b_{1ASTi} solutions, with λ values closer to 0 (resulting in more grouping). In the second exercise about correlated regressors with $N = 20$, for example, the average λ value is .0003 with $\hat{\mathscr{K}}_{1c}$ and .0011 with $\hat{\mathscr{K}}_{1i}$, resulting in more moderate losses and gains relative to b_{OLS} for $\hat{\mathscr{K}}_{1c}$.

Next, I will illustrate how λ can be *ast*imated with $\lambda_0 = 0$ and $\gamma = 0.9$ or 0.99. For the earlier results in Table 4.3, it does not matter whether $\gamma = 0.99$ or 0. This changes when the sample size is $N = 10$, which causes great variations in the estimated cross-correlations and b_{OLS}. Table 4.4 confirms that plain cross-validation then leads to relatively poor results for astimators and the Adaptive Lasso because the penalty term λ becomes too high. Ridge regression is better than b_{OLS} in the subset selection exercise, for example, while its astimated version '2ASTa' is worse than b_{OLS}. Once deviations from $\lambda_0 = 0$ are penalized with 'γ=0.9', the out-of-sample performance of b_{OLS} is considerably improved once more. The results of 12ASTc show that the choice of α had better be astimated as well. It can also be observed again that BIC with a relative simplicity measure for $\hat{\mathscr{K}}$ has promising results in comparison to the degrees of freedom approach and to plain cross-validation.

In sum, there are three overall conclusions about the simulation studies. First, controlling for correlations through $c_{\min}$ leads to excellent results in terms of forecasting accuracy relative to the benchmark methods. Second, the results of BIC are highly competitive to those of cross-validation. Third, an astimator's relative simplicity term can be used as a straightforward measure for the effective number of parameters in an IC.

4.6. Discussion

At the expense of in-sample accuracy, the flexibility of OLS can be restricted by stimulating parameters to be close to a prior β_0 or close to each other via grouping. Astimators that effectively control both ASTs have been shown to outperform other techniques in terms of out-of-sample performance in circumstances where only a subset of regressors is relevant and where there are highly correlated regressors. The ASTs are also of relevance when selecting tuning parameters. Astimators make it easier to define a set of candidate λ. It was also illustrated that cross-validation can lead to volatile choices of λ and that λ can itself be astimated. The hyperparameter λ_0 could also be based on an Information Criterion. The use of ICs has become somewhat controversial due to the measurement of the effective number of parameters ($\mathscr{K}$). I argued that the problem of subset selection uncertainty should be conceptualized in terms of the second AST. Consequently, it was shown that an astimator's relative simplicity term can be used as its measure of $\mathscr{K}$.

Building on the latter result, an exciting area of future research is to employ an astimator as an information criterion. One would succeed in doing so by directly defining a closed-form heuristic choice of λ. This would help to define λ_0 when astimating the tuning parameter, and it would circumvent the trouble of having to optimize over another IC, in which case a large set of candidate λ values needs to be evaluated. Note that BIC can already be rewritten as

$$BIC \propto \log \Big[\text{Relative Accuracy}\Big] + \frac{\log N}{N}\Big[\text{Relative Simplicity}\Big],$$

which closely resembles the loss function of astimators. The logic behind the Adjusted $\bar{R}^2$ criterion of Theil (1961,1971) could be another good starting point for bringing astimators and information criteria closer, since it is defined in terms of $(1-R^2)$, which equals relative accuracy; $\mathscr{K}$, which equals relative simplicity; and the sample size N.

A second direction that might be explored is to improve best subset selection techniques. With the help of a correlation-adjusted $R^{\otimes}$ matrix, one can try to focus on a subset of possibly relevant regressors in each step, so that the search can be performed more quickly. What is more, a $\hat{\mathscr{K}}_{1c(r)}$ type measure of the effective number of parameters might be employed to stimulate highly correlated and relevant regressors to be included as a group.

4.A. Appendix: Degrees of Freedom Ridge Regression

4.A.1: *Show that $\hat{\mathscr{K}}_{DF,Ridge} = \hat{\mathscr{K}}_{1i(r),Ridge}$ when regressors are orthostandard.*

Use that $X'X = (N-1)I_k$ to rewrite the Ridge degrees of freedom as

$$\begin{aligned}\hat{\mathscr{K}}_{\text{DF, Ridge}} &= \text{tr } X(X'X + \lambda I_K)^{-1}X',\\ &= \text{tr } X'X\frac{1}{\lambda + N - 1}I_K,\\ &= K\frac{N-1}{\lambda + N - 1},\end{aligned}$$

where $tr(AB) = tr(BA)$ if A is $m \times n$ and B is $n \times m$.

Next, use that $x_k'x_j = 0$ for all $j \neq k$ and that $x_k'x_k = (N-1)$. To select the k^{th} element of b_{Ridge}, include a $1 \times K$ vector i_k which is 1 at k and 0 otherwise. Finally, note that orthogonality implies that $\sum_{k=1}^{K} \frac{|b_{Ridge,k}|}{|b_{OLSr,k}|} = \sum_{k=1}^{K} \frac{b_{Ridge,k}}{b_{OLS,k}}$. The result is that

$$\begin{aligned}\hat{\mathscr{K}}_{1i(r),\text{Ridge}} &= \sum_{k=1}^{K} \frac{i_k(X'X + \lambda I_K)^{-1}(X'y)}{(x_k'x_k)^{-1}x_k'y},\\ &= K\frac{N-1}{\lambda + N - 1}\sum_{k=1}^{K}\frac{i_k(X'y)}{(x_k'y)},\\ &= K\frac{N-1}{\lambda + N - 1},\end{aligned}$$

which equals $\hat{\mathscr{K}}_{\text{DF, Ridge}}$.

5

Accuracy-Simplicity Tradeoffs and the Weighing of Observations

5.1. Introduction

In time series forecasting, the underlying data generating process might be subject to breaks. The traditional strategy in such a situation is to test for structural breaks and use post-break data (Chow, 1960, Quandt, 1960, Brown et al., 1975, Andrews, 1993). Pesaran and Timmermann (2007) explained, however, that even if one has correctly pinpointed the timing of a break, it might be more optimal in terms of mean squared forecasting errors to include pre-break data. The reason is that a short post-break window can cause a large estimation uncertainty (*ibid.*).

As an alternative, Pesaran and Timmermann (2007) proposed to estimate the timing of the starting point via cross-validation. The idea is to divide the estimation sample in a validation set of recent observations and a training set of more distant observations, and to use the training set to 'predict' the validation set while varying the starting point of the data. In their Best Starting Point method ('SPB'), the starting point with the best pseudo predictions is subsequently selected. One of the main advantages of SPB is that it is easy to apply to a variety of estimation methods. SPB does have three serious drawbacks, and I hope to address them in this chapter.

One problem is that SPB can be slow to respond to a new break, because it can take a while before there is a sufficient number of post-break observations in the validation set. Based on such difficulties in estimating the timing of break points, Pesaran, Pick, and Pranovich (2013, pp. 149) concluded that one should instead use their 'robust optimal weights,' which progressively increase as observations become more recent. Their approximations of exponential weights can indeed be employed when there has been a break most recently or when there are breaks continuously, but in situations where the timing of break points has become sufficiently clear, less recent data need not be wasted so immoderately. After all, data of the distant past could be of relevance in the near future; and an estimation of the timing of discrete break points could be important in trying to understand why the underlying process has changed. ‘Discrete’ weights like SPB result from assigning weights to discrete periods of observations. My proposal is to look for ways in which exponential and discrete weights can be combined.

A second concern is that SPB will ignore large portions of the data based on the smallest of improvements in the accuracy of the validation sample. Such volatility in the selection of a configuration (the best starting point) is a general issue of cross-validation. In the previous chapters I have argued how, at the expense of in-sample Accuracy, Simplicity can be accomplished by penalizing deviations from a given setup. In the current context, I will show that a tuning parameter $\lambda \in [0, 1]$ can be used to intuitively balance between the Accuracy of the validation window and the Simplicity of equal weights. In connection to chapters 2 and 3, where Accuracy-Simplicity Tradeoffs were analyzed in estimating parameters of the linear regression model, I will compare an ℓ_2 norm to an ℓ_1 norm in counting deviations from equal weights. Parallel to Chapter 4, a measure for the effective number of observations will be developed that can be employed in information criteria.

A third issue is that SPB only considers giving positive weights to data after the starting point, whereas individual weights could be assigned to all periods. One advantage of giving each period its proper due is that it will no longer be necessary to adjust the timing of a break point to include pre-break data. Pesaran, Pick, and Pranovich (2013) already derived closed-form expressions for weighing individual periods of observations in the context of a linear regression model with one-step-ahead forecasts. At the cost of analytic tractability, I will present a procedure that is more generally applicable. In doing so, I will also make use of a technique that Bai and Perron (1998, 2003) developed for consistently estimating the timing of multiple breaks. The only downside of this algorithm, is that it can be rather slow for large data sets. I will therefore compare Bai and Perron’s method to a few simple alternatives as an aside.

In short, the goal of this chapter is to improve SPB by responding to a new break more quickly, by discarding old data less quickly, and by assigning individual weights to multiple periods. In dealing with these three aspects, I will develop an algorithm that is designed to provide more robust estimates of the underlying break points and regression parameters. The outline is as follows. In Section 5.2, the benchmark methods are discussed along with the newly proposed algorithm. Simulation studies are presented in Section 5.3. In Section 5.4, I apply the methods on a case study about the rational expectations hypothesis. Section 5.5 concludes.

5.2. Methods

The methods for weighing observations that will be developed below can be applied on many estimation procedures. Still, one can think of applications of the linear regression model,

$$y = X\beta + \epsilon,$$

with a $T \times 1$ dependent variable y, a $T \times K$ matrix of independent variables X, a $K \times 1$ vector of parameters β, and a $T \times 1$ vector of errors ϵ.

The weighing of observations is performed by premultiplying the data by a column vector of weights w that sums to $\sum_{t=1}^{T} w_t = 1$. Assigning an equal weight to all observations in the estimation sample amounts to setting $w_t^{EQ} = {}^1\!/_T$. If a starting point of 2 is selected, only the first observation is left out and the remaining observations receive an equal weight, so that $w = [0 \quad {}^1\!/_{T-1} \quad \dots \quad {}^1\!/_{T-1}]'$.

In case a weighing procedure is applied in the context of the linear regression model, then one can define the matrix $W^{1/2} = \text{diag}(\sqrt{w})$, so that the weights are assigned to the data through $y^w = W^{1/2}y$ and $X^w = W^{1/2}X$. The weighted least squares estimator then becomes $b_{WLS} = (X'WX)^{-1}X'Wy$. Consequently, if y is regressed on a constant $(x_t = 1)$ with weights w_t, the solution is $b_{WLS} = \sum_{t=1}^{T} w_t y_t$, which corresponds to the weighted average of y.

I will now continue by presenting three benchmark methods for weighing observations.

5.2.1. *Benchmark Methods*

The main benchmark to be considered is the best starting point method. This will be called ‘SPB original’, because adjustments will be proposed below. To select the optimal starting point through cross-validation, the estimation sample that runs from 1 to T is split up into two samples. Observations 1 to $T - V$ constitute the training sample, and the last V observations the validation sample. The training sample is used by SPB original to compute h-period-ahead pseudo forecasts regarding the validation sample, whereby the starting point of the training sample is varied. The selected starting point is the one with the best predictions of the validation set. To forecast $\hat{y}_{T+h}$, a sample running from the chosen starting point to T can be employed to estimate the β coefficients.

Regarding the measurement of a starting point's pseudo-out-of-sample prediction accuracy, I will denote an h-step-ahead forecast made with a set of weights w^i as $\hat{y}_{i,v,h}$, where $v \in [T-V+1, T]$ refers to an observation in the validation set. The associated prediction error is then given by $e_{i,v,h} = \hat{y}_{i,v,h} - y_v$, so that the Mean Squared Forecasting Errors can be defined as

$$MSFE^i_t = \frac{1}{V} \sum_{v=T-V+1}^{T} e^2_{i,v,h}.$$

The second benchmark method is also based on cross-validation but takes a *weighted* average of the starting points' predictions. It is called 'SPW original'. The inverse MSFE weight that is assigned to the forecasts of starting point i is given by $\frac{(MSFE^i_t)^{-1}}{\sum_{j=1}^{J}(MSFE^j_t)^{-1}}$, provided that there are J eligible starting points. The technique is based on the forecasting combination literature and is an attempt to diversify risks among starting points. Pesaran and Timmermann (2007) remark that it is likely to work well when there are small breaks.

When applying the SP methods, researchers have to decide upon the minimum length of the training set (minT) and the size of the validation window V. A large minT prevents recent starting points from being selected and a small minT can result in poor estimates of model parameters. The bigger V, the longer it takes for postbreak observations to dominate the validation sample, so the longer it takes for a new break to be identified. When V is too small, on the other hand, 'the ranking of forecasting methods will be too noisy and affected too greatly by random variations' (Pesaran and Timmermann, 2007, pp. 145).

To examine the influence of such tuning parameters, I will define a reference setup and evaluate the effect of altering that reference setup. In the reference setups of the SP methods, the minimum number of observations in the training sample is equated to $minT = 15$, and the validation sample is also specified to have a size of $V = 15$. As I have just explained, the optimal choice of V and minT may depend on the application. To evaluate these manually defined settings, I will study what happens when these sample sizes are allowed to be higher ($minT = 20$, $V = 20$) or lower ($minT = 10$, $V = 10$).

The third benchmark is the 'robust optimal weight' of Pesaran, Pick, and Pranovich (2013, 'PPP'). Under the assumption that the break date is uniformly distributed from 1 to T, the break date is integrated over the entire estimation sample to give,

$$w_t^* = \begin{cases} \frac{-\log(1-t/T)}{T-1} & \text{if } 1 \leq t \leq T-1 \\ \frac{\log(T)}{T-1} & \text{if } t = T, \end{cases}$$

and these weights are normalized to add up to one,

$$w_t^{\text{EXP}} = \frac{w_t^*}{\sum_{t-1}^{T} w_t^*}. \tag{5.1}$$

As it stands, EXP does not require tuning parameters to be cross-validated.

It should be noted that PPP have introduced a number of variants of their robust optimal weights. One such variant is designed to deal with two breaks in a linear regression model conditional on the size of the breaks being known.[1] 'In practice,' write PPP in their conclusion, 'dates and sizes of breaks are unknown and their estimates can be unreliable' (*ibid.*, pp. 149). PPP therefore recommend the robust optimal weights w_t^{EXP} of equation (5.1), since it does not require *a priori* knowledge of break dates or their sizes (*ibid.*).

In evaluating the last two benchmark methods, note that the weighing of starting points by SPW original causes more recent observations to be included more often, which means that they will have a greater influence on the estimation of model parameters than more distant observations. If one wishes to guard against SPB's risk of wrongfully downweighing certain observations, a more direct approach is to penalize deviations from equal weights and to allow for pre-break data to receive a nonnegative weight. The idea of using inverse MSFE scores can then be used to assign weights to periods of observations before and after a break.

Next, EXP assigns progressively higher weights to more recent observations and is related to SPW original in this sense. PPP have shown that the connection between EXP and exponential smoothing is particularly close and that EXP responds more quickly to a single break than SPB original. One might worry about how robust these weights are to changes in the underlying break process. In particular, EXP may not be optimal in case an earlier period of observations is at least as relevant for the current forecast as a more recent period of observations.

With EXP and SPW original in mind, I will focus on three aspects that

[1] An analytic solution of robust optimal weights for two breaks is not available. Note further that PPP also present a version whereby the break sizes are numerically integrated out. The difference with EXP is that the rate of decay is smaller and that the weights initially decrease slightly at the start of the sample. At the end of this section I will present a variant of EXP whereby the rate of decay is determined dynamically by shrinking EXP towards EQ.

might improve the main procedure of interest, SPB original. First, discrete weights can be combined with exponential weights to quicken the response time to new information. Second, SPB original can be made less susceptible to random variations by intuitively penalizing deviations from equal weights through an Accuracy-Simplicity Tradeoff. These features help to circumvent the issue of requiring a small V to make recent starting points eligible and a large V to reduce noise in estimating the starting point. Third, one can enable multiple periods of observations to receive positive weights with the help of inverse MSFE scores. By assigning positive weights to pre-break data, we can deal with the uncertainty of a short post-break window without having to tamper with estimates of the true timing of break points.

The overall goal is to improve upon these three aspects of SPB original. The resulting MB-S algorithm (Multiple Breaks and Shrinkage) incorporates the proposed alterations. It is developed to respond swiftly to a new data generating process and still provide estimates of model parameters and break points that are robust against changes in circumstances. A number of nested procedures will be defined along the way, so that the postulated necessity of each suggestion can be evaluated. Hypotheses about the forecasting performance of the presented methods are formulated at the end of this section.

5.2.2. *Improving the Response Time*

Instead of shortening V and minT to swiftly adapt after a break, one can achieve the same effect by using exponential weights in two unremarkable ways.

First, the predictions errors ('PE') in the validation sample that SPB original uses to select a starting point can be premultiplied by expontial weights to respond more quickly to a recently poor performance. Accordingly, the accuracy measure is defined as

$$MSFE_T^i(w_{\text{PE}}^{\text{EXP}}) = \frac{1}{V} \sum_{v=T-V+1}^{T} w_v^{\text{EXP}} \cdot e_{i,v,h}^2.$$

Weighing the prediction errors in this way is equivalent to weighing the observed y_v and predicted $\hat{y}_{i,v,h}$ with $\sqrt{w_v}$ in weighted least squares. Note further that $w_{\text{PE}}^{\text{EXP}}$ is computed for $t = 1$ to T and that $w_{\text{PE}}^{\text{EXP}}$ uses the last V of those weights after normalizing them to sum to one.

Second, it can be remarked that the original SP methods assign *equal* weights to the observations of the validation sample when the selected starting point is used to estimate a method with which to forecast y_{T+h}. Another simple adjustment to quicken the response to new information is to ascribe exponential weights to the validation window with w_V^{EXP}. When the last V observations receive the last V weights of w_V^{EXP}, then the weights of the training sample do have to be rescaled so that the weights of the entire estimation sample sum to unity.

As described above, a greater emphasis on recent observations could result in noisy approximations of parameters and break points, so I will now look for a way to curtail the influence of cross-validation in weighing observations.

5.2.3. *Penalizing Deviations From Equal Weights*

In SPB original, the tiniest of differences in MSFE can result in much data being excluded, because the MSFE criterion does not penalize deviations from equal weights. One might translate this remark into an Accuracy-Simplicity Tradeoff ('AST'). A method with a good MSFE is defined as being more accurate, and a method that barely deviates from a prior setup of equal weights is defined as more simple. Simplicity can be achieved at the cost of MSFE accuracy. The question is whether it is possible for a researcher to obtain an intuitive control over this AST, so that he can specify how influential the cross-validated weights may be relative to equal weights (EQ).

In accordance with the previous chapters, one can begin by observing that a relative accuracy term can be obtained by dividing the MSFE of w^i by that of EQ. To measure relative simplicity, a deviance measure $D(w^i)$ can be used to quantify how much w^i differs from EQ. For the SP methods, deviances are defined in terms of the proportion of observations in the training sample

$$D_{01}(w^i) = \frac{M - N(w^i)}{M}, \tag{5.2}$$

where $N(w^i)$ is the sample size associated with weights w^i and $M = T - V$ is the maximum number of observations in the training sample. '01' in D_{01} signifies that individual observations are either ignored (0) or included (1). D_{01} equals 0.75 when three quarters of the sample are included, for example. The maximum deviance from equal weights occurs when the minimum amount of observations (minT) is used. $D_{01}(w^i)$ can be divided by $D_{01}(w^{\text{minT}}) = \frac{M - minT}{M}$ to make the simplicity measure relative to the largest permissible deviance from equal weights.

In terms of a general D (other deviance measures will be defined below), the resulting AST loss function becomes

$$L_{AST}(w^i, w_{\text{PE}}^{\text{EXP}}, \lambda, D) = (1-\lambda)\underbrace{\frac{MSFE^i(w_{\text{PE}}^{\text{EXP}})}{MSFE^{\text{EQ}}(w_{\text{PE}}^{\text{EXP}})}}_{\text{Relative accuracy}} + \lambda \underbrace{\frac{D(w^i)}{D(w^{\text{minT}})}}_{\text{Relative simplicity}} . \quad (5.3)$$

The researcher can determine through $\lambda \in [0,1]$ how much influence EQ has relative to cross-validation in weighing observations. EQ is used when $\lambda = 1$, and the cross-validated weights are used when $\lambda = 0$. A λ of $1/3$ means that weights are based for 67% on the accuracy of pseudo forecasts and for 33% on the simplicity of equal weights. The data-optimized solutions might also favor equal weights, of course.

By monotonically transforming equation (5.3) in the following way,

$$L_{AST}(w^i, w_{\text{PE}}^{\text{EXP}}, \lambda, D) \propto MSFE^i(w_{\text{PE}}^{\text{EXP}}) + \underbrace{\frac{\lambda}{1-\lambda} MSFE^{\text{EQ}}(w_{\text{PE}}^{\text{EXP}}) \frac{D(w^i)}{D(w^{\text{minT}})}}_{\text{Penalty term}} .$$

it becomes clear that the $MSFE^i$ of a given set of weights w^i is penalized with the second term on the right hand side. Estimation procedures that can adjust many configurations in order to optimize over the validation sample could be penalized harder to reduce the problem of overfitting the validation sample. Another reason for increasing λ could be that the validation window is rather small and therefore less reliable.

Observe that the penalty is proportional to the $MSFE^{\text{EQ}}(w_{\text{PE}}^{\text{EXP}})$ score of equal weights. When the minimum amount of observations is selected ($w^i = w^{\text{minT}}$), the $MSFE^i$ score has to be at least $\frac{\lambda}{1-\lambda}$ times as good as EQ to be preferred to EQ. This penalty term in conjunction with exponentially weighing forecasting errors reinforces the application of equal weights in case the weighted forecasts suddenly perform poorly relative to equal weights.

The methods that exponentially weigh prediction errors and the validation window and that penalize deviations from EQ are called 'SPB' and 'SPW'. In their reference setups, the AST tuning parameter is set to $\lambda = 1/3$, so that cross-validation has twice as much influence in weighing observations as the prior setup of equal weights. The specification $\lambda = 1/3$ will be compared to $\lambda = 0$ (no penalty for deviating from equal weights), $\lambda = 2/3$, and $\lambda = 1$ (equal weights).

5.2.4. *Multiple Break Points*

To further explore the possibility that observations in the distant past could be more informative about the current underlying process than more recent ones, I will develop a procedure that incorporates multiple break points ('MB') instead of single starting points. To this end, a method needs to be introduced for estimating the timing of breaks and for assigning weights to the resulting periods. The deviance measure D_{01} in L_{AST} will also have to be adjusted, because individual weights will then be allowed to vary among observations that are included.

The method of Bai and Perron ('BP') can be used to find break points. In this procedure, break dates are selected by minimizing the in-sample sum of squared residuals. BP discuss various techniques that choose the *number* of breaks by penalizing the added parameters caused by including break point(s). The Bayesian Information Criterion ('BIC') is known to select too many breaks when there is serial correlation in the errors (Bai and Perron, 2003, pp. 15). By contrast, the modified Schwarz criterion ('LWZ') developed by Liu et al. (1997) tends to be too restrictive in the number of break points it selects. I will use BIC because deviations from equal weights will also be penalized with L_{AST} when weights are assigned to the resulting periods (as I will explain below).

The BP method is known to be rather slow, particularly when the sample size and the maximum number of break points are large. I will therefore set the maximum number of break points allowed to be four, but the optimal choice may depend on the application. A simple alternative that I will consider is that of equally distributing break points over the training sample ('EB'), whereby a researcher can specify the minimum number of observations per period (minT). To compare this method with BP, the maximum number of break points will also be set to four when EB is used. I will also select the single best break point ('BPB') through cross-validation. The main difference between BPB and SPB is that the former allows for weights to be assigned to pre-break data.

To weigh multiple periods, all possible combinations of including some periods while excluding others are evaluated through cross-validation. The best option is subsequently selected. The included periods are either assigned equal weights or individual L_{AST} based weights. In the latter case, L_{AST} scores are computed for each individual period with equation (5.3). Each observation then receives the L_{AST} score of the period it is part of and these weights are subsequently normalized to sum to one ($w_t = \frac{L_{AST,t}}{\sum_{t=1}^{T-V} L_{AST,t}}$). Included periods that have fewer than minT observations receive the average L_{AST} based score of all the included periods. Remember that minT also controls the minimum size of the training sample. As I mentioned before, minT is specified to be 15 in the reference setup and it will be compared to $minT = 10$ and $minT = 20$.

Table 5.1: *Example of Assigning Discrete Weights*

Per.	Obs.	1	2	3	4	5	6	7	8
P1	1:20	EQ	0	EQ	EQ	EQ	0	AST	AST
P2	21:80	EQ	EQ	0	EQ	0	EQ	AST	AST
P3	81:85	EQ	EQ	EQ	0	0	0	AVE	0

Note: this table gives an example of the possible weights that can be assigned to periods P1, P2, and P3 when there are breaks at $t = 21$ and 81 for a training sample of size 85. EQ stands for equal weights, AST for L_{AST} based weights, and AVE is the average of the AST weights of the other periods. Among the 8 alternatives, the one with the best L_{AST} score is selected.

Table 5.1 gives an example for when there are candidate break dates at $t = 21$ and 81 in a training sample of size 85. The training sample is divided into three periods; P1 (observations 1 to 20), P2 (21 to 80), and P3 (81 to 85). The above strategy leads to a total of eight unique ways in which periods of observations can receive weights. In the column labeled 4, the observations of the third period are ignored while the others receive an equal weight of 1/80. From columns 5 and 6 the individual L_{AST} scores of the first two periods can be obtained, since the validation sample is 'predicted' with either the first or the second period, respectively. An individual score is not determined for the third period. The reason is that it only contains five observations, which is less than the required $minT = 15$. When the third period is combined with the other periods on the bases of L_{AST} in column 7, it therefore receives the average L_{AST} score of the first two periods. Of the eight options, the one with the lowest L_{AST} score is selected.

Now that observations can receive individual weights, the deviance measure $D_{01}(w) = \frac{M-N(w)}{M}$ must be refined in the L_{AST} loss function. After all, $D_{01}(w)$ just measures the fraction of observations included without differentiating between the individual weights of the included observations. A deviance measure with an ℓ_2 norm can be formulated as follows

$$D_s(w) = \frac{1}{\sum_{m=1}^{M} w_m^2} \sum_m (w_m - \frac{1}{M})^2, \tag{5.4}$$

given that $\sum_{m=1}^{M} w_m = 1$. Small derivations are provided in Appendix 5.A. The sum of squared deviations is premultiplied by $\frac{1}{\sum_{m=1}^{M} w_m^2}$ to ensure that $D_s(w)$ can be seen as the fraction of observations included. When discrete weights of SPB original are used, for example, $D_s(w) = \frac{M-N(w)}{M}$ and so it also holds that $0 \leq D_s(w) \leq 1$.

To assess the quality of this proposal, D_s will be compared to a measure with absolute deviations from equal weights, which is defined as

$$D_a(w) = \frac{1}{2} \sum_{m=1}^{M} |w_m - \frac{1}{M}|. \tag{5.5}$$

Premultiplication by a half ensures that $D_a(w)$ can also be interpreted as the fraction of included observations ($D_a(w) = \frac{M-N(w)}{M}$ for SPB original). This also implies that $0 \leq D_a(w) \leq 1$.

My preference goes out to $D_s(w)$ for measuring the total deviation from equal weights, because larger individual deviations from EQ are more heavily penalized by the ℓ_2 norm. This can be seen as an attempt to control a second AST between the in-sample Accuracy of the validation sample and the Simplicity of assigning more similar *individual* weights to multiple observations. Conversely, D_a is indifferent to the individual composition of weights, so that it is easier for a few observation to receive extremely high weights. Failing to diversify risks among observations could make $D_a(w)$ more vulnerable to changes in the underlying break process. A small illustration about the differences between D_s and D_a is provided in Appendix 5.A.

In the same Appendix, a useful corollary is also derived, which states that the deviance measures can be employed to obtain heuristics for the 'effective' sample size. This comes in handy for researchers who employ a technique that requires them to specify a sample size when assigning individual weights to observations, like the often applied BIC information criterion. In case the researcher uses exponential weights with $T = M = 100$ observations, for example, a heuristic for the effective number of observations is $\frac{1}{\sum (w^{\text{EXP}})^2} = 51$ under quadratic deviances and $(1 - D_a(w^{\text{EXP}}))M = 63$ under absolute deviances.

5.2.5. *Shrinking MB to EXP and EQ*

So far, I have developed an algorithm that starts by finding candidate break points in the first step. In the second step, it assigns discrete weights to the resulting periods of the training sample. During this second step, the prediction errors and the validation sample receive exponential weights; and deviations from equal weights are penalized with λ_{AST}. This method was called MB.

In case there are breaks continuously or when the timing and size of the break points are unclear, MB may not be preferable. A better option could be to shrink the entire estimation sample towards exponential weights. In a situation with much variability but no breaks, the flexibility of MB to exclude some periods while including others might actually worsen forecasts, in which case equal weights could be more appropriate. As a last step, I will therefore try to refine the first aspect of combining discrete and exponential weights by making such a choice adaptive to changes across times and across applications.

In case a sufficient amount of predictions has been gathered of MB, EXP, and EQ, then one can choose among them based on an L_{AST} measure. As an example, if one has been predicting since $T = 30$ and the current time is $T = 60$, then the last $V = 15$ predictions of each technique can be used to select the best one. To allow for a more gradual transition from MB towards either EQ or EXP, a shrinkage model is applied

$$w_T^{\text{MB-S}}(\mathbb{1}_{\text{EQ}}, \phi) = \begin{cases} (1-\phi)\hat{w}_T^{\text{MB}} + \phi w_T^{\text{EQ}} & \text{if } \mathbb{1}_{\text{EQ}} = 1, \\ (1-\phi)\hat{w}_T^{\text{MB}} + \phi w_T^{\text{EXP}} & \text{if } \mathbb{1}_{\text{EQ}} = 0, \end{cases} \tag{5.6}$$

where ϕ is the shrinkage rate and $\mathbb{1}_{\text{EQ}}$ indicates whether discrete weights are shrunk towards equal (1) or exponential weights (0). The specifications of ϕ and $\mathbb{1}_{\text{EQ}}$ are obtained by varying $\phi \in \{0 : 0.1 : 1\}$ and $\mathbb{1}_{\text{EQ}} \in \{0, 1\}$ and selecting the settings that minimize the L_{AST} loss function. The method that results from shrinking MB is called 'MB-S'. If not enough forecasts of MB are yet available in real-time, a shrinkage of $\phi = 0.5$ towards equal weights is used. Next to MB-S, I will also evaluate 'SPB-S', whereby SPB is shrunk towards EXP or EQ through equation (5.6).

As a final benchmark, I will examine 'EXP-S', which shrinks w_T^{EXP} to w_T^{EQ} based on the L_{AST} loss function like so,

$$w_t^{EXP}(\phi) = (1-\phi)w_t^{EXP} + \phi w_t^{EQ}.$$

The advantage of this shrinkage model over exponential smoothing is that there is no dependency on the first observation in determining the rate of decay, since PPP's weights are not defined recursively.

5.2.6. *Hypotheses*

To summarize, the three main steps in the MB-S algorithm are: find break dates (1), obtain weights for the training sample (2), shrink weights to equal or exponential weights (3). Exponential weights are also assigned to pseudo forecast errors and the observations of the validation window; and deviations from equal weights are only allowed to the extent that pseudo forecasting accuracy sufficiently increases. An overview of all the methods and configurations is presented in Table 5.2.

Table 5.2: *Overview Methods and Tuning Parameters*

1. Benchmark methods	
SPB original	Select best starting point (with minT, V)
SPW original	Take weighted average of starting points (with minT, V)
EXP	PPP's robust optimal weights
2. Exponential weights and penalizations	
SPB	Select best starting point with $w_{\mathrm{V}}^{\mathrm{EXP}}$ and L_{AST} (and minT, V)
SPW	Take weighted average of starting points with $w_{\mathrm{V}}^{\mathrm{EXP}}$ and L_{AST}
BPB	Select best break point with $w_{\mathrm{V}}^{\mathrm{EXP}}$ and L_{AST} through cross-validation
MB	Select multiple break points with $w_{\mathrm{V}}^{\mathrm{EXP}}$ and L_{AST}, BP-BIC, and a maximum number of break points of 4
3. Shrink weights to EXP or EQ	
SPB-S	Shrink SPB to EXP or EQ with $w_{\mathrm{V}}^{\mathrm{EXP}}$ and λ_{AST}
MB-S	Shrink MB to EXP or EQ with $w_{\mathrm{V}}^{\mathrm{EXP}}$ and λ_{AST}
EXP-S	Shrink w_T^{EXP} to w_T^{EQ} with L_{AST}
Tuning parameters (choice in reference setup gets *) and other specifications	
$minT \in \{10, 15^*, 20\}$	Minimum size of the training sample
$V \in \{10, 15^*, 20\}$	Size of the validation sample
$w_{\mathrm{V}}^{\mathrm{EXP}}$	Weigh validation sample exponentially
$w_{\mathrm{PE}}^{\mathrm{EXP}}$	Weigh prediction errors exponentially
$\lambda_{AST} \in \{0, 1/3^*, 2/3, 1\}$	Specify the AST tradeoff to penalize deviations from EQ
D_s or D_a	Use squared or absolute deviations from equal weights
$L_{AST}(w^i, w_{\mathrm{PE}}^{\mathrm{EXP}}, \lambda_{AST}, D_s)$	AST loss function where weights w^i are evaluated

Note: '*BP-BIC*': Bai and Perron procedure with a BIC criterion for selecting the number of breaks. An alternative to BIC is LWZ and an alternative to estimating the timing of breaks through BP is by equally distributing breaks (EB).

Table 5.3 summarizes my hypotheses regarding the benchmark methods.

Table 5.3: *Hypotheses*

H_0	**Description**
1	**SP original methods are slow to respond to a new DGP.**
1.1	Using $w_{\text{PE}}^{\text{EXP}}$ and w_{V}^{EXP} improves the response time.
2	**SP original methods ignore old data too quickly.**
2.1	Setting $\lambda_{AST} = 1/3$ helps to achieve more conservative deviations from EQ.
2.2	When there are multiple breaks in the DGP, the best MSFE follows from estimating multiple discrete breaks.
3	**EXP is robust and optimal.**
3.1	EXP perform best when there is a single break or when there are breaks continuously.

MSFE scores will be used to evaluate these hypotheses. The first main hypothesis is that the SP methods are slow to respond to a new data generating process (‘DGP’). It will be investigated whether exponentially weighing the prediction errors ($w_{\text{PE}}^{\text{EXP}}$) and/or the validation sample (w_{V}^{EXP}) will improve the response time. The second main hypothesis is that SP methods ignore old information too quickly. Here, I will study whether it helps to penalize deviations from equal weights with $\lambda = 1/3$, and whether multiple breaks should be estimated instead of a single starting point. I presume that estimating the timing of multiple breaks works best when the data generating process contains more than one break, but I will also study what happens when a single best break point is used or when the estimation method equally distributes breaks across the training sample.

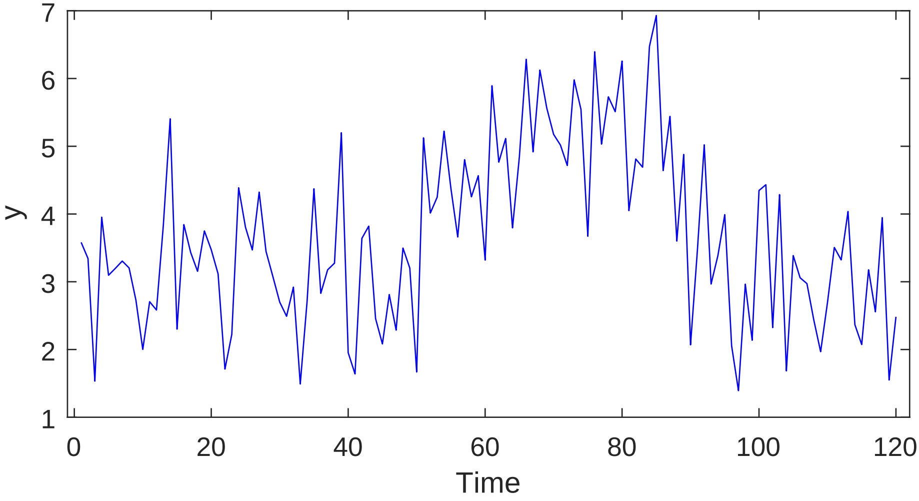
Following PPP, the third main hypothesis is that EXP results in robust optimal weights. I expect that EXP performs best when there is a single break or when there are breaks continuously (due to its connection with exponential smoothing). As an alternative to EXP and the original SP methods, it will be studied whether EQ, EXP, and MB ought to be combined into an MB-S algorithm in order to obtain robust estimates of regression parameters and of (discrete or continuous) breaks.

5.3. Simulation Studies

Before analyzing MSFE accuracies for various simulation problems, the hypotheses are first examined with a single simulation study.

5.3.1. *Two Breaks in Drift*

Figure 5.1: *Example of Simulated Data with Two Breaks in Drift*



Note: This figure presents an example of simulated outcomes when data is generated with $y_t = \mu_t + \epsilon_t$, where $\mu_t = 3 + 2 \cdot \mathbb{1}_{50 \leq t \leq 89}$ and $\epsilon_t \sim \mathcal{N}(0, 1)$.

In the first simulation study, the objective is to estimate a mean ('drift') that is simulated to change from 3 to 5 at $t = 50$ and to revert back to 3 at $t = 90$. That is, the regression model $y_t = \mu_t + \epsilon_t$ is simulated with $\mu_t = 3 + 2 \cdot \mathbb{1}_{50 \leq t \leq 89}$ and a standard normal ϵ_t. The two break dates can be used to define three periods $(1, \ldots, 49; 50, \ldots, 89;$ and $90, \ldots, 120)$, which will be referred to as P1, P2, and P3. The break dates were chosen in this way so that one can observe when methods perform differently compared to equal weights (w_t^{EQ} ='EQ'$= 1/T$) even though the total sample size is small (120 observations). At the cost of a longer estimation time, larger sample sizes could be defined of course.

Forecasts regarding y_{T+1} are generated by taking a weighted average of observations at $t = 1, 2, \ldots, T$, so that $\hat{y}_{T+1} = \sum_{t=1}^{T} w_t \cdot y_t$ with weights w_t. The minimum size of the estimation sample is 30. An illustration of how weights are assigned by SPB original and MB-S will first be given on the basis of a single data set, and the MSFE accuracies of all the methods are subsequently presented for when the exercise is repeated 10,000 times.

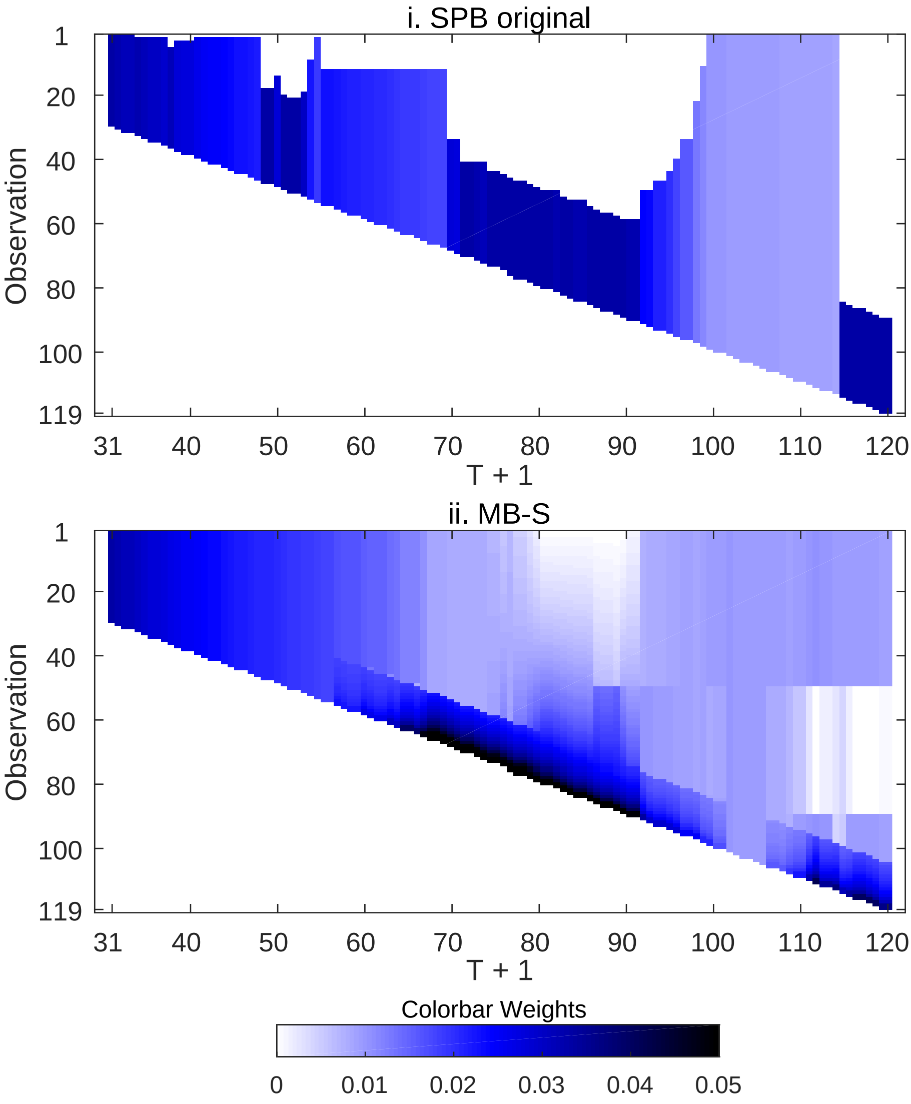
Figure 5.1 gives an example of simulated data according to the data generating process just described. Note that when y_{40} is predicted, all observations should receive about an equal weight in averaging over y. Exponential weights could be used in case there are a few data points after the first break around $t = 50$. During the second period, observations before $t = 50$ need not be discarded altogether, because they could be of relevance in the unknown future. After the second break at $t = 90$, the second period should eventually get a lower weight, while the first period can be more emphasized in estimating the underlying process.

Figure 5.2 shows how SPB original and MB-S assign weights to observations for different forecasts of y_{T+1} across time. The darker a dot, the higher a weight. In the upper panel, SPB original starts with assigning equal weights to all observations. That is, the column of 'Observations' from row 1 to row 30 has the same color at 'T+1' equals 31. As the colorbar indicates, the observations receive a weight of $1/30 \approx .03$. Equal weights are used for $T + 1 = 31$ because $minT + V = 30$, so only the first starting point is eligible. Note that data are already discarded before the first break point has taken place at $T + 1 = 50$. From $T + 1 = 72$ to 92, only the last 30 observations are typically selected. After the second break at $T + 1 = 90$, the included number of observations slowly increases until the entire estimation sample is used. From $T + 1 = 117$ onwards, the smallest possible sample size is employed once more.

The lower panel presents the weights assigned by MB-S. At $T + 1 = 40$, all eligible data receive an equal weight. After the break around $T + 1 = 50$, the validation sample is emphasized more. Once a sufficient amount of post-break data has become available, discrete weights get more pronounced ($T + 1 = 87$). Observations in the first period are not entirely discarded in the second period and this will reduce the prediction error once the simulated mean jumps back from 5 to 3 at $T + 1 = 90$. In the wake of the forecasting inaccuracy at that time, equal weights are quickly emphasized in the third period until the new underlying structure sufficiently reveals itself once more.

Having given an example of how weights are assigned by SPB original and MB-S in case there are two breaks in the drift, I will now repeat this simulation study 10,000 times. Starting with the three benchmark methods, Table 5.4 shows the average MSFE of the benchmark methods at a particular point or period in time. All scores are relative to the MSFE of equal weights (EQ).

Figure 5.2: *Heatmap of Weights Across Time: Two Breaks in Drift*



Note: this heatmaps show the weights w_t assigned to Observations (vertical axis) that were used by SPB original (panel i) and MB-S (panel ii) in estimating μ_t at a certain time (horizontal axis). Simulated model: $y_t = \mu_t + \epsilon_t$, with $\mu_t = 3 + 2 \cdot \mathbb{1}_{50 \leq t \leq 89}$ and $\epsilon_t \sim \mathcal{N}(0, 1)$. Prediction model: $\hat{y}_{T+1} = \hat{\mu}_T$.

Table 5.4: *Relative MSFE for Two Breaks in Drift: Benchmark methods*

	t=55	t=89	t=90	t=95	t=120	P1	P2	P3	All
SPB original	**0.88**	**0.46**	**2.67**	**1.63**	**0.70**	**1.01**	**0.66**	**1.23**	**0.84**
minT=20 V=20	0.94	0.46	2.63	1.88	0.96	1.00	0.76	1.36	0.93
minT=20 V=10	0.88	0.47	2.62	1.25	0.72	1.01	0.65	1.14	0.81
minT=10 V=20	0.89	0.46	2.70	1.88	0.70	1.01	0.68	1.33	0.87
minT=10 V=10	0.79	0.47	2.68	1.25	0.72	1.02	0.57	1.07	0.74
SPW original	**0.95**	**0.54**	**1.92**	**1.57**	**1.05**	**1.00**	**0.79**	**1.38**	**0.96**
minT=20 V=20	0.97	0.61	1.68	1.52	1.17	1.00	0.86	1.37	1.00
minT=20 V=10	0.95	0.55	1.90	1.51	1.06	1.00	0.78	1.36	0.95
minT=10 V=20	0.95	0.54	1.94	1.62	1.05	1.00	0.79	1.40	0.96
minT=10 V=10	0.91	0.52	2.05	1.56	0.94	1.01	0.70	1.32	0.89
EXP	**0.69**	**0.52**	**2.00**	**1.48**	**0.94**	**1.02**	**0.60**	**1.24**	**0.80**

Note: the table reports MSFEs relative to those of the equal weighted prediction, $\text{MSFE}_i/\text{MSFE}_{EQ}$, where MSFE_i is the MSFE of forecasting method i in the first column. The header 't=55' refers to the one-period ahead prediction regarding $t = 55$. The reference setup for each method is boldfaced. Variations to a reference setup, like 'minT=20 V=20', are presented below it. Simulated model: $y_t = \mu_t + \epsilon_t$, with $\mu_t = 3 + 2 \cdot \mathbb{1}_{50 \leq t \leq 89}$ and $\epsilon_t \sim \mathcal{N}(0, 1)$. Prediction model: $\hat{y}_{T+1} = \hat{\mu}_T$. Repetitions: $10,000$ times. P1: obs. 31-49. P2: obs. 50-89. P3: obs. 90-12. All: obs. 31-12.

'SPB original' is the best starting point method. The results highlight that it can indeed be too slow in adapting to changes in the data generating process (H_0^1) and that it is too quick in ignoring old information (H_0^2). I will give a short walk-through of the outcomes of SPB original. For the one-period ahead prediction regarding the observation at $t = 55$, the relative MSFE of SPB original is 0.88. This improvement of 12% relative to EQ is considerably smaller than EXP. Once more post-break data has become available, at $t = 89$, SPB original gets an excellent score by ignoring old parts of the data. As a consequence, SPB original is no less than 2.67 times worse than EQ when the data generating process goes back from 5 to 3 at $t = 90$. At the fifth forecast after the second break ($t = 95$), EQ still outperforms SPB original by a large amount. In the period before the first break ('P1'), SPB original has about the same forecasting accuracy as EQ, in the second period it outperforms EQ by 34%, and in the third period it is 23% worse than EQ on average. Column 'All' shows that the overall MSFE score relative to EQ is 0.84.

The four rows directly below the boldfaced SPB original allow us to analyze how forecasts change when the reference setup of SPB original is altered in terms of the minimum number of observations in the training ($minT$) and validation (V) samples. The global performance of SPB original improves in case they are decreased from fifteen to ten ('minT=10 V=10'), because such small sample sizes help in responding more quickly to a new data generating process (or random noise).

Table 5.5: *Relative MSFE for Two Breaks in Drift: SPB, SPW and MB*

Method	t=55	t=89	t=90	t=95	t=120	P1	P2	P3	All
SPB original	.88	.46	2.67	1.63	.70	1.01	.66	1.23	.84
SPW original	.95	.54	1.92	1.57	1.05	1.00	.79	1.38	.96
EXP	.69	.52	2.00	1.48	.94	1.02	.60	1.24	.80
SPB	**.69**	**.48**	**2.48**	**1.08**	**.82**	**1.02**	**.58**	**1.00**	**.73**
w_{PE}=EQ	.69	.47	2.49	1.17	.82	1.02	.58	1.05	.74
w_V=EQ	.99	.49	2.40	1.03	.93	1.00	.81	1.09	.90
w_{PE} & w_V=EQ	1.00	.48	2.42	1.16	.94	1.00	.81	1.16	.91
minT=20 V=20	.69	.48	2.49	1.16	.81	1.02	.60	1.03	.75
minT=20 V=10	.69	.50	2.40	.98	.84	1.02	.60	.98	.74
minT=10 V=20	.69	.47	2.55	1.17	.79	1.02	.58	1.04	.74
minT=10 V=10	.68	.49	2.44	.99	.82	1.02	.55	.97	.70
λ=0	.66	.46	2.66	1.18	.71	1.03	.55	1.04	.72
λ=2/3	.69	.68	1.49	1.06	.84	1.02	.65	.94	.76
SPW	**.69**	**.52**	**2.02**	**1.39**	**.93**	**1.02**	**.60**	**1.18**	**.79**
MB	**.69**	**.53**	**2.23**	**1.07**	**.72**	**1.02**	**.62**	**.90**	**.73**
$w_{PE} = EQ$	.69	.53	2.21	1.16	.71	1.02	.62	.93	.74
$w_V = EQ$	1.00	.58	2.05	1.01	.73	1.00	.91	.87	.91
w_{PE} & w_V=EQ	1.00	.58	2.04	1.15	.73	1.00	.91	.94	.93
minT=20 V=20	.69	.61	2.07	1.16	.71	1.02	.62	.94	.74
minT=20 V=10	.69	.54	2.12	.92	.73	1.02	.65	.85	.74
minT=10 V=20	.69	.51	2.37	1.16	.71	1.02	.60	.95	.73
minT=10 V=10	.68	.53	2.24	.92	.73	1.02	.60	.85	.71
λ=0	.68	.47	2.63	1.07	.71	1.02	.60	.93	.73
λ=2/3	.69	.64	1.59	1.07	.82	1.02	.64	.94	.75
D_a(w)	.69	.52	2.29	1.08	.72	1.02	.62	.91	.73
BP-LWZ	.69	.53	2.23	1.07	.72	1.02	.62	.90	.73
EB	.69	.52	2.22	1.07	.74	1.02	.60	.91	.72
BPB	.68	.52	2.24	1.09	.74	1.02	.59	.91	.71

Note: the table reports MSFEs relative to those of the equal weighted prediction, $\text{MSFE}_i/\text{MSFE}_{EQ}$, where MSFE_i is the MSFE of forecasting method i in the first column. The reference setup for each method is boldfaced. Variations to a reference setup, like 'minT=20 V=20', are presented directly below it. Simulated model: $y_t = \mu_t + \epsilon_t$, with $\mu_t = 3 + 2 \cdot \mathbb{1}_{50 \leq t \leq 89}$ and $\epsilon_t \sim \mathcal{N}(0, 1)$. Prediction model: $\hat{y}_{T+1} = \hat{\mu}_T$. Repetitions: $10,000$ times. P1: obs. 31-49. P2: obs. 50-89. P3: obs. 90-12. All: obs. 31-12.

The same story goes for SPW original, although its scores are generally worse than SPB original. Despite the fact that the errors of SPW original are smaller at the second break, SPW original performs poorly in the third period. At $t = 120$, for example, it is still unable to beat EQ.

Turning to the third benchmark method, EXP responds admirably to the first break at $t = 50$. Its forecasting error at $t = 90$ is quite a bit smaller than the other benchmark methods, and yet EXP remains substantially less accurate than EQ in the third period. This is a first indication that PPP's exponential weights are not robust to varying break processes (H_0^3). Next to the fact that old information is too easily discarded at times, another disadvantage is that EXP does not inform us about when the underlying break (approximately) occurs.

Now I will switch to three methods that incorporate exponential weights in a discrete weighing scheme on the one hand, and penalize deviations from equal weights on the other hand. These techniques either select the best starting point ('SPB'), weigh starting points ('SPW'), or use multiple break points ('MB'). Their forecasting accuracies are compared to the benchmark methods in Table 5.5.

SPB is slightly worse than EQ in the first period because it exponentially weighs the validation sample. Compared to the original SPB procedure, the method adjusts the mean more quickly after the first break. Due to a quick recovery after the second break, SPB manages to avoid worsening EQ in the third period on average. The rows directly below the boldfaced SPB again show what happens when certain configurations are altered in the reference setup. If the prediction errors are not weighed exponentially but equally ('w_{PE}=EQ'), the response to the second break gets worse at $t = 95$. The row labeled 'w_V=EQ' clearly shows that exponentially weighing the validation sample has a large positive effect on overall forecasting accuracy.

Next, it is confirmed that penalizing deviations from equal weights with λ_{AST} helps to achieve more conservative deviations from EQ ($H_0^{2.1}$). Setting 'λ=0' instead of 1/3 results in a good overall MSFE, but a poorer MSFE during the break at $t = 90$. At the expense of a slightly worse overall score, the forecasting performance of SPB becomes even less volatile when $\lambda = 2/3$. SPW is improved considerably by using exponential weights, but its predictions during the third period remain poor.

As could be expected, MB performs particularly well in the third period. Exponentially weighing the validation sample in MB has the same large influence on forecasting accuracies as in SPB. Regarding hypothesis 2.2, a striking result is that, in terms of forecasting accuracy, one might as well equally distribute

Table 5.6: *Relative MSFE for Two Breaks in Drift: SPB-S, MB-S, and EXP-S*

Method	t=55	t=89	t=90	t=95	t=120	P1	P2	P3	All
SPB original	.88	.46	2.67	1.63	.70	1.01	.66	1.23	.84
SPW original	.95	.54	1.92	1.57	1.05	1.00	.79	1.38	.96
EXP	.69	.52	2.00	1.48	.94	1.02	.60	1.24	.80
SPB	.69	.48	2.48	1.08	.82	1.02	.58	1.00	.73
SPW	.69	.52	2.02	1.39	.93	1.02	.60	1.18	.79
MB	.69	.53	2.23	1.07	.72	1.02	.62	.90	.73
SPB-S	**.94**	**.52**	**2.23**	**1.03**	**.88**	**1.00**	**.67**	**1.05**	**.80**
MB-S	**.94**	**.53**	**2.07**	**1.04**	**.78**	**1.00**	**.68**	**.97**	**.78**
w_{PE}=EQ	.97	.52	2.04	1.13	.77	1.00	.71	1.03	.82
w_V=EQ	.94	.53	1.99	1.02	.76	1.00	.68	.93	.78
w_{PE} & w_V=EQ	.97	.53	1.97	1.17	.75	1.00	.71	1.03	.82
minT=20 V=20	.95	.54	1.96	1.06	.78	1.00	.72	1.02	.82
minT=20 V=10	.93	.55	2.01	1.00	.80	1.00	.67	.94	.77
minT=10 V=20	.94	.52	2.16	1.06	.78	1.00	.68	1.01	.80
minT=10 V=10	.88	.54	2.08	1.00	.79	1.00	.65	.93	.75
λ=0	.69	.49	2.42	1.07	.73	1.00	.60	.97	.73
λ=2/3	.99	.91	1.10	1.01	.97	1.00	.94	1.00	.96
$D_a(w)$	.93	.52	2.08	1.05	.76	1.00	.65	.97	.77
BP-LWZ	.94	.53	2.07	1.04	.78	1.00	.68	.97	.78
EB	.94	.53	2.07	1.04	.79	1.00	.68	.98	.79
BPB	.94	.54	2.09	1.04	.80	1.00	.67	.98	.78
EXP-S	**.89**	**.53**	**1.96**	**1.09**	**.99**	**1.00**	**.65**	**1.11**	**.80**
w_{PE}=EQ	.94	.53	1.98	1.28	.99	1.00	.67	1.15	.82
V=20	.92	.53	1.97	1.12	.99	1.00	.66	1.11	.81
V=10	.86	.54	1.94	1.05	.99	1.00	.64	1.09	.79
λ=0	.70	.52	1.99	1.18	.95	1.00	.61	1.13	.78
λ=2/3	.98	.80	1.25	1.03	1.00	1.00	.89	1.03	.94
$D_a(w)$	.98	.54	1.93	1.05	1.00	1.00	.68	1.10	.82

Note: the table reports MSFEs relative to those of the equal weighted prediction, $\text{MSFE}_i/\text{MSFE}_{EQ}$, where MSFE_i is the MSFE of forecasting method i in the first column. The reference setup for each method is boldfaced. Variations to a reference setup, like 'minT=20 V=20', are presented directly below it. Simulated model: $y_t = \mu_t + \epsilon_t$, with $\mu_t = 3 + 2 \cdot \mathbb{1}_{50 \leq t \leq 89}$ and $\epsilon_t \sim \mathcal{N}(0, 1)$. Prediction model: $\hat{y}_{T+1} = \hat{\mu}_T$. Repetitions: $10,000$ times. P1: obs. 31-49. P2: obs. 50-89. P3: obs. 90-12. All: obs. 31-12.

breaks (‘EB’) across the training sample instead of employing more sophisticated techniques. In MB, I have used Bai and Perron’s procedure with a BIC criterion for determining the number of breaks included. Changing BIC into LWZ does not affect mean prediction errors. Selecting the single best break point (‘BPB’) through cross-validation while allowing for pre-break data to receive positive weights leads to results that are more closely akin to MB than to SPB. From this we might tentatively conclude that the manner in which weights are assigned to periods of observations has a larger influence on the results than the manner in which break dates are identified.

Lastly, I will analyze in Table 5.6 what happens when the weights of SPB, MB, and EXP are shrunk towards equal or exponential weights based on the their latest $V = 15$ predictions. The shrinkage step decreases forecasting errors of the three procedures at the second break ($t = 90$) and leads to smaller improvements relative to EQ at the end of the third period ($t = 120$). For MB-S, the largest difference in predictive quality results from altering the tuning parameter λ, which allows the user to balance cross-validated accuracy with the simplicity of EQ.

Can we leave out discrete weights altogether here and simply optimize over the extent to which exponential weights are used? ‘EXP-S’ shows that the addition of discrete weights does help to improve forecasts. Particularly in the third period, is it clear that break points or starting points should be estimated before shrinking discrete weights towards EQ or EXP.

Overall, the first impression about the main hypotheses is that SPB original indeed responds slowly to new information. This might be explained as a cautionary strategy when the choice is to shorten the sample after a recent number of aberrant observations. Yet, once a second break is introduced, the return to the full sample also takes a long time. Exponential weights were shown to respond quickly to the first break point, but to ignore old data too soon as well. As an alternative, MB or MB-S algorithm can be used, which appear to be less volatile than EXP and SPB original. Will the same conclusions hold when different simulation exercises are considered?

5.3.2. *Alternative Simulation Exercises*

Having analyzed one simulation study in-depth, I will now present more global results of other simulation models. Table 5.7 gives an overview of the exercises. With ‘Drift-1’, the mean in the exercise of the previous section has a simulated break of size 1 instead of 2. The exercise labeled ‘X’ replaces the mean by a regressor with a standard normal distribution. The ‘AR’ task is loosely inspired by Pesaran and Timmermann (2005), who showed that the use of post-break data can lead to poor forecasts with autoregressive models, which is why it is interesting to see whether the weighing schemes are robust when the AR coefficient becomes less or more persistent. The fourth simulation exercise analyses the continuous break process of a standard random walk model.

Table 5.8 reports forecasting performances across the three periods, which are defined by breaks at $t = 50$ and $t = 90$. Panel i shows that when the average of y is simulated to jump from 3 to 4 to 3 instead of from 3 to 5 to 3, the overall performance of the three benchmark methods (SPB original, SPW original, and EXP) are more volatile than MB and MB-S. The same holds for panel ii. with the 'X' exercise. Methods like BPB and MB, that assign individual weights to pre-break observations, particularly outperform others in the third period.

The lower left panel of Table 5.8 indicates that when there is a break in an autoregressive parameter, it is best to weigh observations equally. Consequently, PPP's exponential weights have much difficulties here. In case weights are shrunk towards EQ or EXP based on L_{AST} scores, equal weights are often used, which is why SPB-S, MB-S, and EXP-S perform about as well as EQ here. These methods also produce good forecasts in the continuous break process of the random walk model (panel iv), particularly once a sufficient number of forecasts is available to combine discrete weights with EXP.

In general, MB and MB-S have the most consistent performance among the different techniques for weighing observations. It is also noteworthy that assigning individual weights to pre-break data through BPB often improves SPB; and that equally distributing breaks across the training sample, '(EB) MB', results in forecasts that are about the same as the Bai and Perron procedure for selecting the timing and number of breaks ($H_0^{2.2}$) in all of the exercises. Next, I will turn to an empirical case study.

Table 5.7: *Overview Alternative Simulation Exercises*

	Simulation model	Specify	Prediction model
Drift-1	$y_t = \mu_t + \epsilon_t$	$\mu_t = 3 + \mathbb{1}_{50 \leq t \leq 89}$	$\hat{y}_t = \hat{\mu}_t$
X	$y_t = \beta_{1,t} X_t + \epsilon_t$	$\beta_{1,t} = 3 + 2 \cdot \mathbb{1}_{50 \leq t \leq 89}$	$\hat{y}_t = \hat{\beta}_{1,t} X_t$
AR	$y_t = \phi_{1,t} y_{t-1} + \epsilon_t$	$\phi_{1,t} = 0.3 + 0.2 \cdot \mathbb{1}_{50 \leq t \leq 89}$	$\hat{y}_t = \hat{\mu}_t + \hat{\phi}_{1,t} y_{t-1}$
RW	$y_t = \mu_t + \epsilon_t$	$\mu_t = \mu_{t-1} + v_t$	$\hat{y}_t = \hat{\mu}_t$

$X_t, v_t, \epsilon_t \sim N(0,1)$, $t = 1, 2, \ldots, 120$, forecasting horizon $h = 1$.

Table 5.8: *Relative MSFE for Alternative Simulation Exercises*

	i. Drift-1				ii. X			
	P1	P2	P3	All	P1	P2	P3	All
SPB original	1.01	.83	1.09	.94	1.01	.61	1.34	.80
SPW original	1.00	.90	1.14	.99	1.00	.74	1.55	.94
EXP	1.02	.80	1.09	.92	1.02	.54	1.36	.76
SPB	1.02	.81	1.00	.90	1.02	.51	1.04	.66
SPW	1.02	.80	1.08	.92	1.02	.54	1.27	.74
MB	1.02	.82	.98	.90	1.02	.56	.86	.65
(EB) MB	1.02	.81	.98	.90	1.02	.53	.88	.64
BPB	1.02	.81	.98	.90	1.02	.51	.89	.63
SPB-S	1.00	.90	1.02	.95	1.00	.61	1.09	.74
MB-S	1.00	.90	1.01	.95	1.00	.62	.96	.72
EXP-S	1.00	.88	1.03	.94	1.00	.60	1.18	.75
	iii. AR				iv. RW			
	P1	P2	P3	All	P1	P2	P3	All
SPB original	1.01	.99	1.05	1.01	.79	.46	.30	.43
SPW original	1.00	.99	1.03	1.00	.88	.64	.49	.60
EXP	1.05	.98	1.04	1.01	.47	.44	.42	.43
SPB	1.05	.98	1.02	1.01	.48	.37	.29	.34
SPW	1.05	.98	1.04	1.01	.48	.44	.40	.43
MB	1.05	.98	1.02	1.01	.48	.37	.28	.34
(EB) MB	1.05	.98	1.02	1.01	.48	.39	.31	.36
BPB	1.05	.98	1.02	1.01	.48	.36	.27	.33
SPB-S	1.00	.99	1.00	1.00	.90	.40	.30	.41
MB-S	1.00	.99	1.00	1.00	.90	.39	.29	.40
EXP-S	1.00	.99	1.01	1.00	.90	.44	.42	.48

Note: the table reports MSFEs relative to those of the equal weighted prediction, $\text{MSFE}_i/\text{MSFE}_{EQ}$, where MSFE_i is the MSFE of forecasting method i in the first column. The title of each panel refers to the simulation exercise defined in Table 5.7. Repetitions: $10,000$ times. P1: obs. 31-49. P2: obs. 50-89. P3: obs. 90-12. All: obs. 31-12.

5.4. Empirical Application

In his paper, Croushore (2008) describes how survey forecasts were ignored as a source of data on people's expectations as a result of the rational expectations literature in the 1980s. The surveys appeared to contain irrational expectations, in the sense that inflation forecasts were systematically too high or too low. One way of testing whether the mean survey forecasts are biased, is to estimate the bias-adjusted model

$$\bar{f}^{ba}_{t+h} = \alpha + \beta \bar{f}_{t+h},$$

where the mean survey forecast, $\bar{f}_{t+h} = \frac{1}{N}\sum_{i=1}^{N} f_{t+h,i}$, gives the average h-quarter-ahead prediction at time t of all the participating experts $i = 1, 2, \ldots, N$. If expert forecasts are unbiased, then one should estimate $\alpha = 0$ and $\beta = 1$. Since experts might tend to overpredict in one period and underpredict in another (Capistrán and Timmermann, 2009a), the underlying data generating process could be subject to breaks. I will examine whether the bias-adjusted model improves upon the mean survey forecast and whether the bias-adjusted model produces better forecasting accuracies when observations receive unequal weights.

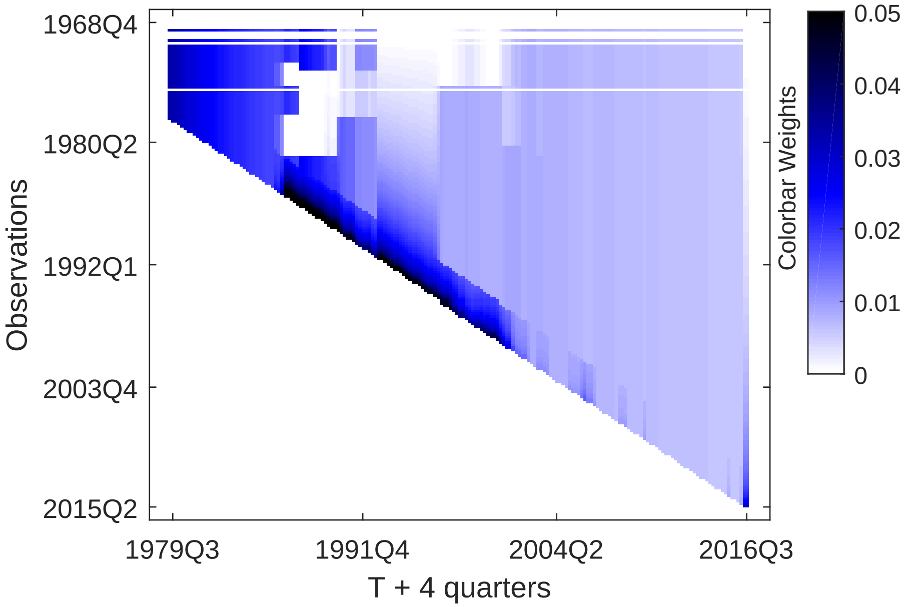
The data on expert predictions is obtained from the Survey of Professional Forecasters ('SPF'), which has been conducting the survey about macroeconomics variables on a quarterly basis since 1968Q4. Inflation forecasts are obtained by transforming the quarterly data on the Price Index for the Gross Domestic Product ('PGDP') with

$$x_{t+h} = 400 \cdot \ln \frac{X_{t+h}}{X_{t+h-1}}, \tag{5.7}$$

following (Capistrán and Timmermann, 2009b). Next to the one quarter and one year ahead predictions of PGDP, I will analyze expert forecasts of the Nominal Gross Domestic Product ('NGDP'), which were also transformed with equation (5.7).

Figure 5.3 shows the weights that MB-S assigned to observations (vertical axis) when predicting inflation at a given time (horizontal axis) for one-year-ahead forecasts. The validation sample and the minimum training sample are both of size 15. Notice that some data points are missing (white horizontal lines). This is because four-quarters-ahead forecasts made at 1968Q4 can only be evaluated with real-time data five quarters later; and because, for some reason, one-year-ahead forecasts were not collected on a number of occasions.

Figure 5.3: *Heatmap of Weights Across Time: SPF Data*



This heatmap shows the weights that were assigned to 'Observations' (vertical axis) in producing 'T + 4 quarters' ahead forecasts (horizontal axis) of inflation with the bias-adjusted model.

Table 5.9: *MSFE for Bias-Adjusted Model: Relative to Mean Survey Forecast*

	PGDP		NGDP	
	1 qr ahead	1 yr ahead	1 qr ahead	1 yr ahead
EQ	**1.12**	**1.67**	**1.01**	**1.02**
SPB original	1.00	1.36	1.02	1.23
SPW original	1.02	1.39	1.04	1.04
EXP	0.98	1.28	1.05	1.06
SPB	1.01	1.31	1.04	1.07
SPW	1.00	1.26	1.05	1.09
MB	1.01	1.15	1.04	1.09
EB	1.02	1.26	1.03	1.10
SPB-S	1.02	1.38	1.02	1.02
MB-S	1.03	1.35	1.02	1.01
EXP-S	1.01	1.35	1.01	1.01

Note: the table reports MSFEs of the bias-adjusted model whereby different schemes (first column) were used to weigh observations in estimating the regression parameters. The scores are made relative to the MSFE of the mean survey forecasts $\bar{f}$. Data are PGDP and NGDP (1968Q4-2016Q3). The first 30 available mean survey forecasts are used in the initial estimation sample.

MB-S used equal weights to estimate $\bar{f}^{ba}_{t+4}$ until these weights resulted in some relatively poor forecasts in the 1980s. As of 1986Q4, groups of observations before around 1980 were often disregarded in the second step of the algorithm, and these discrete weights were shrunk to exponential ones in the third step. From 2001Q3 onwards, equal weights were generally employed once more.

Table 5.9 reports the MSFE results of the bias-adjusted model relative to the mean survey forecast. When α and β are estimated with equally weighted observations ('EQ'), it clearly worsens mean survey predictions for PGDP. Assigning unequal weights to observations clearly helps to improve the bias-adjusted model in the case of PGDP, although the results are still worse than the unadjusted mean survey forecasts.

In the case of NGDP, the bias-adjusted model with equally weighted observations has a similar performance as the mean survey forecast ($\bar{f}$). Assigning weights to observations deteriorates estimates of the bias-adjusted model, particularly for the original SPB method. Penalizing deviations from equal weights helps to avoid such bad results, especially when the possibility is added to shrink previously obtained weights to EQ or EXP in the third part of the algorithm (SPB-S, MB-S, EXP-S).

What we learn from this application is that the presumed biases in survey forecasts of PGDP and NGDP are not of such a structural nature that they can be corrected for in real-time. Regarding the MB-S algorithm, it is pleasing to observe that its weights only deviate from equal weights if the accuracy of the validation window is markedly better (PGDP); and that they do not deviate from EQ when the relative improvement in MSFE is small (NGDP).

5.5. Discussion

In this chapter, I have studied three ways to improve upon the best starting point method. First, it was shown how the slow response time of SP methods could be improved with the help of exponential weights. Second, I explained how a tuning parameter $\lambda_{AST} \in [0, 1]$ enables a researcher to make the tradeoff between cross-validation and a prior setup, so that more conservative deviations from equal weights can be stimulated. Third, a broadly applicable procedure was introduced for weighing observations of multiple periods instead of just assigning a positive weight to post-break data. Overall, it was found that the resulting MB-S algorithm offers robust estimates of breaks dates and regression parameters.

A possible limitation is that the Bai and Perron (2003) procedure used for estimating the timing of breaks can take a long time to run. When a plain model, like regressing y on a constant, is estimated for a sample of 2000 observations, MB-S takes around 45 sec. while SPB takes 0.7 sec., for example. A suggestion for further research is to look for ways in which multiple break points can be found more quickly. I have presented simple alternatives to BP, like equally distributing breaks across the estimation sample. Such a simple alternative could function as a prior setup if one desires to estimate the timing of breaks with an Accuracy-Simplicity Tradeoff.

5.A. Appendix: Heuristics for Determining the Effective Sample Size

I will now show why $D_s(w)$ and $D_a(w)$ may be interpreted as measures for the relative deviance from equal weights; how a heuristic for the 'effective' number of observations follows; and how differences between the two measures can be illustrated.

I begin with

$$D_s(w) = \frac{1}{\sum_m w_m^2} \sum_m (w_m - \frac{1}{M})^2,$$

where $\sum_m w_m = \sum_m^{M=1} w_m = 1$. Without further assumptions, D_s can be rewritten as

$$\begin{aligned} D_s(w) &= \frac{1}{\sum_m w_m^2} \sum_m (w_m - \frac{1}{M})^2 \\ &= \frac{1}{\sum_m w_m^2} \Big[\sum_m (w_m^2) + \frac{1}{M} - 2\frac{1}{M} \sum_m (w_m) \Big], \\ &= \frac{M - \hat{\mathbb{N}}_2}{M}, \end{aligned}$$

where $\hat{\mathbb{N}}_2 = \frac{1}{\sum_m w_m^2}$ is the effective sample size with an ℓ_2 norm and where I have used that $\sum_{m=1}^{M} c = Mc$. It follows that $0 \leq D_s(w) \leq 1$.

In case the weights still need to be normalized, one would get

$$\begin{aligned} D_s(w) &= \frac{(\sum_m w_m)^2}{\sum_m w_m^2} \sum_m (\frac{w_m}{\sum_m w_m} - \frac{1}{M})^2, \\ &= \frac{M - \frac{(\sum_m w_m)^2}{\sum_m w_m^2}}{M}, \end{aligned}$$

and the heuristic for the effective sample equals $\frac{(\sum_m w_m)^2}{\sum_m w_m^2}$. The same expression was derived in the design effect literature by determining what adjusted sample size is required to equate the variance of an individually weighted average $var(\tilde{X}) = \sigma^2 \frac{\sum_m w_m^2}{(\sum_m w_m)^2}$ to the variance of an equal weighted average $var(\bar{X}) = \frac{\sigma^2}{N}$ under the assumption of independent and identically distributed observations (Kish, 1965).

Next, I turn to the absolute deviance measure,

$$D_a(w) = \frac{1}{2} \sum_m |w_m - \frac{1}{M}|.$$

Note that $D_a(w) = 0$ when $w_m = \frac{1}{M}$ for all $m = 1, 2, \ldots, M$. When the first N observations are included and the others are ignored, this can be represented as $w^{1:N}$. Such weights are an example of when N observations receive an equal weight of $\frac{1}{N}$ and the other $(M - N)$ observations receive a weight of 0. If one also uses that $N \leq M$, the following can be derived

$$\begin{aligned} D_a(w^{1:N}) &= \frac{1}{2} \sum_m |w^{1:N} - \frac{1}{M}|, \\ &= \frac{1}{2} \Big[(M - N) \cdot |0 - \frac{1}{M}| + N \cdot |\frac{1}{N} - \frac{1}{M}| \Big], \\ &= \frac{1}{2} \Big[\frac{M - N}{M} + 1 - \frac{N}{M} \Big], \\ &= \frac{M - N}{M} = D_{01}(w), \end{aligned}$$

and this proportion of observations is between 0 and 1. Since the largest deviation from equal weights occurs when only the minimum amount of observations is included, it also follows that $0 \leq D_a(w) \leq 1$. The heuristic for the effective sample size $\hat{\mathbb{N}}_1 = (1 - D_a(w))M$ results from solving $D_a(w) = \frac{M - \hat{\mathbb{N}}_1}{M}$ for $\hat{\mathbb{N}}_1$.

Table 5.10 compares the two deviance measures when $N = 3$ observations receive weights of $w = [p\ q\ r]'/(p + q + r)$. The letter r is associated with the

Table 5.10: *Measuring Deviations from Equal Weights*

	i. D_s						ii. D_a					
p\q	0	1	2	3	4	5	0	1	2	3	4	5
5						0						0
4					.01	.01					.05	.05
3				.06	.04	.05				.12	.08	.10
2			.18	.12	.10	.11			.22	.17	.15	.17
1		.40	.29	.23	.21	.21		.38	.29	.22	.23	.24
0	$\frac{2}{3}$	.54	.44	.37	.34	$\frac{1}{3}$	$\frac{2}{3}$	.50	.38	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

Note: this table shows how D_s and D_a measure the total deviation from equal weights. The three weights are given by $w = [p\ q\ r]'/(p + q + r)$, where $r = 5$.

third observation and is fixed at 5 while p and q of observations one and two vary from 0 to 5. When both p and q equal zero (and $r = 5$), for example, then the deviance from equal weights is 2/3 for both measures, because two out of three observations are excluded. The total deviances from equal weight is 0 in case $p = q = r = 5$. As one would expect, D_a gives higher penalties than D_s to small deviations from equal weights, while D_s has more discriminatory power for large deviations from equal weights. For instance, when $p = 0$, D_a is *exactly* equal to 1/3 for $q = 3, 4$, and 5; while D_s gives higher penalties than 1/3 the more q diverges from $r = 5$. In this way, D_s stimulates periods of observations to receive similar weights because larger deviations from EQ are more heavily penalized by the ℓ_2 norm.

6

A Quick and Easy Search for Optimal Configurations

6.1. Introduction

In econometric analyses one often has to make choices about configurations like the starting point of a data set, the penalty term of an estimator, or weights that are used when combining forecasts. One popular procedure for the selection of configurations is called cross-validation. The data is first split into a training and a validation set. Varying configurations are then used to produce pseudo forecasts with the training set about the observations of the validation set. The configurations with the best pseudo predictions are subsequently selected for producing out-of-sample forecasts.

When a single statistical decision like the starting point of a sample is selected by optimizing over a validation set, it might already take a long time to evaluate the full grid of all eligible candidates. The computation time will increase considerably when multiple statistical choices are cross-validated. For these reasons, tools are needed that identify for which range of settings it is worthwhile to investigate many candidate configurations, and which areas do not have to be studied in close detail.

To give an example, say a researcher optimizes over a tuning parameter $c \in [0, 1]$ by applying cross-validation on the grid $\{0, 0.01, 0.02, \ldots, 0.99, 1\}$. By equally distributing the candidate values in this way, areas of configurations that produce forecasts which barely deviate from each other are examined just as closely as areas that produce highly dissimilar forecasts. Turning the tables, one can also choose new candidate configurations such that the resulting predictions vary by a similar degree across all neighbouring configurations. Just find out where the average forecasting deviance (‘FD’) between two contiguous configurations is largest, add the point that lies in the middle, and repeat this process.

Once more settings have been globally spread across the configuration space in this way, one might become more interested in focusing on local areas with good pseudo predictions. In the latter case, a configuration c can be added that lies between two neighboring configurations which have on average the best pseudo forecasting accuracy (‘FA’) according to some loss function.

By giving the user an intuitive control over a gradual transition from a selection based on forecasting deviance towards a selection based on forecasting accuracy, configuration searches can be performed far more efficiently. The new ‘FAD’ (Forecasting Accuracy Deviance) that I aspire to introduce, is a global to local strategy. A choice that is based on data-optimization will be referred to as an ‘item’. The items may be discrete or continuous and there may be multiple items. FAD can also be applied when information criteria like AIC or BIC are used to select configurations, or when cross-validation is employed with multiple folds.[1]

The main requirement for an FD search is that neighboring configurations ($c = 0$ and $c = 0.25$) produce more similar forecasts on average than configurations that are further apart ($c = 0$ and $c = 1$). FD is unaffected by further conditions such as the convexity of the accuracy measure. An FA search is barely troubled by forecasts of contiguous configurations being more similar or not, but does require the optimization problem to be convex. By combining FD and FA in FAD, the global and local search techniques help to overcome each other's liabilities, so that FAD can find multiple (local) minima.

Widely used methods for selecting trial configurations are grid and random searches (Bergstra and Bengio, 2012). With a grid search, the user manually chooses candidate sets of configurations for each item and computes every possible combination among these configurations. Its main advantage is that it is straightforward to apply and interpret. One disadvantage is that the grid search suffers from the curse of dimensionality, which means that the number of permutations grows exponentially as more items are added. A practitioner could manually adjust the set of configurations to be examined, but this can be a cumbersome exercise in practice.

In a random search, configurations are independently drawn from a uniform density with the same manually defined configuration space as the one spanned by a regular grid, like $\mathscr{U}(0, 1)$ (Bergstra and Bengio, 2012). The random search has

[1] In g-fold cross-validation, the estimation sample is split into g folds and each fold is predicted with the other folds precisely one time.

the same efficiency in the relevant set of parameters as if the search algorithm had only been applied to those relevant dimensions, because many unique configurations are evaluated per item. When there are many items, the random search may thus find a better set of configurations than a grid search for a given number of runs. A disadvantage could be that the entire estimation procedure needs to be rerun for each unique set of configurations, which might slow the program down.

As a third benchmark method, I will consider a more sophisticated search algorithm which starts with a random search and continues by iteratively predicting which set of configurations will lead to the largest Expected Improvement ('EI') (Jones et al., 1998, Bartz-Beielstein et al., 2005, Hutter et al., 2011). Here the assumption is that configurations that are closer together have a more similar predictive performance. Extending the random search with EI is known to improve in-sample accuracy for a given number of runs. The time required for estimating expected improvements through a stochastic model might be much larger though. Bergstra and Bengio (2012) write that 'of course, random search can probably be improved by automating what manual search does, i.e., a sequential optimization, but this is left to future work' (pp. 283).

Regarding the structure of this chapter, Section 6.3 explains how FAD can help to automate what a manual search does with a sequential optimization procedure. In Section 6.3, the search techniques will be evaluated with a variety of simulation exercises. Section 6.4 applies the search methods to multiple choices that can be made when combining a number of top-ranked expert forecasts. Section 6.5 concludes.

6.2. Search Methods

6.2.1. *Benchmark Methods*

As I mentioned above, the grid search typically spreads candidate configurations equally across a manually defined space, while the random search selects candidate configurations from a uniform distribution. Although both methods seem easy to apply, there are several reasons why the grid and random searches could be inefficient when searching within a single item. For one, the procedures heavily depend on the initial range that is given. For another, they might include many configurations that lead to highly similar forecasts. Lastly, they do not discriminate in selecting configurations between those that lead to good forecasts and those that lead to poor ones.

In a setting where multiple items are to be optimized over, the grid search will equally distribute configurations across the space by examining all combinations between the sets of configurations examined. When there are two items $A \in [0, 1]$ and $B \in [10, 100]$, for example, one could define $A \in \{0, 0.1, \ldots, 1\}$ and $B \in \{10, 20, \ldots 100\}$ and take combinations $(A = 0, B = 10)$, $(A = 0, B = 20)$, and so on. Note that this means that only 11 unique configurations are considered for item A and 10 for item B. When more items are added, it may only be feasible to assess fewer unique configurations per item.

In case the candidate configurations are drawn from a uniform distribution, by contrast, many unique configurations of a given item are selected at least once, which is why the random search could be more efficient when the number of items increases. As Bergstra et al. (2012) write, the 'random search has the same efficiency in the relevant subspace as if it had been used to search only the relevant dimensions' (pp. 284). The idea that the random search could beat the grid because it evaluates many unique configurations for each item could also lead one to expect that the random search is slower to execute. The reason is that when only a few settings are altered while others remain the same, it may often be unnecessary to recompute large parts of the algorithm. This could be a comparative advantage of the grid search.

The Expected Improvement ('EI') search attempts to refine these aspects of the random search by adding new configurations based on their expected performance. The performance of configuration c is measured with a score function $f(c)$ that is typically based on the accuracy of the model. It is defined in such a way that a lower score is better than a higher one. Think of Root Mean Squared Forecast Errors (RMSFE). After running an estimation procedure with a number of randomly drawn configurations, one has obtained a data set consisting of those initial configurations and their associated scores. The relation between the scores and the configurations can then be estimated with this data set, and these estimates can subsequently be used to predict the unknown scores of configurations that have not been evaluated yet. EI iteratively adds configurations whose Expected Improvement in the score is largest.

A set of configurations with M items is summarized in the M-vector c. In case the vector has not been evaluated yet, it is denoted as c^*. Given the best score that has been observed so far ($f_{\min}$), the expected improvement of an unknown vector of configurations c^* is given by

$$\mathbb{E}[I(c^*)] := \mathbb{E}\Big[\max\{0, f_{\min} - f(c^*)\}\Big].$$

To predict scores of unknown configurations, one can regress

$$f(c_i) = \mu + \epsilon_i, \tag{6.1}$$

where μ is an unknown constant, c_i is a set of configurations for which the scores have been computed, and ϵ_i is the error term. The trick is to note that configurations that are closer together may be expected to have a more similar error (ϵ_i) from the mean score μ.

To capture these correlations, the distribution of ϵ_i is given by $\mathscr{N}(0, \sigma^2)$ with covariance $V(c_i, c_j) = \sigma^2 A(\theta, \epsilon(c_i), \epsilon(c_j))$. The Gaussian correlation function A is formulated as

$$A(\theta, \epsilon(c_i), \epsilon(c_j)) = \prod_{m=1}^{M} \exp(-\theta_m (c_i^m - c_j^m)^2), \tag{6.2}$$

where c_i^m refers to item m of a set of configurations i. The parameters μ, σ^2, and θ_m can be estimated by applying maximum likelihood on the data of sampled configurations and scores (Jones et al., 1998, pp. 460). Note that the errors of c_i and c_j are indeed more closely related in A when c_i and c_j are closer to each other. The parameter θ_m can be regarded as a measure of importance of item m (*ibid.*, pp. 459).[2] The stochastic model defined by equations (6.1) and (6.2) is called 'DACE' ('Design and Analysis of Computer Experiments') following (Sacks et al., 1989).

If $r = 1, 2, \ldots, R$ is the number of runs that has been performed so far (the number of configurations for which scores have been computed), the DACE predictor of the unknown score of configuration c^* is given by

$$\hat{f}(c^*) = \hat{\mu} + a' A^{-1}(f(c) - \mathbf{1}\hat{\mu}),$$

where $\mathbf{1}$ is an r-vector of ones and a is an r-vector with correlations between the error terms of c^* and the previously sampled points c_i. The mean squared error of the predictor is

$$s^2(c^*) = \sigma^2 \left[1 - a' A^{-1} a + \frac{(1 - \mathbf{1} A^{-1} a)^2}{\mathbf{1}' A^{-1} \mathbf{1}} \right],$$

which is larger the less it is correlated to the sampled points and which is zero at the sampled points, because $a' A^{-1} a$ will then be 1 (Jones et al., 1998, pp. 462).

[2] I follow Hutter et al. (2009) and others in using a power of 2 in equation (6.2), instead of estimating this parameter as well, like Jones et al. (1998).

To arrive at the expected improvement $\mathbb{E}[I(c)]$, one can now define $u := \frac{f_{\min} - \hat{f}(c^*)}{s}$ and compute

$$\mathbb{E}[I(c)] = s \cdot \Big[u \cdot \Phi(u) + \phi(u) \Big],$$

for standard normal pdf (ϕ) and cdf (Φ). The expected improvement of configurations c^* is large when it is far removed from a set of configurations whose score is known (through s) and when the expected score $\hat{f}(c^*)$ is low (*ibid.*, pp. 470-473). In this way, EI gives an automated tradeoff between exploiting known good areas and exploring unknown areas (Hutter et al., 2011, pp. 515). The scores are often log-transformed, in which case the optimization criterion becomes

$$I_{\text{exp}}(c) := \max\{0, f_{\min} - e^{f(c)}\}.$$

Defining $v := \frac{\log(f_{\min}) - \hat{f}(c_i)}{s}$, one then obtains

$$\mathbb{E}(I_{\text{exp}}(c)) = f_{\min}\Phi(v) - e^{\frac{1}{2}s^2 + \hat{f}(c_i)} \cdot \Phi(v - s),$$

following (Hutter et al., 2009).

Jones et al. (1998) formulated an Efficient Global Optimization procedure ('EGO') for continuous parameters. Many other variations were developed to deal with discrete and categorical configurations; and scores that can vary when they are recomputed (such as computation times), see Hutter et al. (2011) for an overview. Following Jones et al., I will initialize the EI procedure with a random sample of size $10M$, where M is the number of items.[3] Subsequently, a new set of configurations that has the largest Expected Improvement is added until a prespecified maximum number of runs is reached. I will use a well-documented Matlab toolbox to estimate the DACE model (Lophaven et al., 2002). Since it is relatively cheap (takes little time) to predict unknown scores of configurations c^* *once* the DACE model is estimated, I will assess the expected improvement of 10^4 randomly drawn configurations each time a new configuration is to be added (Hutter et al., 2009, Bartz-Beielstein et al., 2005).

Although the EI procedure is known to improve upon the random search when it comes to selecting promising new configurations, its estimation procedure is more complicated. It could also take a much longer time to run because the stochastic model needs to be estimated and because EI tends to select unique

[3] One may use a random Latin Hypercube design to initialize EI with continuous configurations, but Hutter et al. (2009) remark that there is 'very little variation in predictive quality due to procedures used for constructing the initial design.'

configurations for multiple items. The computation time will be empirically examined in the following sections. Like the grid and random searches, the exploration of unknown areas by EI is largely based on the distance between configurations rather than the average deviance between their resulting predictions; and I will argue that such a strategy for a global search may be inefficient.

6.2.2. *A Global to Local Search: FAD*

To develop an alternative approach that is quick and simple, one could start by observing that the grid and random searches more or less equally distribute configurations across a space. Switching perspectives, one can also choose configurations such that the average deviances between the resulting forecasts are more equally distributed. Once there is a sufficient number of configurations to avoid local minima, the researcher can subsequently focus on subsets of configurations with a good forecasting accuracy.

In selecting candidate configurations such that the deviance between forecasts becomes more equal, I will use a forecasting deviance measure that describes the relevance of statistical choices in terms of their influence on forecasts. Let $\hat{y}_i$ denote predictions that are generated with a set of configurations c_i regarding the observations in the validation sample $v = 1, 2, \ldots, V$. An average absolute deviance measure between the forecasts of two sets of configurations labeled i and j can then be given by

$$FD_a(\hat{y}_i, \hat{y}_j) = \frac{1}{V} \sum_{v=1}^{V} |\hat{y}_{i,v} - \hat{y}_{j,v}|. \tag{6.3}$$

A value of $FD_a = .1$ means that the average absolute deviation among the predictions of c_i and c_j is .1. When the FD_a between the highest and lowest configuration of a statistical choice is zero, this means that the two forecasts are exactly the same for all $v \in [1, V]$. It would then make little sense to study such a decision in further detail. Conversely, if the average forecasting deviance is relatively large between two extremes of a statistical decision, then it might be interesting to analyze more of its configurations. Since I will be optimizing over squared errors (RMSFE), I will consider squared deviances,

$$FD_s(\hat{y}_i, \hat{y}_j) = \sqrt{\frac{1}{V} \sum_{v=1}^{V} (\hat{y}_{i,v} - \hat{y}_{j,v})^2}. \tag{6.4}$$

An FD measure can be used to develop an automated procedure for selecting candidate configurations. Continuing the example with $c \in [0, 1]$, one can start by comparing the pseudo predictions of the extremes $c = 0$ and $c = 1$. In case the forecasting deviance is high, a candidate weight is added that is halfway between 0 and 1; namely, $c = 0.5$. Subsequently, the pseudo predictions of $c = 0.5$ can be compared to those of $c = 0$ and 1 to determine whether $c = 0.25$ or 0.75 should be examined. The next configuration that is added is the one where the forecasting deviance between two contiguous configurations is largest.

Note that a focus on forecasting deviances does not imply that the distance between configurations is irrelevant. After all, forecasting deviances are only evaluated between configurations that are closest neighbors. In fact, the main assumption is that predictions of neighbouring configurations i and $i + 1$ are more alike on average than those of i and $i + 2$. An accuracy measure plays no role here, so assumptions regarding convexity do not apply. Possible stopping criteria are a minimum amount of forecasting deviance, a maximum number of runs, and/or a maximum amount of computation time.

New configurations can also be selected based on forecasting accuracy. One may, for example, use a root mean squared prediction error,

$$RMSFE(\hat{y}_i) = \sqrt{\frac{1}{V}\sum_{v=1}^{V}(y_v - \hat{y}_{i,v})^2}, \tag{6.5}$$

as a performance measure. The average pseudo forecasting accuracy between two consecutive configurations can then be given by

$$FA_s(\hat{y}_i^{\mathrm{V}}, \hat{y}_i^{\mathrm{V}}) = \left[\frac{1}{2}\Big(\mathrm{RMSFE}(\hat{y}_i) + \mathrm{RMSFE}(\hat{y}_j)\Big)\right]^{-1},$$

where an inverse is used to ensure that FA is high when the average accuracy is high. To avoid extreme values, I will add the rule that if the lowest average between RSMFEs is ξ less than .001, then ξ will be added to all RMSFEs before FA_s is computed.[4] A high average forecasting accuracy between two neighboring configurations could indicate that the accuracy of the configuration that lies in the middle will be high as well. So, a sequential procedure to find an optimal configuration could be to iteratively select the middle of two configurations whose FA_s is largest.

Now there are two measures. The FD measure may search the global space efficiently, but it does not focus on configurations that have good pseudo forecasts.

[4] The inverse is taken over the average $\frac{1}{2}\Big(\mathrm{RMSFE}(\hat{y}_{i,t}) + \mathrm{RMSFE}(\hat{y}_{j,t})\Big)$ instead of the individual RMSFEs because the average is less prone to extreme values.

The FA measure concentrates on promising subsets, but it might easily end up in a local minimum or even get stuck along the way. By using a global to local strategy, one first focuses on an FD approach at the start of the procedure and one gradually turns to an FA approach as more configurations have efficiently been spread across the space.

With a tuning parameter $\phi \in [0,1]$, the researcher can specify how important the relative FD scores are compared to the relative FA scores in the following weighted average

$$FAD(y_i, y_j) = (1-\phi)\ \log\ \overbrace{\frac{FA_{ij}}{\max(FA)}}^{\text{Relative accuracy}} + \phi\ \log\ \overbrace{\frac{FD_{ij}}{\max(FD)}}^{\text{Relative deviance}}\ , \tag{6.6}$$

where $\max(FA)$ is the maximum value of all the FA's. $FAD(y_i, y_j)$ is high when the average deviance between forecasts of two consecutive configurations is high and/or when their average forecasting accuracy is high. The log is taken to mitigate the influence of extreme dissimilarities in forecasts.[5] The configuration that lies in the middle of two consecutive settings with the highest FAD is added. I will use the rule that if the maximum forecasting deviance is zero, configurations will be added based on FA.

The researcher can choose ϕ by specifying after how many runs (R_u) he wants to focus more on FA than FD by defining $\phi = \frac{R_u}{R_u+r}$, where r is the number of runs at a given time. This means that configurations will first ($r < R_u$) mostly be selected based on FD and that FA will gradually become more important as the number of runs increases ($r > R_u$). In the reference setup, I will set $R_u = \frac{1}{2}R$ and I will compare this specification to others.[6]

When there are multiple items, like $A \in [0,1]$ and $B \in [10,100]$, the basic FAD procedure circles through two steps again. In step 1, pseudo predictions are generated for all new sets of configurations. At first, these are all of the possible combinations of the initial configuration sets of each item. Step 2 is to select a set of configurations based on FAD scores. In case configuration $A = 0.5$ is added conditional on $B = 50$, the extremes ($A = 0, B = 50$) and ($A = 1, B = 50$) will also be added if they were not already included before, in order to make sure that this apparently relevant dimension can be fully investigated. Steps 1 and 2 are iterated until some stopping criterion is reached, such as a maximum number

[5] Since $\alpha \log \beta = \log \beta^{\alpha}$, it is clear that relative differences between FD values decrease as $\phi \in [0,1]$ decreases.

[6] One could also define ϕ in terms of computation time $\theta \in [0,\Theta]$, by choosing after which amount of time, $\Theta_u = \frac{1}{2}\Theta$, FA should be as important as FD in $\phi = \frac{\Theta_u}{\Theta_u+\theta}$.

of runs R. I will investigate whether it is necessary to round off continuous configurations to two decimal places in order to prevent the algorithm from getting stuck in one dimension.

In case FAD is applied to multiple items, situations might arise whereby there is only a single configuration of one item conditional on the other items. One might then be concerned about not being able to expand each set of configurations in all directions. To study this potentially relevant issue, a small extensions ('step 3') is included in the reference setup of FAD. The third step involves placing 'anchors' around the current best set of configurations (c_b) and it is performed on each occasion that step 2 has been repeated another 10 times. Anchors are placed by adding the extremes of each configuration in c_b and by adding two configurations that are 'nearby'. To find configurations that are nearby, a list for each item m will be made of all the unique configurations that have been included so far, and the configuration that is closest above and below c_b^m will be added (if they are new). I will examine empirically whether this addition results in a quicker convergence to the optimal value.

Lastly, I will make two remarks. First, FAD might be more sensitive to the assumption that neighboring configurations produce more similar forecasts than EI is to the assumption that scores of neighboring configurations are more similar. The main reason is that EI predicts the expected improvement of a set of configurations by making use of all other sets that have been evaluated up to that point, whereas FAD only uses the predictions of two direct neighbours. Second, it is good to remark that there is no tree structure in the optimization procedure of FAD, in the sense that solutions do not depend on a user-specified order in which items are optimized over. A tree structure can nevertheless be implemented for the FAD procedure as well.

6.2.3. *Hypotheses*

Table 6.1: *Hypotheses*

	Few items	Many items	Time
Grid	+	-	+
Random	0	+	+
EI	+	+	-
FAD	+	+	+

Few/Many items: Methods is good (+), neutral (0), bad (-) in finding optimal configurations.

Time: Methods is quick (+), unremarkable (0), or slow (-) in finding a set of configurations.

Table 6.1 summarizes the main hypotheses that I will examine in the simulation studies and the empirical application. I will compare the efficiency of the existing search methods with the alternative FAD procedure in terms of the RMSFE accuracy of the selected configurations and in terms of the time it takes for the procedures to run. As aforementioned, the grid search is regarded to be an efficient and easy-to-apply search technique when there are a few items, and the random search is known to perform better when there are many items. Neither of the two techniques is known for resulting in a long computation time. The more complicated EI search may require fewer configurations to be evaluated before an optimal set is found in both cases, although its computation time is known to be longer. As an alternative, it will be examined whether the FAD search is a quick method for finding optimal configurations.

In evaluating the search techniques, the size of the training and validation samples will both be 15. One can think of a time series application whereby the 15 most recent observations are used as a validation set to select configurations. Larger validation samples will just make it easier for EI and FAD to predict where the most relevant configurations can be found. This assertion will be evaluated empirically as well.

Simulation studies are used to compare the performance of the search methods when the object is to find a single configuration. An empirical application is subsequently used to examine how well they perform when there are multiple items.

6.3. Simulation Studies

As a first step in analyzing the above hypotheses for search problems with a single item, I will apply the procedures to Ridge regression and show how configurations are selected. Subsequently, the MSFE accuracy and the running time of the techniques are analyzed based on three simulation studies. There will be one exercise where forecasts barely differ for large parts of the sample space, one exercise where the change in forecasts is proportional to the change in configurations, and one exercise where forecasts deviate strongly in large parts of the sample space.

6.3.1. *Choosing the Tuning Parameter of Ridge Regression*

I will first study the selection of the Ridge tuning parameter. Let the linear regression model be given by

$$y = X\beta + \epsilon, \tag{6.7}$$

with an $N \times 1$ vector representing N observations of the dependent variable y, an $N \times K$ matrix of $k = 1, 2, \ldots, K$ explanatory variables X, an $N \times 1$ vector of disturbances ϵ, and a $K \times 1$ vector of parameters β. In Ridge regression, the loss function for estimating β is given by

$$L_{Ridge} = (y - Xb)'(y - Xb) + \lambda b'b,$$

so that the size of $b_{OLS} = (X'X)^{-1}X'y$ is restrained by penalizing large $b'b$. Solving the first-order condition for b leads to the following solution

$$b_{\text{Ridge}} = (X'X + \lambda I_K)^{-1}(X'y).$$

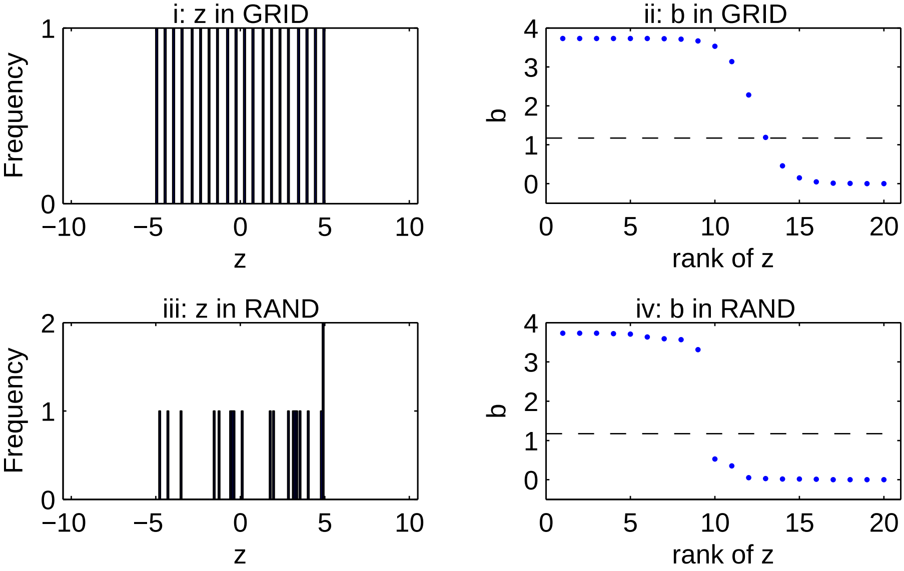
The penalty term can be any positive number, $\lambda > 0$, which makes it difficult to anticipate which configurations are good candidates.

Cross-validation is frequently applied to select the penalty term for these types of methods, and oftentimes, a grid of the form $\lambda = 10^z$ is used for a hundred equally distributed values of z (Zou and Hastie, 2005), whereby the range of the grid may change per application (Friedman et al., 2010, pp. 17). After dividing the estimation sample into a training and validation sample, the training sample is used to estimate $b^{\text{T}}_{\text{ridge}}$ for various values of λ. Forecasts are then generated regarding the validation sample with $\hat{y}^{\text{V}} = b^{\text{T}}_{\text{ridge}}X^{\text{V}}$. The penalty term with the lowest RMSFE is subsequently selected.

In the following, I will use a single predictor X in equation (6.7) and simulate the column vectors X and ϵ to be standard normally distributed. A sample of $N = 30$ observations is split halfway into a training set used for estimating $b^{\text{T}}_{\text{ridge}}$ and a validation set used for evaluating z in $\lambda = 10^z$. The coefficient in the training sample will be set at $\beta^{\text{T}} = 4$ and in the validation sample at a fraction q of this parameter, so that $\beta^{\text{V}} = q \cdot \beta^{\text{T}}$. The object is to find the right penalty λ for shrinking the regression parameter of the training sample ($b^{\text{T}}_{\text{ridge}}$) towards the optimal regression parameter of the validation sample $b^{\text{V}}_{\text{OLS}}$.

Each procedure will be allowed to perform $R = 20$ runs. To avoid excluding relevant values of z, the initial bounds on z were made as large as $[-10, 10]$. For the grid and random procedures, this space was made twice as small to facilitate more instructive comparisons. So, in the grid approach, the R candidate configurations were equally distributed over $z \in [-5, 5]$ and in the random search, they were drawn from $\mathscr{U}(-5, 5)$.

Figure 6.1: *Ridge regression: Grid and Random Searches*



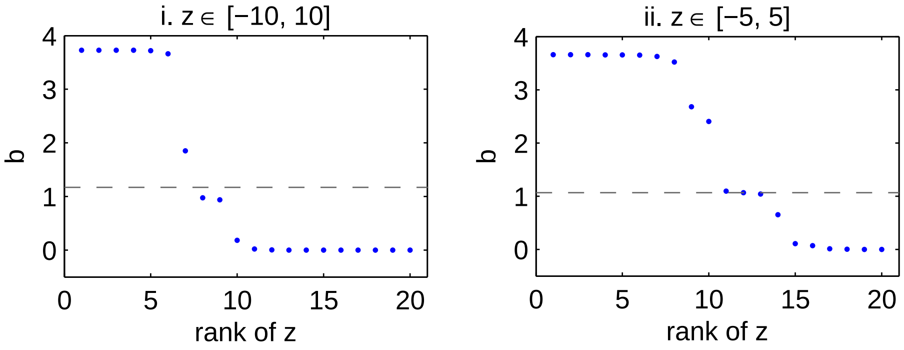
This Figure illustrates how the grid and random searches select configurations for the Ridge exercise. Panels to the left show which values of configuration $z \in \lambda = 10^z$ were selected. The panels to the right show the resulting values $b^{\mathrm{T}}_{\mathrm{ridge}}(\lambda(z))$ for sorted $z_{(\mathrm{rank})}$. The lowest z in the left panel has a rank of 1 in the right panel, the second lowest z in the left panel has a rank of 2 in the right panel, and so on. The first row of panels corresponds to the grid search, and the second row to the random search. Simulated series: $X, e \sim \mathscr{N}(0, 1)$, $b^{\mathrm{T}}_{\mathrm{OLS}} = 3.73$ and $b^{\mathrm{V}}_{\mathrm{OLS}} = 1.17$ (dashed line), $N = 30$, validation sample of $V = 15$. Maximum number of runs is $R = 20$.

As I will explain, Figure 6.1 illustrates that equally distributing configurations can lead to many forecasts of the validation sample that are too similar to be of relevance. The data-optimized value is $b^{\mathrm{T}}_{\mathrm{OLS}} = 3.73$ in the training set and $b^{\mathrm{V}}_{\mathrm{OLS}} = 1.17$ in the validation set.

The left panels of Figure 6.1 indicate which values of the configuration $z \in \lambda = 10^z$ were selected, and the right panels show the resulting $b^{\mathrm{T}}_{\mathrm{ridge}}$ estimates that were obtained based on $z_{(\mathrm{rank})}$. The lowest z has a rank of 1, the second lowest z has a rank of 2, and so on. Configurations were equally distributed for the grid search and selected from a uniform distribution in the random search, as the left panels indicate.

The right panels of Figure 6.1 confirm that many choices of z result in solutions of $b^{\mathrm{T}}_{\mathrm{ridge}}$ that are highly similar. A value of $z = -5$ (first bar in upper left panel) leads to a $b^{\mathrm{T}}_{\mathrm{ridge}}$ of 3.73 (first point in upper right panel). A value of $z = -4.5$ (second bar in upper left panel) also leads to a $b^{\mathrm{T}}_{\mathrm{ridge}}$ of 3.73 (second point in upper right panel). In this example, it so happens that one of the configurations of the grid search is close to the optimal value. That is, the thirteenth lowest value of z results in a $b^{\mathrm{T}}_{\mathrm{ridge}}$ of 1.19, which is close to $b^{\mathrm{V}}_{\mathrm{OLS}} = 1.17$ (dashed line in upper right panel). The lower right panel indicates that the $b^{\mathrm{T}}_{\mathrm{ridge}}$ values of the random search are far removed from the optimal value at the dashed line in this data set.

Figure 6.2: *Ridge Regression: EI*

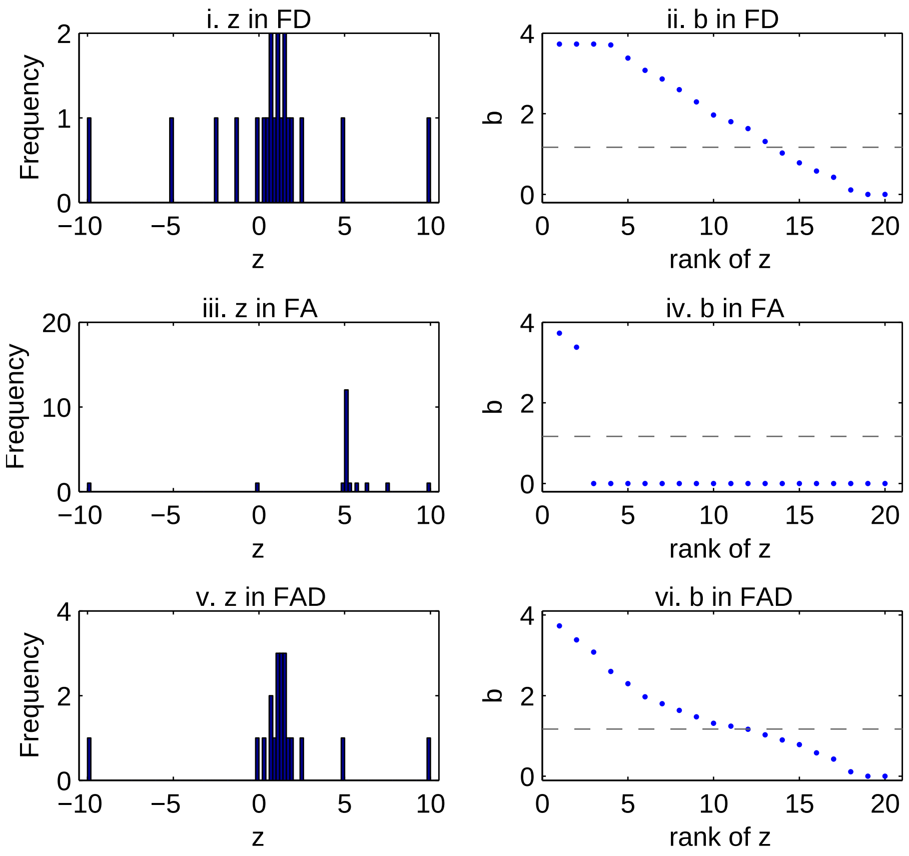


The panels show the values $b^{\mathrm{T}}_{\mathrm{ridge}}(\lambda(z))$ for sorted $z_{(\mathrm{rank})}$ when applying the EI search. The lowest z has a rank of 1, the second lowest z a rank of 2, and so on. The left panel gives uses bounds $z \in [-10, 10]$, the right panel $z \in [-5, 5]$.Simulated series: $X, e \sim \mathcal{N}(0, 1)$, $b^{\mathrm{T}}_{\mathrm{OLS}} = 3.73$ and $b^{\mathrm{V}}_{\mathrm{OLS}} = 1.17$ (dotted line), $N = 30$, validation sample of $V = 15$. Maximum number of runs is $R = 20$. EI is initialized with 10 randomly drawn configurations. Outcomes can differ when search is repeated with same simulated data.

Figure 6.2 gives an example of how the EI search chooses candidate z. Ten configurations were selected uniformly at random, and ten configurations were subsequently added based on the expected improvement. Even when the initial range is [-10, 10], EI does quite well to approximate the optimal value at the dashed line in panel i. This plot can change substantially when the search is repeated. In case the initial range of z is made twice as small (panel ii), EI gets even closer to 1.17. The spread of candidate $b^{\mathrm{T}}_{\mathrm{ridge}}$ is not so well, because there are many values around the edges and not so many values in between.

The reason is that EI's global search is largely based on the distance between configurations.

Figure 6.3: *Ridge Regression:* FD_s*,* FA_s*, and* FAD_s

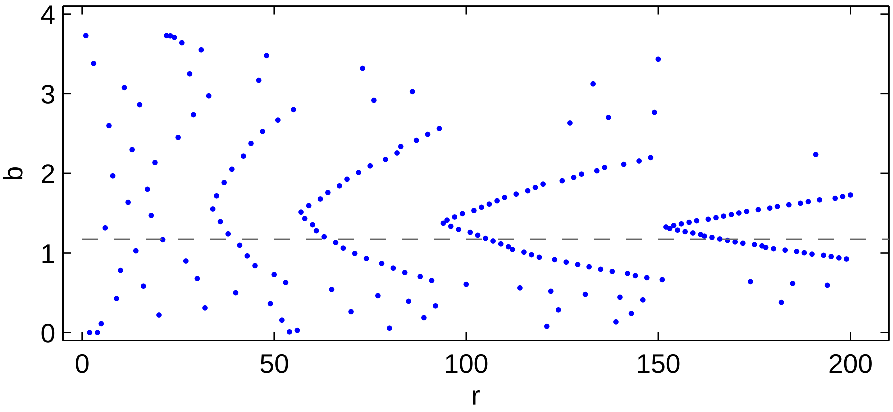


This figure illustrates the selection of configurations for FD, FA, and FAD for the Ridge example with $z \in [-10, 10]$. Panels to the left show the frequency with which certain values of z were selected. The panels to the right show the resulting values $b^{\mathrm{T}}_{\mathrm{ridge}}(\lambda(z))$ for sorted $z_{(\mathrm{rank})}$. The lowest z has a rank of 1, the second lowest z a rank of 2, and so on. In the first row configurations are selected based on forecasting deviances (FD_s), in the second row based on forecasting accuracy (FA_s), and in the third row based on FAD_s with $R_u = 10$. Simulated series: $X, e \sim \mathcal{N}(0, 1)$, $b^{\mathrm{T}}_{OLS} = 3.73$ and $b^{\mathrm{V}}_{\mathrm{OLS}} = 1.17$ (dotted line), $N = 30$, validation sample of $V = 15$. Maximum number of runs is $R = 20$.

The upper panels of Figure 6.3 present the selection of configurations for the FD search. Panel i indicates that values of z are unevenly distributed in the area $[-10, 10]$. Most of the configurations are between $z = 0$ and 3. Panel ii shows that the resulting forecasts $(X^{\mathrm{V}})b^{\mathrm{T}}_{\mathrm{ridge}}$ are quite evenly spread across the space, since $b^{\mathrm{T}}_{\mathrm{ridge}}$ is generally shrunk in a linear fashion from around $b^{\mathrm{T}}_{\mathrm{OLS}} = 3.73$ to 0. There are also a couple of configurations that lead to similar forecasts. The four lowest values of z all lead to $b^{\mathrm{T}}_{\mathrm{ridge}}$ values close to 3.73, for example. The number of such superfluous cases is often less when the simulation exercise is repeated.

The second row of Figure 6.3 presents the selection of candidate z for FA_s. The FA search does not converge to $b^{\mathrm{V}}_{\mathrm{OLS}} = 1.17$ but sticks to around zero instead, because, for $z_{(\mathrm{rank})}$ in panel ii, $FA_s(\hat{y}^{\mathrm{V}}_{z_3}, \hat{y}^{\mathrm{V}}_{z_4}) > FA_s(\hat{y}^{\mathrm{V}}_{z_2}, \hat{y}^{\mathrm{V}}_{z_3})$. The lower panels illustrate how FAD selects configurations for $R_u = \frac{1}{2}R = 10$. Note that FA helps FD so that there are no redundant configurations at the start, and FD helps FA to find the minimum at $b^{\mathrm{V}}_{\mathrm{OLS}} = 1.17$.

Figure 6.4: *Ridge Regression: Sequence of Configurations Added by FAD*



This figure illustrates the sequence with which configurations are selected by FAD in the Ridge exercise for a total of $R = 200$ runs and $z \in [-10, 10]$. A dot represents the $b^{\mathrm{T}}_{\mathrm{ridge}}(\lambda(z))$ value of the r^{th} configuration that was added. Simulated series: $X, e \sim \mathcal{N}(0, 1)$, $b^{\mathrm{T}}_{OLS} = 3.73$ and $b^{\mathrm{V}}_{\mathrm{OLS}} = 1.17$ (dotted line), $N = 30$, validation sample of $V = 15$.

Figure 6.4 presents in what sequence configurations are added for $R = 200$ and $R_u = 100$. At the start of this global to local procedure, FAD chooses configurations in such a way that the resulting forecasts $(X^{\mathrm{V}})b^{\mathrm{T}}_{\mathrm{ridge}}$ are evenly spread. As more configurations are added, the walls around the optimal value start to close. Observe that, rather than adding more configurations redundantly close to the best configuration so far, FAD will keep on looking for alternative local minima in a systematic fashion. The FD measure determines what 'redundantly close' is and the tuning parameter $\phi = \frac{R_u}{R_u + r}$ specifies that the FD measure is emphasized less as more configurations are added.

Having illustrated the search methods with Ridge regression, the performance of the procedures will now be studied in terms of RMSFE for when there are large subspaces of configurations with similar forecasts; when the sample space is well-defined; and when there are large dissimilarities between forecasts in a large part of the configurations space.

6.3.2. *Large Subspace with Similar Forecasts*

In the first assessment, I will continue with the Ridge example. The validation sample will be simulated by $\beta^{\mathrm{V}} = q \cdot \beta^{\mathrm{T}}_{\mathrm{sim}}$, where q is drawn from $\mathscr{U}(0,1)$ or specified otherwise. The penalty term $\lambda(z)$ should be chosen in such a way that $b^{\mathrm{T}}_{\mathrm{ridge}}$ is sufficiently shrunk towards zero. The optimal solution with an RMSFE accuracy measure occurs when $b^{\mathrm{T}}_{\mathrm{ridge}} = b^{\mathrm{V}}_{\mathrm{OLS}}$. For each search procedure, I will select the candidate configuration with the lowest RMSFE. In the 'reference' setup, the search procedures are allowed to have $R = 10$ runs. The grid and random procedures choose z from $[-5, 5]$ and the EI and FAD searches are based on $z \in [-10, 10]$. The simulation exercise is repeated 10,000 times.

Table 6.2: *Accuracy Search Methods for Similar Forecasts*

	Ref.	q=0	q=.1	q=.2	q=.3	q=.4	q=.5	q=.6	q=.7	q=.8	q=.9	q=1
GRID	1.14	1.03	1.04	1.06	1.10	1.23	1.34	1.29	1.15	1.06	1.04	1.05
RAND	1.21	1.06	1.07	1.15	1.26	1.34	1.40	1.35	1.25	1.15	1.07	1.06
EI	1.26	1.09	1.13	1.25	1.34	1.38	1.35	1.37	1.35	1.23	1.12	1.08
(no scale) EI	1.27	1.09	1.13	1.25	1.37	1.39	1.36	1.39	1.37	1.24	1.12	1.08
($z \in [-5, 5]$) EI	1.06	1.07	1.06	1.07	1.07	1.05	1.04	1.07	1.09	1.08	1.05	1.05
FAD	1.01	1.02	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.04
FA	1.42	1.04	1.12	1.35	1.64	1.85	1.80	1.70	1.45	1.21	1.07	1.05
($\phi = 0.5$) FAD	1.02	1.02	1.01	1.01	1.02	1.02	1.02	1.02	1.01	1.01	1.01	1.04
FD	1.02	1.02	1.01	1.02	1.02	1.02	1.02	1.02	1.02	1.02	1.02	1.05
(R_u=0.25R) FAD	1.01	1.02	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.01	1.04
(R_u=0.75R) FAD	1.01	1.02	1.01	1.01	1.01	1.02	1.02	1.01	1.01	1.01	1.01	1.04
(R_u=R) FAD	1.02	1.02	1.01	1.01	1.02	1.02	1.02	1.02	1.01	1.01	1.01	1.04
(MSFE) FAD	1.01	1.02	1.00	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.04
(no log) FAD	1.01	1.02	1.00	1.00	1.00	1.01	1.01	1.00	1.00	1.00	1.01	1.04
	R=5	R=6	R=7	R=8	R=9	R=15	R=20	R=100	N=10	N=100	$\lambda \in [0, 10^5]$	R=20, $\lambda \in [0, 10^5]$
GRID	1.41	1.19	1.23	1.19	1.14	1.07	1.04	1.00	1.22	1.10	1.55	1.54
RAND	1.39	1.33	1.28	1.26	1.23	1.14	1.11	1.01	1.34	1.19	2.44	2.41
EI	1.53	1.50	1.40	1.36	1.29	1.09	1.02	NaN	1.40	1.23	1.41	1.31
(no scale) EI	1.56	1.52	1.40	1.37	1.30	1.09	1.02	NaN	1.42	1.23	1.41	1.32
($z \in [-5, 5]$) EI	1.42	1.35	1.23	1.17	1.11	1.00	1.00	NaN	1.11	1.06	1.39	1.31
FAD	1.41	1.14	1.05	1.02	1.01	1.00	1.00	1.00	1.04	1.01	1.50	1.01
FA	1.47	1.42	1.42	1.43	1.42	1.43	1.43	1.42	1.38	1.46	1.55	1.37
(ϕ=0.5) FAD	1.41	1.14	1.05	1.02	1.02	1.01	1.00	1.00	1.04	1.01	1.50	1.01
FD	1.41	1.14	1.08	1.05	1.03	1.01	1.01	1.00	1.08	1.02	1.50	1.01
(R_u=0.25R) FAD	1.41	1.14	1.05	1.02	1.01	1.00	1.00	1.00	1.03	1.00	1.50	1.00
(R_u=0.75R) FAD	1.41	1.14	1.05	1.02	1.02	1.01	1.00	1.00	1.04	1.01	1.50	1.01
(R_u=R) FAD	1.41	1.14	1.05	1.03	1.02	1.01	1.00	1.00	1.05	1.01	1.50	1.01
(MSFE) FAD	1.41	1.14	1.05	1.02	1.01	1.00	1.00	1.00	1.04	1.01	1.50	1.01
(no log) FAD	1.42	1.14	1.05	1.02	1.01	1.00	1.00	1.00	1.03	1.00	1.51	1.00

This table reports RMSFE performances relative to the RMSFE of $b^{\mathrm{V}}_{\mathrm{OLS}}$. The reference setup ('Ref.') for ridge regression: $X, \epsilon \sim \mathscr{N}(0,1)$, $\beta^{\mathrm{T}} = 4$ and $\beta^{\mathrm{V}} = q \cdot \beta^{\mathrm{T}}$ where q is drawn from $\mathscr{U}(0,1)$, estimation sample of $N = 30$, validation sample of $V = 15$, and a total of $R = 10$ runs. The column headers show variations of the simulation exercise, and the row headers enlist the different search techniques. The header 'q=0' signifies that q is set to zero instead of being randomly drawn. '(No scale) EI' means, for example, that the reference setup of the EI method is altered in that no log transformation is applied to scale the scores. In the reference setup for GRID and RAND I define $z \in [-5, 5]$ and for the EI and FAD procedures $z \in [-10, 10]$. EI uses $\log y$. FAD uses $R_u = 0.5R$ to define ϕ. The subsequent rows show what happens when certain settings in a reference setup are altered. Computation time (sec.) in reference setup: GRID (.0007), RAND (.0025), EI (.5117), FAD (.0042).

Table 6.2 reports the RMSFE of a search technique relative to the optimal RMSFE when $b^{\mathrm{V}}_{\mathrm{OLS}}$ is estimated based on the validation sample. In the above defined 'reference' setup, the relative RMSFE score of the grid search is 1.14, which means that it is 14% less accurate than $b^{\mathrm{V}}_{\mathrm{OLS}}$ on average. The grid performs better than the random search. EI has some trouble in finding optimal configurations in case the initial range in its reference setup is as large as $z \in [-10, 10]$. By comparing 'EI' with '(No scale) EI', it is clear that it makes no difference whether the scores that EI predicts are log-transformed or not. When EI also employs the boundaries $z \in [-5, 5]$ instead of $[-10, 10]$, it does find more accurate settings than the grid and random searches.

The reference FAD search, with $R = 10$ runs and a large initial space of $z \in [-10, 10]$, is only 1% removed from an optimal choice. The subsequent rows show how sensitive the FAD procedure is to changes in ϕ, the tuning parameter that regulates how much FD is emphasized relative to FA. It is clear that choosing a different ϕ than $\frac{R_u}{R_u+r}$ does not improve results. The FD procedure works quite well whereas the FA search is no less than 42% removed from the optimal RMSFE. Increasing or decreasing the speed with which FA is emphasized with $R_u = 0.25R$ and $R_u = R$, respectively, does not appear to be necessary. The row labeled '(MSFE) FAD' shows what happens when one does not take a square root of MSFE in FA and a square root in FD in equation (6.4). This decision does not seem to influence results much. A FAD measure that does not include logs, '(no log) FAD', performs excellent as well here.

The columns labeled 'q=0', 'q=.1', ..., 'q=1' illustrate that the accuracies of the search procedures depend on the degree q with which $b^{\mathrm{T}}_{\mathrm{OLS}}$ is shrunk. As could be expected from Figure 6.1, GRID and RAND have more difficulties in finding a good z when $b^{\mathrm{T}}_{\mathrm{ridge}}$ is shrunk half-way ($q = 0.5$). By contrast, the relative performance of the FAD search is 1.01 for all q except $q = 0$ and $q = 1$. The anomalies arise at the extremes of q because $b^{\mathrm{V}}_{\mathrm{OLS}}$ can be below zero for low q or above $b^{\mathrm{T}}_{\mathrm{OLS}}$ for high q. This is something to think about when doing cross-validation.[7]

The columns labeled 'R=6' up to 'R=100' show what happens when the maximum number of configurations is altered. In case $R \leq 5$, all search techniques have relative scores that are higher than 1.30. The random procedure with a 100 runs has a score of 1.01, which is the same as FAD's score after ten runs.

Decreasing the sample size from $N = 30$ to $N = 10$ (with $V = 5$) worsens the

[7] One might use a transformation so that cross-validated outcomes that are close to the extremes become more similar to the extremes.

performance of all search procedures while results improve in case the estimation sample is increased to $N = 100$ (with $V = 50$). When we optimize over the extremely large space of $\lambda \in [0, 10^5]$' instead of over $z \in [10, 10]$, then none of the methods perform well. It is only when the maximum number of runs is increased from $R = 10$ to 20 that the FAD procedure is near-optimal again. In the even more extreme case that $\lambda \in [0, 10^{10}]$, it takes around 40 runs for FAD to converge.

Finally, EI's computation times for $R = 10$, 100, and 200 runs are 0.5, 12, and 40 seconds, respectively. To compare, the grid and random search take around 0.02 seconds to do 200 runs, and FAD around 0.05 seconds. This means that search techniques other than EI can evaluate far more configurations in the same amount of time. In fact, I have not actually computed the EI procedures for a hundred runs ('NaN'), but we can be confident that it will have converged to a score of 1.00 by then.

Table 6.3: *Accuracy Search Methods for Well-Defined Space*

	Ref.	δ=0	δ=.1	δ=.2	δ=.3	δ=.4	δ=.5	δ=.6	δ=.7	δ=.8	δ=.9	δ=1
GRID	1.02	1.03	1.02	1.02	1.02	1.02	1.02	1.02	1.02	1.02	1.02	1.03
RAND	1.09	1.25	1.10	1.07	1.07	1.07	1.07	1.06	1.07	1.07	1.09	1.24
EI	1.00	1.02	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.02
(no scale) EI	1.00	1.02	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.02
FAD	1.00	1.02	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.02
FA	1.00	1.02	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.02
(ϕ=0.5) FAD	1.01	1.02	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.02
FD	1.02	1.03	1.02	1.02	1.02	1.02	1.02	1.02	1.03	1.03	1.02	1.03
(R_u=0.25R) FAD	1.00	1.02	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.02
(R_u=0.75R) FAD	1.01	1.02	1.00	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.00	1.02
(R_u=R) FAD	1.01	1.02	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.02
(MSFE) FAD	1.00	1.02	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.02
(no log) FAD	1.00	1.02	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.02
	R=3	R=4	R=5	R=6	R=7	R=8	R=15	R=20	R=100	N=10	N=100	$X \sim \mathscr{N}(0,1)$
GRID	1.31	1.16	1.09	1.06	1.04	1.03	1.01	1.01	1.00	1.04	1.02	1.01
RAND	1.48	1.34	1.25	1.18	1.15	1.10	1.04	1.03	1.00	1.16	1.07	1.01
EI	1.70	1.37	1.09	1.02	1.00	1.00	1.00	1.00	NaN	1.01	1.00	1.00
(no scale) EI	1.67	1.38	1.08	1.02	1.00	1.00	1.00	1.00	NaN	1.01	1.00	1.00
FAD	1.31	1.09	1.03	1.01	1.01	1.01	1.00	1.00	1.00	1.01	1.00	1.00
FA	1.31	1.09	1.03	1.01	1.00	1.00	1.00	1.00	1.00	1.01	1.00	1.00
(ϕ=0.5) FAD	1.31	1.09	1.04	1.02	1.02	1.01	1.00	1.00	1.00	1.02	1.01	1.00
FD	1.31	1.20	1.09	1.08	1.06	1.03	1.01	1.01	1.00	1.05	1.02	1.01
(R_u=0.25R) FAD	1.31	1.09	1.03	1.01	1.00	1.00	1.00	1.00	1.00	1.01	1.00	1.00
(R_u=0.75R) FAD	1.31	1.09	1.03	1.02	1.02	1.01	1.00	1.00	1.00	1.02	1.01	1.00
(R_u=R) FAD	1.31	1.09	1.05	1.02	1.02	1.01	1.00	1.00	1.00	1.02	1.01	1.00
(MSFE) FAD	1.31	1.09	1.03	1.01	1.01	1.00	1.00	1.00	1.00	1.01	1.00	1.00
(no log) FAD	1.31	1.09	1.03	1.01	1.00	1.00	1.00	1.00	1.00	1.01	1.00	1.00

This table reports RMSFE performances relative to the RMSFE of the optimal $\tilde{\delta}_{OLS}$. Reference setup ('Ref.') for combining forecasts (equation 6.8): $X, \epsilon \sim \mathscr{N}(0,4)$, and q is drawn from $\mathscr{U}(0,1)$, estimation sample of $N = 30$, validation sample of $V = 15$. Maximum number of runs is $R = 10$. The configurations space is $\delta \in \{0,1\}$. The column headers show variations of the simulation exercise, and the row headers enlist the different search techniques. The header 'δ=0' signifies that δ is set to zero instead of being randomly drawn. '(No scale) EI' means, for example, that the reference setup of the EI method is altered in that no log transformation is applied to scale the scores. EI uses $\log y$. FAD uses $R_u = 0.5R$ to define ϕ. Computation time (sec.) for reference setups: GRID (.0006), RAND (.0021), EI (.5068), FAD (.0027).

6.3.3. *Well-Defined Space*

In the second simulation exercise I will investigate a situation where defining the relevant initial range is less of a problem than in Ridge regression. A linear regression model will be estimated with the restrictions that $0 \leq \delta_k \leq 1$, $\sum_{k=1}^{K} \delta_k = 1$, and $K = 2$. This may be interpreted as taking a weighted average between two forecasts x_1 and x_2. One can write

$$y = \delta x_1 + (1 - \delta) x_2 + \epsilon, \tag{6.8}$$

$$y - x_2 = \delta(x_1 - x_2) + \epsilon, \tag{6.9}$$

and define $\tilde{y} = y - x_2$ and $\tilde{x} = x_1 - x_2$, so that the solution that minimizes the sum of squared errors becomes $\tilde{\delta}_{OLS} = (\tilde{x}'\tilde{x})^{-1}\tilde{x}'\tilde{y}$. Furthermore, I will simulate $X \sim \mathcal{N}(0, 4)$ and $\epsilon \sim \mathcal{N}(0, 1)$. The maximum number of candidate configurations in the reference setup is $R = 10$, the sample size is $N = 30$, and the last $V = 15$ observations constitute the validation window on the basis of which δ is selected.

Table 6.3 gives the results for this example. In the reference setup, the random search is quite far removed from an optimal RMSFE score, whereas the grid search performs good and EI and FAD have near solutions after a few runs. I have again not computed EI for a hundred runs. The column labeled $X \sim \mathcal{N}(0, 1)$ shows that if X is simulated with a variance of 1 instead of 4, the exercise has hardly any discriminatory power.

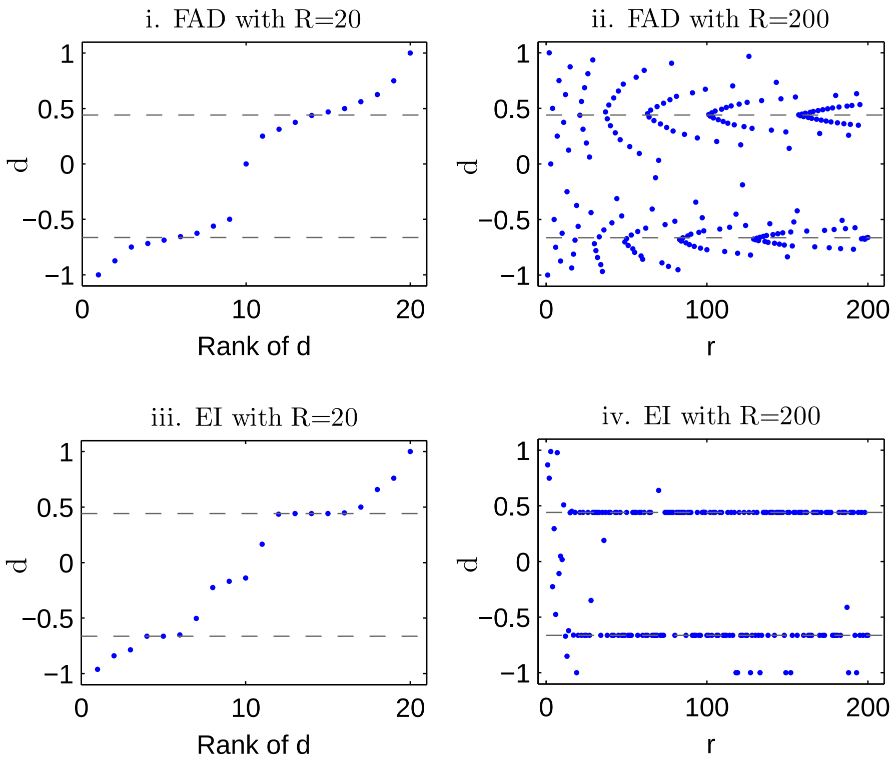
To briefly illustrate what happens when there are multiple local minima, I will generate the data in the same fashion as equation (6.9), but alter the estimation model. Namely, I will specify that

$$\hat{\delta} = \begin{cases} d & \text{if } d \geq 0 \\ d^2 & \text{otherwise} \end{cases}$$

and optimize over $d \in [-1, 1]$. Figure 6.5 presents the selection of configurations for FAD with $R_u = R/2$. The optimal values are $d = \tilde{\delta}^{\text{V}}_{OLS} = .44$ and $d = -\sqrt{\tilde{\delta}^{\text{V}}_{OLS}} = -.66$.

The upper left panel of Figure 6.5 shows how twenty candidate configurations are selected in this example where a *linear* regression parameter is used for d between 0 and 1 and a *squared* parameter between -1 and 0. In the reference setup, about 10 candidate configurations are employed to evenly spread configurations across the space; and the other 10 candidates are distributed around the two optimal values. The upper right panel clearly illustrates how the gradual

Figure 6.5: *Combining Forecasts: FAD and EI*



This plot illustrate how FAD and EI select configurations when data is simulated by $y = \delta X_1 + (1-\delta)X_2 + \epsilon$ with $X \sim \mathscr{N}(0,4)$, $\epsilon \sim \mathscr{N}(0,1)$. The estimated $\hat{\delta} = d$ if $d \geq 0$ and $\hat{\delta} = d^2$ for $d < 0$, and $d \in [-1,1]$. Optimal values are $d = .44$ and $d = -.66$ (dashed lines). The left panels shows which value of $\hat{\delta}$ are associated with the rank of $d_{\text{(rank)}}$, where the lowest d gets a rank of 1, and the second lowest d a rank of 2, and so on. The right panels show which values of $\hat{\delta}$ are obtained when the r^{th} configuration is added.

transition from FD to FA leads FAD to close up on the two global minima. If R is smaller, FAD will also emphasize FA more quickly. The lower panels show that EI does well to find both minima. Where FAD continues to search for possible alternative configurations after locating the two minima, EI mostly keeps on adding configurations closely around the optimal values as the number of runs increases.

Table 6.4: *Accuracy Search Methods for Dissimilar Forecasts*

	Ref.	δ=0	δ=.1	δ=.2	δ=.3	δ=.4	δ=.5	δ=.6	δ=.7	δ=.8	δ=.9	δ=1
GRID	2.60	1.03	1.03	1.22	1.62	2.13	2.68	3.25	3.82	4.25	3.93	3.36
RAND	5.09	4.27	4.44	4.59	4.65	4.62	4.77	4.91	5.52	5.38	5.58	5.80
EI	1.22	1.02	1.02	1.05	1.09	1.13	1.20	1.26	1.32	1.39	1.46	1.52
(no scale) EI	1.56	1.03	1.09	1.23	1.36	1.49	1.62	1.71	1.80	1.87	1.96	1.99
FAD	1.02	1.02	1.01	1.01	1.01	1.01	1.01	1.02	1.03	1.03	1.03	1.05
FA	1.11	1.03	1.04	1.06	1.08	1.05	1.13	1.23	1.12	1.05	1.21	1.12
(ϕ=0.5) FAD	3.30	1.04	1.20	1.60	2.10	2.65	3.22	3.80	4.39	4.98	5.58	6.18
FD	2.58	1.03	1.02	1.08	1.37	1.83	2.36	2.92	3.49	4.08	4.67	5.27
(R_u=0.25R) FAD	1.09	1.02	1.00	1.07	1.22	1.04	1.00	1.01	1.01	1.05	1.28	1.29
(R_u=0.75R) FAD	1.35	1.03	1.04	1.08	1.37	1.81	1.90	1.48	1.13	1.05	1.30	1.73
(R_u=R) FAD	2.58	1.03	1.04	1.08	1.37	1.83	2.36	2.92	3.49	4.08	4.67	5.27
(MSFE) FAD	1.07	1.02	1.01	1.03	1.04	1.04	1.05	1.07	1.10	1.05	1.22	1.11
(no log) FAD	3.30	1.04	1.20	1.60	2.10	2.65	3.22	3.80	4.39	4.98	5.58	6.18

	R=6	R=7	R=8	R=9	R=15	R=20	R=100	N=10	N=100	$X \sim \mathcal{N}$(0,1)
GRID	1.91	1.65	2.19	2.58	1.92	1.57	1.03	3.12	2.50	1.19
RAND	18.70	12.10	9.10	6.35	2.59	1.93	1.11	6.13	4.59	1.82
EI	6.08	2.11	1.47	1.41	1.05	1.01	NaN	1.34	1.19	1.02
(no scale) EI	4.84	2.67	1.90	1.87	1.06	1.02	NaN	1.77	1.52	1.07
FAD	1.35	1.35	1.17	1.07	1.01	1.00	1.00	1.04	1.02	1.01
FA	1.35	1.27	1.24	1.15	1.03	1.02	1.00	1.12	1.10	1.03
(ϕ=0.5) FAD	3.29	3.30	3.31	3.29	3.30	3.31	3.29	4.04	3.14	1.34
FD	2.58	2.59	2.59	2.58	2.59	2.59	2.57	3.12	2.47	1.21
(R_u=0.25R) FAD	1.40	1.32	1.20	1.13	1.03	1.00	1.00	1.15	1.08	1.00
(R_u=0.75R) FAD	2.58	2.59	1.35	1.35	1.03	1.02	1.00	1.53	1.31	1.04
(R_u=R) FAD	3.29	3.30	3.31	2.58	1.35	1.05	1.00	3.13	2.48	1.22
(MSFE) FAD	1.35	1.35	1.17	1.07	1.01	1.00	1.00	1.12	1.06	1.01
(no log) FAD	3.29	3.30	3.31	3.29	3.30	3.31	2.38	4.04	3.14	1.33

This table reports RMSFE performances relative to the RMSFE of the optimal $\tilde{\delta}_{OLS}$. Reference setup ('Ref'): Data simulated by $y = \delta X_1 + (1 - \delta)X_2 + \epsilon$ where $X \sim \mathcal{N}(0, 4)$, $\epsilon \sim \mathcal{N}(0, 1)$, δ is drawn from $\mathcal{U}(0, 1)$ unless specified otherwise, sample size $N = 30$, and validation window $V = 15$. Maximum number of configurations is $R = 10$. Find $b \in [0, 5]$ when estimating $y = \beta^4 X_1 + (1 - \beta^4)X_2 + \nu$. The column headers show variations of the simulation exercise, and the row headers enlist the different search techniques. The header 'δ=0' signifies that δ is set to zero instead of being randomly drawn. '(No scale) EI' means, for example, that the reference setup of the EI method is altered in that no log transformation is applied to scale the scores. EI uses $\log y$. FAD uses $R_u = 0.5R$ to define ϕ. Computation time (sec.) for reference setups: GRID (.0006), RAND (.0019), EI (.45), FAD (.0023).

6.3.4. *Large Subspace with Dissimilar Forecasts*

In the third simulation exercise, a case will be studied with large forecasting deviances in relatively large subspaces. The data will again be simulated using equation (6.8),

$$y = \delta X_1 + (1 - \delta)X_2 + \epsilon,$$

with $X \sim \mathcal{N}(0, 4)$ and $\epsilon \sim \mathcal{N}(0, 1)$. The difference is that predictions are generated by estimating β in

$$y = \beta^4 x_1 + (1 - \beta^4)x_2 + \nu, \tag{6.10}$$

where $\beta \in [0, 5]$. With β^4, a small change in β can result in a large change in forecasts.

Table 6.4 reports the accuracy of the various search methods. The grid and random searches have much difficulties with this task and the EI search starts to perform well when the number of runs exceeds $R = 20$. FAD is close to an optimal solution after 10 runs already. In the reference setup, the computation time of EI (.45 seconds) is again much longer than that of grid (.0006), rand (.0019), and FAD (.0023).

Focusing on FAD, the results do not improve by altering ϕ. '(MSFE) FAD' makes it clear that excluding a square root in FD and FA worsens the search in these extreme settings. Results are also much poorer for '(no log) FAD', so a log should indeed be included in the FAD measure to deal with forecasts that are highly dissimilar.

The three simulation exercises lead me to reject the hypotheses that grid, random, and EI are efficient search techniques when a single item is optimized over through cross-validation. The newly proposed FAD search has outperformed each of them. In the next section I will evaluate the search techniques for when there are multiple items to be optimized over.

6.4. Empirical Application

Having analyzed the search techniques for a single item, I will now use an empirical case study to examine the how well the search methods perform in a situation with multiple items. In the case study, I will analyze choices to do with combining a number of best-ranked expert forecasts of the US Survey of Professional Forecasters ('SPF'), see Capistrán and Timmermann (2009b). Three of such statistical decisions are the maximum number of best-ranked experts included ($E \in [1, 40]$), the required number of pseudo predictions used for ranking and weighing expert forecasts ($G \in [1, 10]$), and a shrinkage rate which determines to what extent individual weights ($S = 0$) or equal weights ($S = 1$) are assigned to the E forecasts. A fourth item specifies to what degree absolute ($P = 1$) or squared ($P = 2$) prediction errors are used in assessing expert forecasts. E and G are discrete items, S and P are continuous, and some items may be more relevant than others.

To be more concrete, the forecasts of expert i are ranked and weighed on the basis of scores ω_t^i that are derived from his h-step-ahead predictions like so

$$\omega_t^i = \frac{1}{G} \sum_{t=T-h-G}^{T-h-1} |y_t - \hat{y}_{t,h}^i|^P.$$

Note that a prediction about y_t can only be evaluated in the following quarter at y_{t+1}, because we are dealing with a real-time data set. Individual weights are then determined by taking the inverse of these expert scores ω^i and by shrinking them to equal weights,

$$w_t^i = (1 - S)\frac{(\omega_t^i)^{-1}}{\sum_{j=1}^{E} (\omega_t^j)^{-1}} + S\frac{1}{E}. \tag{6.11}$$

The weights assigned to forecasts of the selected top-ranked experts vary less when $S \to 1$ (shrink towards equal weights) and when $P \to 1$ (use absolute rather than squared prediction errors).

Forecasts of configurations that are closer together need not be more similar than those of more distant neighbors. A pooled forecast with 5 top-ranked experts could be more similar to a forecast with 20 rather than 10 top-ranked experts; although on average, it usually won't be.[8] Similarly, when experts are compared based on a track record of 2, the resulting forecasts could be more similar when a track record of 5 is used rather than a track record of 3. It will be interesting to see how the FAD and EI approaches perform under these circumstances.

I will start by looking at h-quarter ahead predictions of the USA Price index of the Gross Domestic Product ('PGDP').[9] Following Capistrán and Timmermann

[8] The more experts that are included in the average, the less the average will change due to the addition of a new expert (law of large numbers).

[9] Data available at: https://www.philadelphiafed.org/research-and-data/real-time-center/survey-of-professional-forecasters/

(2009b), PGDP is transformed by

$$x_{t+h} = 400 \cdot \ln \frac{X_{t+h}}{X_{t+h-1}}. \tag{6.12}$$

Thirteen other predictands will be evaluated later to assess the overall quality of each search technique. The main goal is again to efficiently cross-validate over statistical decisions, so I want to minimize pseudo prediction errors of the validation sample by selecting an optimal set of configurations regarding the four items just mentioned.

Figure 6.6 shows how the search procedures distribute 49 sets of configurations across the number of experts E and the length of the track record G for a validation window of twenty observations. The other items are set to $S = 0$ and $P = 2$ because it is more convenient to present a two dimensional graph than a four dimensional one.

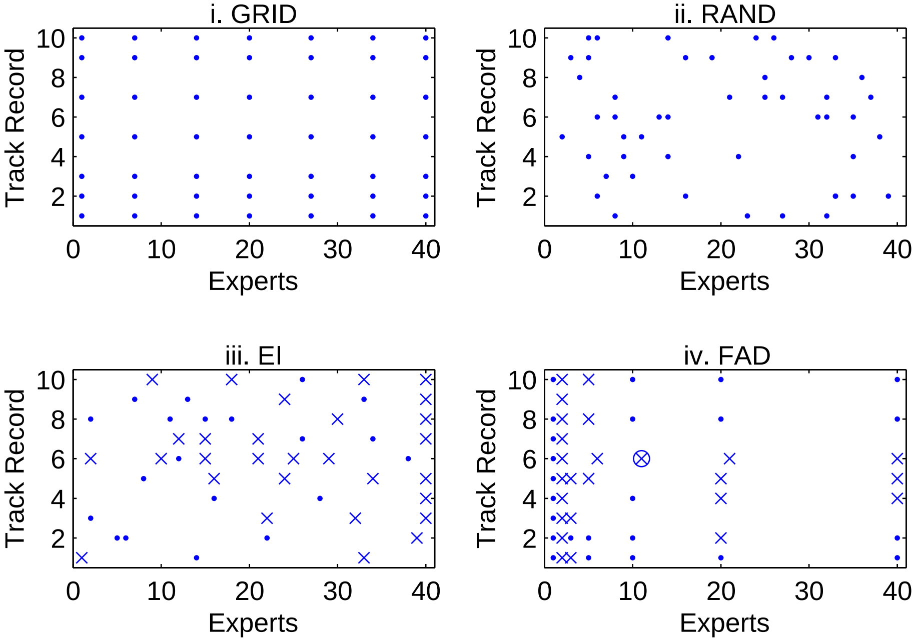
The grid and random searches evaluate numerous less relevant configurations. For example, pooling over $E = 31$ top-ranked experts instead of 35 probably amounts to similar forecasts, whereas the choice between including 1 or 5 experts could result in large differences. So, based on forecasting deviances, it seems more worthwhile to evaluate the average of 3 rather than 33 expert forecasts. Moreover, experts frequently enter and exit the survey, so that many forecasts are excluded when the required track record is too long. The decision on the maximum number of experts might therefore only be relevant conditional on the track record being short. When the EI search is repeated for this example, it often focuses on one good area in the space and continues by adding points along the perimeter. In Figure 6.6.iii I have shown an example where EI gets close to the optimal value ($E = 11, G = 6$).

Rather than equally distributing configurations across the space or adding configurations along the perimeter, an FD approach can be used to efficiently spread sets of configurations. The best forecasts might nevertheless result from combining many eligible experts with long track records. That is why, as more configurations are added, the FAD search gradually focuses on areas with promising FA's. To illustrate step 2, where configurations are selected based on FAD scores, consider the following hypothetical results for experts E and track records G:

$$FAD(y_{E=1,\ \underline{G=1}}, y_{E=40,\ \underline{G=1}}) = \log 0.8, \qquad FAD(y_{\underline{E=1},\ G=1}, y_{\underline{E=1},\ G=10}) = \log 0.6,$$
$$FAD(y_{E=1,\underline{G=10}}, y_{E=40,\underline{G=10}}) = \log 0.4, \qquad FAD(y_{\underline{E=40},G=1}, y_{\underline{E=40},G=10}) = \log 0.2,$$

where $\underline{G = 1}$ is underlined to emphasize that the forecasts of $E = 1$ are compared

Figure 6.6: *Example of Selection Track Records and Experts*



This figure shows the selection of E and G for grid, random, EI, and FAD searches, where $E \in [1, 40]$ is the number of best-ranked experts included, and $G \in [1, 10]$ is the length of the track records on the basis of which the experts are compared. The other items are set at $S = 0$ and $P = 2$. In panel iii, the dots represent the randomly drawn initial configurations and the × represent the selection based on EI. In panel iv. the dots represent FAD configurations when FD was predominant and the × represent configurations when FA was predominant. The optimal set of configurations with $E = 11$ experts and a track record of $G = 6$ is encircled. Data: PGDP, $h = 4$, window [161,180].

to $E = 40$ conditional on $G = 1$. In this example, the difference between using 1 and 40 experts for a track record of 1 leads to the highest FAD, which is why a number of experts will next be added that is half-way between 1 and 40, so $(E = 20, \underline{G = 1})$. When adding a new E conditional on, say, $\underline{G = 3}$, it might be the case that the extremes $(E = 1, \underline{G = 3})$ and/or $(E = 40, \underline{G = 3})$ have not yet been included. As aforesaid, those settings will then be added as well to ensure that this apparently relevant subspace can be properly investigated.

Figure 6.6.iv gives an example of 49 sets of configurations that were selected by the FAD procedure. When the FD search was predominant (dots), few experts with short to intermediate track records were mainly selected. This tendency was continued when FA became more influential (crosses). The optimal configuration $(E = 11, G = 6)$ was only included after 46 runs, which explains why it has not been anchored yet. Performing step 3 means adding $(E = 11, G = 1)$ and $(11, 10)$ at the extremes; and $(E = 10, G = 6)$, $(20, 6)$, $(11, 5)$, and $(11, 7)$ close-by. It would have been more usual for FAD to have selected $(E = 10, G = 5)$ here, but I specifically looked for an example that helps to illustrate the third FAD step. Regarding the lack of tree structure in FAD, it can be remarked that in Figure 6.6.iv, $E = 20$ was added conditional on $G = 1$, and $G = 4$ was added conditional on $E = 20$.

Moving on to evaluate the accuracy of the search methods, Table 6.5 compresses the results for $h = 0, 1, 2, 3$, and 4 quarter ahead predictions of PGDP. Associating observation 1 with 1968Q4 and observation 187 with 2015Q3, I applied the techniques on windows [21, 40], [31, 50], ... [161, 180]. The windows are of size twenty, and I iteratively move them up with 10 observations. Since h-quarter-ahead forecasts with a track record of 10 are only available in real-time one quarter later, I will define the first window as [11+h, 10+h+20]. Finally, to also use the last observations, a window of [168, 187] is included. The RMSFEs reported in Table 6.5 are averages over 85 RMSFE scores.

In the column labeled 'G, E, S' I have set the power P to 2 and tried to find the optimal configurations of the track record, the number of top-ranked experts included, and the shrinkage rate, respectively. The grid is defined by equally distributing 5 configurations for each item and by taking all possible combinations between the items. The other search techniques are also allowed to consider $R = 5^3 = 125$ runs. This chapter is about finding the optimal set of configurations. So, for each search procedure, I have selected the set of configurations with the lowest (pseudo) RMSFE of the validation window. The scores are made relative to a large grid procedure, whereby

Table 6.5: *Combining SPF Forecasts for PGDP*

	G, E, S	G, E, S, P
GRID	1.063 (0.060)	1.083 (0.073)
RAND	1.038 (0.041)	1.050 (0.056)
RAND	1.055 (0.056)	1.064 (0.059)
EI	1.016 (0.028)	1.023 (0.045)
EI	1.009 (0.015)	1.018 (0.044)
(R/2) EI	1.032 (0.047)	1.029 (0.048)
(no trans) EI	1.013 (0.030)	1.018 (0.039)
FAD	1.002 (0.004)	1.002 (0.006)
(R/2) FAD	1.011 (0.018)	1.009 (0.017)
FA	1.076 (0.063)	1.110 (0.084)
FD	1.003 (0.006)	1.004 (0.009)
(no step 3) FAD	1.001 (0.003)	1.004 (0.006)
(round discr. up)FAD	1.002 (0.006)	1.002 (0.004)
(no round cont.) FAD	1.002 (0.004)	1.016 (0.025)
(no log) FAD	1.001 (0.002)	1.002 (0.006)

This table reports the average (and standard deviation) of the 85 RMSFE results of a search technique relative to the RMSFE of a large grid with $G \in [1 : 1 : 10], E \in [1 : 1 : 40],$, $S \in [0 : .1 : 1]$, (and $P = [1 : .1 : 2]$) for the PGDP data. $R = 125$ in the left panel and 256 in the right panel. The windows are of size V=20. Remarks in brackets indicate how a reference setup defined in the main text is altered. '(no step 3) FAD' means that the third FAD step is not executed, for example. For four items the average computation times in seconds are 0.6 (grid), 32.0 (large grid), 7.5 (random), 115.3 (EI), and 2.5 (FAD).

$G \in [1:1:10], E \in [1:1:40]$, and $S \in [0:.1:1]$.[10].

Starting with the benchmark methods, the RMSFE of the grid is on average 6.3% removed from the large grid when the search includes three items, with a standard deviation of .060 among the RMSFSEs. The random search is slightly better than the grid search. I have repeated the random search and EI twice to highlight that results can vary. The accuracy of EI search is quite close to the large grid. '(R/2) EI' indicates that, in case the maximum number of runs R is halved, EI is still better than the grid and random searches. FAD nearly always finds an optimal set of configurations which is why its average RMSFE is near optimal. The quality of this performance is mainly due to FD's efficient global search, as the rows labeled 'FA' and 'FD' indicate. The repercussions of including the third FAD step, of rounding discrete configuration to the lowest integer, or of rounding continuous variables to the second decimal place are negligible here.

In the right column, all four items are included. In that case, the small grid is defined by equally distributing 4 configurations for each item, so that the maximum number of runs becomes $R = 4^4 = 256$. RMSFE scores are now made relative to a large grid which also includes $P = [1:.1:2]$. The large grid contains 48,400 sets of configurations in total. The scores of the grid, random, and EI searches deteriorate as a fourth item is added, while FAD finds near optimal configurations again. '(R/2) FAD' shows that in case the maximum number of runs R is halved, FAD is still better than the benchmark methods. The row labeled '(no step 3) FAD' suggests that when the anchoring step is excluded, the results of FAD become slightly worse. It is interesting to observe that the FAD scores deteriorate when continuous configurations are not rounded off to the second decimal place. The reason is that a small difference in the continuous item P can have a large discontinuous effect on the ranking of experts.

The average computation time in seconds varies greatly between the search techniques; for four items they are 0.6 (grid), 32.0 (large grid), 7.5 (random), 115.3 (EI), and 2.5 (FAD). The ranking of experts need only be determined afresh when the track record (G) or power (P) are altered. The random and EI searches often use unique configurations for each item in each run, so that experts need to be ranked again for each new set of configurations. FAD is far less susceptible to this issue, because it only varies one item a time (conditional on other items).

For PGDP, the 'optimal' choices of the large grid often involves few experts, either high or low shrinkage rates, and all kinds of track records and powers. In

[10] 1:1:10 means from 1 to 10 with increments of 1.

Table 6.6: *Overview Variables*

Name	Description	Transformation	Starting Point
PGDP	Price Index of GDP	logYEAR	1968Q4
NGDP	Nominal GDP	logYEAR	1968Q4
HOUSING	Housing starts	logYEAR	1968Q4
INDPROD	Index of industrial production	logYEAR	1968Q4
RGDP	Real GDP	logYEAR	1968Q4
UNEMP	Civilian unemployment rate	-	1968Q4
CPROF	Corporate profits after tax	logYEAR	1968Q4
CPI	CPI inflation rate	-	1981Q3
RCBI	Real change in private inventories	-	1981Q3
RCONSUM	Real personal consumption expenditures	logYEAR	1981Q3
RSLGOV	Real state and local government consumption & gross investment investment	logYEAR	1981Q3
RFEDGOV	Real federal government consumption & gross investment	logYEAR	1981Q3
RRESINV	Real residential fixed investment	logYEAR	1981Q3
TBILL	Three-month Treasury bill	-	1981Q3

Transformation logYEAR: $x_{t+h} = 400 \cdot \ln X_{t+h}/X_{t+h-1}$. End of sample: 2015Q3

Table 6.7: *Combining SPF Forecasts: 14 Variables and 4 Items*

	$V = 10$	$V = 20$
GRID	1.084 (0.103)	1.052 (0.055)
RAND	1.063 (0.081)	1.037 (0.044)
EI	1.028 (0.059)	1.013 (0.027)
FAD	1.006 (0.022)	1.004 (0.011)

This table reports the average RMSFE results relative to that of the large GRID. The average is taken over 5 horizons, 18 or 12 windows (depending on the starting point of the data set), and fourteen variables. The average computation times in seconds are 0.6 (grid), 28.0 (large grid), 8.7 (random), 97.8 (EI), and 2.5 (FAD) for $V = 20$.

selecting candidate configurations, the FAD procedure mostly focuses on adding new expert numbers and track records. The power is also varied quite often and the shrinkage rate is largely ignored. In the few instances that shrinkage rates are added by FAD, they are mostly introduced conditional on large P. That is because shrinkage rates S are more relevant for high powers, in the sense that $P = 2$ results in more extreme deviances between the weights of expert forecasts than $P = 1$. FAD automatically focuses on such a relevant subset of configurations.

Lastly, I will study the overall performance of grid, random, EI, and FAD for a total of fourteen macroeconomic variables, see Table 6.6.[11] I use the same data and transformations as in Capistrán and Timmermann (2009b). Note that for seven of the regressors, expert forecasts are only available as of observation number 52 (1981Q3). In that case, the first window is defined as [63+h, 62+h+V] and the next windows continue with [71, 70+V], [81, 80+V], . . . , [181-V, 180], [188-V, 187]. To study the effect of the size of the validation window, I will evaluate window sizes of $V = 20$ and $V = 10$.

Table 6.7 shows the RMSFE relative to that of the large grid for each search method. All four items G, E, S, and P are included and I average over the RMSFEs of the different windows for all five horizons and fourteen variables (1050 results in total). The results for $V = 20$ are quite similar to the ones of PGDP and grow worse when the window size is decreased to $V = 10$. For $V = 20$, the average computation times in seconds of the large grid (28.0), the random search (8.7), and EI (97.8) are again quite a bit longer than those of the small grid (0.6) and FAD (2.5).

Overall, this application shows that the FAD search can be a quick procedure for finding optimal configurations of multiple statistical decisions.

6.5. Discussion

In this chapter, it has been investigated how to efficiently select statistical settings based on cross-validation. The standard practice of using grid or random searches was shown to be inefficient, in the sense that many (manually defined) sets of configurations need to be evaluated to find the optimal set of configurations. The more sophisticated Expected Improvement search results in more accurate

[11] Regarding CPROF, experts predicted the corporate profits after tax *without* IVA and CCadj prior to 2006 and *with* IVA and CCAdj from 2006 to present. Prior to 2006, the real-time data set ‘NCPROFAT’ was used, and as of 2006, I used realizations from https://fred.stlouisfed.org/series/CPATAX.

solution than the grid and random searches for a given number runs, but its estimation procedure is more complex and takes a longer time to run.

Instead of equally distributing forecasts across a space, I have suggested to use the average forecasting deviance between neighboring configurations to decide which point to add next. A global to local approach was developed by gradually focusing on forecasting accuracies as more configurations get included. This simple FAD search was shown to be quick and accurate for a variety of challenges.

In future research, I recommend that the global to local approach is applied to other problems so that its quality can be further assessed. A sequential optimization procedure requires that forecasts of nearby configurations are more similar than forecasts of distant configurations. In applying the method, the practitioner may need to formulate the statistical problem such that configurations can be ordered. Rather than comparing mean and median forecasting accuracy measures, for example, I have tried to capture the idea of downweighing extreme forecasting errors by comparing absolute ($P = 1$) to squared errors ($P = 2$).

I will give three more suggestions for further research. First, the initial grid now contains 2^K sets of configurations. To make FAD feasible for more than 10 items, adjustments need to be made. Second, when considering an item c_1 conditional on some choice for $c_2 = \kappa$, it appears wasteful to only use direct neighbors in predicting the deviance and accuracy of the middle configuration while other configurations of c_1 are available conditional on $c_2 = \kappa$ as well. One may take a weighted average of various FA and FD scores, whereby the weights are determined based on the (inverse) distances between the configurations. This will smooth FAD scores, thereby making them less vulnerably to volatile behavior in y as c_1 is varied. Third, a measure of the relative deviance between configurations can be included in the FAD measure to deal with cases where average forecasts often get more similar when the distance between configurations increases.

Finally, I would like to make some remarks regarding the use of cross-validation to evaluate statistical settings, since the FAD procedure appears to promote its use (although FAD can also be applied when employing information criteria). In many cases, cross-validation need not be applied if the tuning parameter is defined more intuitively. A FAD tuning parameter $\phi \in [0, 1]$ was for instance specified in terms of $\frac{R_u}{R_u + r}$ to make a gradual transition from forecasting deviance towards forecasting accuracy. Setting $R_u = \frac{R}{2}$ has appeared to result in a stable performance.

A related point is that adding more complexity to an algorithm may worsen out-of-sample results. Instead of employing a loss function that merely focuses on in-sample accuracy, I recommend that a researcher defines a loss function which makes an intuitive tradeoff between relative accuracy and relative simplicity. Such tradeoffs were examined for linear regression models in Chapter 2, 3 and 4 and for weighing observation in Chapter 5. ASTs help to make optimization problems more convex and thereby less random. Cross-validation is an excellent technique in dealing with statistical decisions, particularly when an efficient search procedure like FAD is employed to quickly find the optimal set of configurations.

Samenvatting (Dutch)

In het cumulatieve proces van wetenschap wordt kennis over de onderliggende waarheid voortdurend bijgesteld door hypothesen te testen met nieuwe empirische data. Huidige statistische benaderingen maken het moeilijk voor onderzoekers om een goede balans te vinden tussen de *eenvoud* van het onveranderd laten van de hypothesen en de *accuraatheid* van data-geoptimaliseerde waarden. Een overkoepelend doel van deze dissertatie is om een dergelijke afweging tussen eenvoud en accuraatheid intuïtief te formuleren, zodat een onderzoeker beter kan anticiperen en beïnvloeden hoe modellen geschat worden.

Het boek is geschreven vanuit een econometrisch perspectief, gezien deze discipline er in het bijzonder op toegerust is om op basis van data analyses uitspraken te doen over de onderliggende waarheid. Naast de algemene inleiding in hoofdstuk 1 bestaat dit proefschrift uit vijf hoofdstukken.

Hoofdstukken 2, 3 en 4 gaan over het lineaire regressiemodel. Dit werkpaard van de econometrie stelt dat een afhankelijke variabele y lineair gerelateerd is aan een onafhankelijke variabele x via parameters α en β en een residu ϵ. Observaties $n = 1, 2, \ldots, N$ worden aldus als volgt gegenereerd:

$$y_n = \alpha + \beta x_n + \epsilon_n.$$

De ware en onveranderlijke parameters α en β zijn onbekend. Om die te kunnen achterhalen, stelt een onderzoeker allereerst hypothesen op, zoals $\alpha_0 = 2$ en $\beta_0 = 1$. Deze verwachtingen kunnen vervolgens aangepast worden door een willekeurige steekproef te analyseren.

Verscheidene statistische methoden maken gebruik van een penaltieparameter λ om te bepalen hoeveel de uiteindelijk geschatte waarden mogen afwijken van de hypothesen. Het is daarbij pas achteraf mogelijk om te bepalen hoeveel invloed een bepaalde waarde van λ toekent aan een hypothese. Bij de Bayesiaanse benadering kan nog geprobeerd worden om variabelen op een handige manier

te schalen, maar een dergelijke schalingsprocedure wordt doorgaans vermeden omdat het (extreem) tijdrovend kan zijn. In de nieuwe schattingsmethode die ik voorstel geeft $\lambda \cdot 100\%$ in percentages aan wat de minimale invloed van de hypothesen is. Alleen relevante afwijkingen van de hypothesen worden toegestaan. De schattingsmethode berekent de relevantie van iedere onafhankelijke variabele impliciet aan de hand van diens bijdrage aan een welbekende maat voor de accuraatheid van het model (R^2). Terzijde laat ik zien hoe deze individuele bijdragen expliciet gemeten kunnen worden.

Een tweede manier om de eenvoud van modellen te bevorderen, is om dezelfde waarde toe te kennen aan verschillende parameters. Indien de ene variabele gemiddeld genomen hoog (of laag) is als een andere variabele dat ook is, dan spreekt men van een hoge positieve correlatie tussen die variabelen. Bij een negatieve correlatie bewegen de variabelen gemiddeld juist in tegengestelde richting. De nieuwe schattingsmethode zorgt ervoor dat parameters van sterk positief of negatief gecorreleerde variabelen gegroepeerd kunnen worden, wat betekent dat ze gestimuleerd worden om dezelfde afwijking van de hypothesen te krijgen. De schattingsmethode stelt de onderzoeker in staat om van te voren aan te geven wat een 'sterke' correlatie is. Voorheen werd er geen onderscheid gemaakt tussen zwakke en sterke correlaties tussen variabelen, waardoor irrelevante aanpassingen van de hypothesen bevorderd werden.

In hoofdstuk 2 introduceer ik een schattingsmethode die een directe controle geeft over de bovenstaande twee manieren om een model eenvoudiger te maken. De eenvoud van een model wordt hier gekwantificeerd door de gekwadrateerde afstand te berekenen tussen de hypothesen en de data-geoptimaliseerde waarden. Deze schatter heeft een exacte oplossing, staat irrelevante wijzigingen van hypothesen (bij benadering) niet toe en groepeert parameters van sterk gecorreleerde variabelen.

In hoofdstuk 3 wordt een absolute in plaats van een kwadrateerde afstand gebruikt voor het meten van de eenvoud van een model. De resulterende schatter heeft een bijzondere eigenschap die helpt bij het bepalen van welke variabelen relevant zijn voor het schatten van y. Zelfs al voordat λ de maximale waarde bereikt heeft, zullen parameters exact gelijkgesteld worden aan de gehypothetiseerde waarden. Het moment waarop dat gebeurt was voorheen onduidelijk en ik laat zien dat dit direct samenhangt met de bijdrage die een onafhankelijke variabele levert aan R^2. Zo kan de onderzoeker dus goed anticiperen en beïnvloeden hoe de afweging tussen hypothesen en variabelen gemaakt worden. Ook deze schatter zal op een effectieve manier parameters van sterk gecorreleerde variabelen groeperen.

In hoofdstuk 4 onderzoek ik methoden die de onderzoeker helpen in het bepalen van een geschikte waarde van de penaltieparameter λ. Wederom laat ik zien hoe op een praktische manier de afweging gemaakt kan worden tussen een waarde die door de onderzoeker opgegeven is en een waarde die door data-optimalisatie tot stand is gekomen. Daarnaast bespreek ik hoe λ gekozen kan worden met een informatiecriterium. Zo'n criterium maakt een afweging tussen de accuraatheid van het model en het effectieve aantal parameters in het model aan de hand van het aantal observaties in de dataset.

Op het moment is er geen onomstreden manier om het effectieve aantal parameters van een model te berekenen. In de literatuur is over het hoofd gezien dat het aantal parameters van een model vermindert naarmate dezelfde waarde wordt toegekend aan meerdere parameters. Methoden zoals de F-toets hebben daardoor de neiging om variabelen uit te sluiten als ze al hoog correleren met andere variabelen in het model. Als gevolg hiervan kunnen variabelen genegeerd worden die wel relevant zijn in het onderliggende data-genererende proces. Bovendien zullen de risico's tussen hoog gecorreleerde variabelen niet gespreid worden, waardoor de voorspelkracht van het model kan afnemen. Om dit probleem op te lossen zal ik laten zien dat de term die door de eerder geïntroduceerde schatters gebruikt werd om de eenvoud van het model te bepalen, direct toegepast kan worden om het effectieve aantal parameters te meten. Simulatie studies tonen aan dat het voorspelvermogen en de interpretatie van modellen verbeteren door de methoden ik ontwikkeld heb om lineaire regressie- en penaltieparameters te schatten.

Waar in de voorgaande hoofdstukken is aangenomen dat het data genererende proces onveranderlijk is, onderzoek ik in hoofdstuk 5 hoe parameters geschat kunnen worden als er breuken zijn in het onderliggende proces. De werkelijke α kan na 50 observaties bijvoorbeeld veranderen van $\alpha = 3$ naar $\alpha = 5$. Een veelgebruikte benadering is om allereerst het moment en de grootte van de breuk te schatten en om vervolgens de data na de breuk te gebruiken voor het schatten van modelparameters. De beste startpunt methode selecteert de optimale breuk door te bepalen welk startpunt de grootste accuraatheid heeft in het voorspellen van de meest recente observaties die beschikbaar zijn. Dit wordt ook wel cross-validatie genoemd. De drie voornaamste tekortkomingen van deze methode zijn dat het te traag is in het reageren op een nieuwe breuk, dat het te snel is in het negeren van observaties uit het verleden en dat het enkel observaties na de breuk in ogenschouw neemt.

De reactiesnelheid op een nieuwe breuk kan worden verkort door gewichten aan observaties te geven die groter worden naarmate de observaties meer recent worden. Om ervoor te zorgen dat de invloed van data alleen afneemt als daar voldoende aanleiding toe is, formuleer ik een afweging tussen de eenvoud van alle data een gelijk gewicht te geven en de accuraatheid van cross-validatie. Om het derde punt te adresseren wordt er een breed toepasbare methode geïntroduceerd waarmee meerdere periodes in het verleden hun eigen gewicht toegekend krijgen. De voorgestelde aanpassingen aan de beste startpunt methode worden verwerkt in een algoritme dat aan de hand van simulatiestudies en een empirische toepassingen geëvalueerd wordt. Het eerste en het derde punt bouwen voort op het werk van Pesaran, Pick en Pranovich (2013).

Tenslotte wordt in hoofdstuk 6 besproken hoe op een eenvoudige, snelle, en accurate wijze de waarde van een statistische keuze door data-optimalisatie gevonden kan worden. Denk aan de keuze van de penaltieparameter λ of van het startpunt van de dataset. Om hierover te optimaliseren wordt een set van kandidaatwaarden opgesteld. De beste waarde kan hier bijvoorbeeld geselecteerd worden door vast te stellen wanneer één deel van de data het beste 'voorspeld' kan worden aan de hand van een andere deel van de data (cross-validatie). Meestal worden de kandidaatwaarden gelijk verdeeld over de dimensies, zoals $\lambda = 0, 0.01, 0.02, \ldots, 0.99, 1$. Het gevolg is dat gebieden van configuraties waarbij de resulterende voorspellingen nauwelijks verschillen net zo nauw onderzocht worden als gebieden waarbij de voorspellingen sterk verschillend zijn. Bovendien wordt er geen speciale aandacht gegeven aan gebieden die betere voorspellingen genereren dan andere. Wel bestaat er een meer geavanceerde techniek die configuraties selecteert op basis van de verwachtte vooruitgang in accuraatheid die het oplevert, maar het duurt erg lang voordat computers de benodigde berekeningen hiervoor gemaakt hebben.

De methode die ik voorstel voegt eerst kandidaatconfiguraties toe in gebieden waar de voorspellingen het sterkst verschillen. Naarmate meer configuraties op deze manier globaal verdeeld worden, zal de accuraatheid van de voorspelling steeds meer invloed krijgen in het selecteren van nieuwe kandidaatwaarden. De voornaamste veronderstelling hierbij is dat voorspellingen van aangrenzende configuraties die reeds geëavalueerd zijn, zeg $\lambda = 0$ en $\lambda = 0.25$, meer op elkaar lijken dan de voorspellingen van configuraties die verder uit elkaar liggen, zoals $\lambda = 0$ en $\lambda = 1$. De methode is eenvoudig en kan worden toegepast om meerdere statistische beslissingen tegelijk te onderzoeken. Simulatie studies en een empirische applicatie laten veelbelovende resultaten zien. In plaats van 101 waarden van λ te evalueren, zijn er bijvoorbeeld slechts 10 nodig om de juiste waarde te selecteren.

Het schatten van regressie- en penaltieparameters, het wegen van observaties en het efficiënt selecteren van configuraties zijn de voornaamste toepassingen waar deze dissertatie over balansoefeningen in de econometrie over gaat. Er bestaan verschillende opvattingen over de manier waarop er uitspraken gedaan kunnen worden over de onderliggende waarheid. Toch is er algemene overeenstemming over de globale stappen die genomen dienen te worden bij het doen van onderzoek. De wetenschapper begint met het opstellen van een onderzoeksvraag, stelt op basis van eerdere kennis hypothesen op en specificeert de methoden waarmee die hypothesen onderzocht gaan worden. Vervolgens verzamelt hij willekeurig geselecteerde data en past hij de methoden op de data toe om de hoofdhypothesen te evalueren terwijl de overige aannamen onveranderd blijven. Tenslotte geeft de onderzoeker aan welke conclusies er wel en niet getrokken kunnen worden en bespreekt hij hoe eventuele tekortkomingen van de studie in de toekomst verholpen kunnen worden.

De bovenstaande procedure staat bekend als de wetenschappelijke methode. Deze opvatting van wetenschap is niet zonder problemen en die zal ik verder onderzoeken in het binnenkort te verschijnen boek *Science: Under Submission*. De hoofdstukken van de huidige dissertatie zijn geschreven volgens de geldende normen van de wetenschap. Met behulp van de statistische benadering die hier gepresenteerd wordt zal het eenvoudiger worden voor onderzoekers om vooraf aan te geven in hoeverre ze bereid zijn om hun hypothesen over de onderliggende waarheid bij te stellen aan de hand van de nog onbekende resultaten van een nieuwe dataset.

Bibliography

Andrews, D. W. (1993). Tests for parameter instability and structural change with unknown change point. *Econometrica: Journal of the Econometric Society*, 61(4):821–856.

Bai, J. and Perron, P. (1998). Estimating and testing linear models with multiple structural changes. *Econometrica*, 66(1):47–78.

Bai, J. and Perron, P. (2003). Computation and analysis of multiple structural change models. *Journal of Applied Econometrics*, 18(1):1–22.

Bartz-Beielstein, T., Lasarczyk, C. W., and Preuß, M. (2005). Sequential parameter optimization. In *IEEE Congress on Evolutionary Computation*, volume 1, pages 773–780. Institute of Electrical and Electronics Engineers.

Bates, J. M. and Granger, C. W. (1969). The combination of forecasts. *Journal of the Operations Research Society*, 20(4):451–468.

Bergstra, J. and Bengio, Y. (2012). Random search for hyper-parameter optimization. *The Journal of Machine Learning Research*, 13(1):281–305.

Boyd, S. and Vandenberghe, L. (2009 [2004]). *Convex Optimization.* Cambridge University Press.

Brown, R. L., Durbin, J., and Evans, J. M. (1975). Techniques for testing the constancy of regression relationships over time. *Journal of the Royal Statistical Society. Series B (Methodological)*, 37(2):149–192.

Capistrán, C. and Timmermann, A. (2009a). Disagreement and biases in inflation expectations. *Journal of Money, Credit and Banking*, 41(2-3):365–396.

Capistrán, C. and Timmermann, A. (2009b). Forecast combination with entry and exit of experts. *Journal of Business & Economic Statistics*, 27(4):428–440.

Chamberlain, G. and Leamer, E. E. (1976). Matrix weighted averages and posterior bounds. *Journal of the Royal Statistical Society. Series B (Methodological)*, 38:73–84.

Chow, G. C. (1960). Tests of equality between sets of coefficients in two linear regressions. *Econometrica: Journal of the Econometric Society*, 28(3):591–605.

Croushore, D. (2008). An evaluation of inflation forecasts from surveys using real-time data. *Federal Reserve Bank of Philadelphia.*

Efron, B. (1986). How biased is the apparent error rate of a prediction rule? *Journal of the American Statistical Association*, 81(394):461–470.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., et al. (2004). Least angle regression. *The Annals of Statistics*, 32(2):407–499.

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.

Friedman, J., Hastie, T., Höfling, H., Tibshirani, R., et al. (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–332.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2014). *Bayesian Data Analysis*, volume 2. Chapman & Hall/CRC Boca Raton, FL, USA.

Hastie, T., Tibshirani, R., and Friedman, J. (2013 [2009]). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer.

Heij, C., de Boer, P., Franses, P., Kloek, T., and van Dijk, H. K. (2004). *Econometric Methods with Applications in Business and Economics.* Oxford University Press.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 12:55–67.

Hutter, F., Hoos, H. H., and Leyton-Brown, K. (2011). Sequential model-based optimization for general algorithm configuration. In *Learning and Intelligent Optimization*, pages 507–523. Springer.

Hutter, F., Hoos, H. H., Leyton-Brown, K., and Murphy, K. P. (2009). An experimental investigation of model-based parameter optimisation: SPO and beyond. In *Proceedings of the 11th Annual Conference on Genetic and Evolutionary Computation*, pages 271–278. Association for Computing Machinery.

Janson, L., Fithian, W., and Hastie, T. J. (2015). Effective degrees of freedom: a flawed metaphor. *Biometrika*, 102(2):479–485.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering sciences*, 186(1007):453–461.

Jones, D. R., Schonlau, M., and Welch, W. J. (1998). Efficient global optimization of expensive black-box functions. *Journal of Global optimization*, 13(4):455–492.

Kaufman, S. and Rosset, S. (2014). When does more regularization imply fewer degrees of freedom? Sufficient conditions and counterexamples. *Biometrika*, 101(4):771–784.

Kish, L. ([1995] 1965). *Survey Sampling.* John Wiley & Sons.

Kreyszig, E. (1999). *Advanced Engineering Mathematics, 8th Edition.* John Wiley & Sons.

Leamer, E. E. (1981). Coordinate-free Ridge regression bounds. *Journal of the American Statistical Association*, 76(376):842–849.

Liu, J., Wu, S., and Zidek, J. V. (1997). On segmented multivariate regression. *Statistica Sinica*, 7(2):497–525.

Lophaven, S. N., Nielsen, H. B., and Søndergaard, J. (2002). DACE: A Matlab Kriging toolbox, version 2.0. Technical report.

Mallows, C. L. (1973). Some comments on c_p. *Technometrics*, 15(4):661–675.

Marquardt, D. W. and Snee, R. D. (1975). Ridge regression in practice. *The American Statistician*, 29(1):3–20.

Osborne, M. R., Presnell, B., and Turlach, B. A. (2000). A new approach to variable selection in least squares problems. *IMA Journal of Numerical Analysis*, 20(3):389–403.

Pan, C.-C. (2012). Significance of prostatic capsular status in radical prostatectomy. *Urological Science*, 23(1):15–17.

Pesaran, M. H., Pick, A., and Pranovich, M. (2013). Optimal forecasts in the presence of structural breaks. *Journal of Econometrics*, 177:134–152.

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*, 137:134–161.

Potter, S. R., Epstein, J. I., and Partin, A. W. (2000). Seminal vesicle invasion by prostate cancer: prognostic significance and therapeutic implications. *Reviews in Urology*, 2(3):190–195.

Quandt, R. E. (1960). Tests of the hypothesis that a linear regression system obeys two separate regimes. *Journal of the American Statistical Association*, 55(290):324–330.

Sacks, J., Welch, W. J., Mitchell, T. J., and Wynn, H. P. (1989). Design and analysis of computer experiments. *Statistical Science*, 4(4):409–423.

Schlaifer, R. and Raiffa, H. (1961). *Applied Statistical Decision Theory.* Harvard University.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

Smith, G. and Campbell, F. (1980). A critique of some Ridge regression methods. *Journal of the American Statistical Association*, 75(369):74–81.

Smith, J. and Wallis, K. F. (2009). A simple explanation of the forecast combination puzzle. *Oxford Bulletin of Economics and Statistics*, 71(3):331–355.

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9(6):1135–1151.

Swindel, B. F. (1976). Good Ridge estimators based on prior information. *Communications in Statistics-Theory and Methods*, 5(11):1065–1075.

Theil, H. (1961). *Economic Forecasts and Policy.* North-Holland Publishing Company.

Theil, H. (1971). *Principles of Econometrics.* Wiley.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the Fused Lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108.

Tibshirani, R. J. (2011). *The Solution Path of the Generalized Lasso*. Stanford University.

Tibshirani, R. J. (2015). Degrees of freedom and model search. *Statistica Sinica*, 25(3):1265–1296.

Tseng, P. (1988). Coordinate ascent for maximizing nondifferentiable concave functions. Technical report.

Wang, S., Nan, B., Rosset, S., and Zhu, J. (2011). Random Lasso. *The Annals of Applied Statistics*, 5(1):468.

Yang, X. (2000). A matrix trace inequality. *Journal of Mathematical Analysis and Applications*, 250(1):372–374.

Ye, J. (1998). On measuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association*, 93(441):120–131.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno De Finetti*, 6:233–243.

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.

Zou, H., Hastie, T., and Tibshirani, R. (2007). On the ‘degrees of freedom’ of the Lasso. *The Annals of Statistics*, 35(5):2173–2192.

Tinbergen Publications List

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:

667 H. SCHMITTDIEL, Paid to Quit, Cheat, and Confess

668 A. DIMITROPOULOS, Low Emission Vehicles: Consumer Demand and Fiscal Policy

669 G.H. VAN HEUVELEN, Export Prices, Trade Dynamics and Economic Development

670 A. RUSECKAITE, New Flexible Models and Design Construction Algorithms for Mixtures and Binary Dependent Variables

671 Y. LIU, Time-varying Correlation and Common Structures in Volatility

672 S. HE, Cooperation, Coordination and Competition: Theory and Experiment

673 C.G.F. VAN DER KWAAK, The Macroeconomics of Banking

674 D.H.J. CHEN, Essays on Collective Funded Pension Schemes

675 F.J.T. SNIEKERS, On the Functioning of Markets with Frictions

676 F. GOMEZ MARTINEZ, Essays in Experimental Industrial Organization: How Information and Communication affect Market Outcomes

677 J.A. ATTEY, Causes and Macroeconomic Consequences of Time Variations in Wage Indexation

678 T. BOOT, Macroeconomic Forecasting under Regime Switching, Structural Breaks and High-dimensional Data

679 I. TIKOUDIS, Urban Second-best Road Pricing: Spatial General Equilibrium Perspectives

680 F.A. FELSÅŘ, Empirical Studies of Consumer and Government Purchase Decisions

681 Y. GAO, Stability and Adaptivity: Preferences over Time and under Risk

682 M.J. ZAMOJSKI, Panta Rhei, Measurement and Discovery of Change in Financial Markets

683 P.R. DENDERSKI, Essays on Information and Heterogeneity in Macroeconomics

684 U. TURMUNKH, Ambiguity in Social Dilemmas

685 U. KESKIN, Essays on Decision Making: Intertemporal Choice and Uncertainty

686 M. LAMMERS, Financial Incentives and Job Choice

687 Z. ZHANG, Topics in Forecasting Macroeconomic Time Series

688 X. XIAO, Options and Higher Order Risk Premiums

689 D.C. SMERDON, âĂŸEverybodyâĂŹs doing itâĂŹ: Essays on Trust, Norms and Integration

690 S. SINGH, Three Essays on the Insurance of Income Risk and Monetary Policy

691 E. SILDE, The Econometrics of Financial Comovement

692 G. DE OLIVEIRA, Coercion and Integration

693 S. CHAN, Wake Me up before you CoCo: Implications of Contingent Convertible Capital for Financial Regulation

694 P. GAL, Essays on the role of frictions for firms, sectors and the macroeconomy

695 Z. FAN, Essays on International Portfolio Choice and Asset Pricing under Financial Contagion

696 H. ZHANG, Dealing with Health and Health Care System Challenges in China: Assessing Health Determinants and Health Care Reforms

697 M. VAN LENT, Essays on Intrinsic Motivation of Students and Workers

698 R.W. POLDERMANS, Accuracy of Method of Moments Based Inference

699 J.E. LUSTENHOUWER, Monetary and Fiscal Policy under Bounded Rationality and Heterogeneous Expectations

700 W. HUANG, Trading and Clearing in Modern Times

701 N. DE GROOT, Evaluating Labor Market Policy in the Netherlands

702 R.E.F. VAN MAURIK, The Economics of Pension Reforms

703 I. AYDOGAN, Decisions from Experience and from Description: Beliefs and Probability Weighting

704 T.B. CHILD, Political Economy of Development, Conflict, and Business Networks

705 O. HERLEM, Three Stories on Influence

706 J.D. ZHENG, Social Identity and Social Preferences: An Empirical Exploration

707 B.A. LOERAKKER, On the Role of Bonding, Emotional Leadership, and Partner Choice in Games of Cooperation and Conflict

708 L. ZIEGLER, Social Networks, Marital Sorting and Job Matching. Three Essays in Labor Economics

709 M.O. HOYER, Social Preferences and Emotions in Repeated Interactions

710 N. GHEBRIHIWET, Multinational Firms, Technology Transfer, and FDI Policy

711 H.FANG, Multivariate Density Forecast Evaluation and Nonparametric Granger Causality Testing

712 Y. KANTOR, Urban Form and the Labor Market

713 R.M. TEULINGS, Untangling Gravity

714 K.J.VAN WILGENBURG, Beliefs, Preferences and Health Insurance Behavior

715 L. SWART, Less Now or More Later? Essays on the Measurement of Time Preferences in Economic Experiments

716 D. NIBBERING, The Gains from Dimensionality