Of Needles and Haystacks:

Novel Techniques for Data-Rich Economic Forecasting

ISBN: 978 90 3610 267 4

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

This book is no. 516 of the Tinbergen Institute Research Series, established through cooperation between Thela Thesis and the Tinbergen Institute. A list of books which already appeared in the series can be found in the back.

Of Needles and Haystacks:
Novel Techniques for Data-Rich Economic Forecasting

Over naalden en hooibergen:

Nieuwe technieken voor economisch voorspellen op basis van zeer veel gegevens

Proefschrift

ter verkrijging van de graad van doctor aan de
Erasmus Universiteit Rotterdam
op gezag van de
rector magnificus

Prof.dr. H.G. Schmidt

en volgens besluit van het College voor Promoties.

De openbare verdediging zal plaatsvinden op
dinsdag 13 december 2011 om 11:30 uur

door

Peter Exterkate
geboren te Naarden.

Promotiecommissie

Promotoren:	Prof.dr. P.J.F. Groenen Prof.dr. D.J.C. van Dijk
Overige leden:	Prof.dr. C. Croux Dr. D. Fok Prof.dr. S.J. Koopman
Copromotor:	Dr. C. Heij

Acknowledgements

This book is the result of over three years of hard work. Although it only bears my name, there can be no doubt that several others have been of invaluable help during this process. I would like to take this opportunity to acknowledge their support.

First, I thank my supervisors, Dick van Dijk, Patrick Groenen, and Christiaan Heij. All three of them have done an outstanding job in getting me started in the academic world, in co-authoring Chapters 2 and 4 of this thesis, in giving detailed and constructive criticism on the other chapters, and in stimulating me to present my work at no fewer than nine international conferences over these three years.

Dick is a busy person, as everyone who knows him will surely agree. Nevertheless, he always manages to find time for his ten (!) Ph.D. students, even if this means working long hours after midnight. He has a great knowledge of, and a major impact on, current developments in financial and macro-economic forecasting. In addition to his valuable research ideas, methodological suggestions, and contributions to the academic writing style of this book, I owe Dick many thanks for introducing me to Niels Haldrup, leading to an exciting new step in my career.

I have been lucky to have both Dick and Patrick around. Patrick has made similarly useful contributions in terms of ideas, methods, and writing style, but from a very different perspective: his great familiarity with many branches of multivariate statistics has proven very valuable to me over these three years. Moreover, his emphasis on clear formulations, readable tables, and illustrative figures has definitely enhanced the readability of this thesis. I finally thank Patrick for introducing me to Christophe Croux.

The very fact that this book exists is in no small part due to Christiaan, my co-supervisor. During my years as a B.Sc. and M.Phil. student, he managed to spark my interest in pursuing a career in academia. In 2008, he agreed that I would become his first Ph.D. student since 1995. Our weekly discussions have been entertaining and fruitful, particularly because of his emphasis on mathematical rigor, his determination to understand the deepest details of methods, and his precautions against reading too much into results. In addition, if it had not been for Christiaan's continuous reminders of timing, I would probably not have finished writing this book in anything near three years' time.

I thank Christophe Croux for his kind agreement to host me in Leuven for four months, and for co-authoring Chapter 3 of this thesis. This visit, and the discussions that we had, have broadened my view of the econometrics and statistics profession. My stay in Leuven was a crash course in robust statistics and in doing academic research extremely quickly, two skills that will be of great help in my further career. Moreover, it was fun.

Many thanks go to Siem Jan Koopman, Dennis Fok, and again Christophe Croux for their willingness to comment on this manuscript. The input of these experts in different branches of statistics and econometrics has been very valuable.

During my last year as a Ph.D. student, I have taken some time to think about where my career should take me next. My decision has finally been to apply for a position at CREATES, at Aarhus University. Without having seen more than a brief CV, Niels Haldrup, the director of this institute, has been willing to invest a considerable amount of time and effort into helping me to get there. I thank Niels for his confidence and support, and I look forward to a fruitful collaboration with him and his team.

I also thank all the people at Tinbergen Institute and Econometric Institute. I have always found someone willing to engage in entertaining discussions, about all sorts of subjects, and at all imaginable levels of seriousness. Special thanks go to the administrative staff for putting up with my apparent inability to follow bureaucratic guidelines, and to Oleg, Steffi, and Rui for putting up with my apparent inability to stick to regular working hours while sharing an office.

A special word of thanks goes to my parents, brother, sisters, and all the other members of my family. Although not all of them understand exactly what I am doing (and why all of it is taking so long), they have always supported my decision to pursue this career, stimulated me to make the most out of myself, and understood if I told them I had to spend yet another weekend behind my desk.

My friends outside the Tinbergen Institute have provided the right distractions when I was working too hard, and the right encouragements when things were not going well. I thank them for the fun that we have had and the support that they have been giving over the last eight years, in Rotterdam, Porto, Marsala, or Ploče, while walking or swimming, over beer or *poffertjes*.

Finally, I am extremely happy that my friend Evelien and my sister Anneke have agreed to be my paranymphs at the defence ceremony of this thesis. Thanks for sharing this special moment with me, and I hope to return the favor one day.

Peter Exterkate
Rotterdam, September 2011

Contents

Chapter 1

Introduction and Outline

1.1 Introduction

From an economist's perspective, we live in fascinating times. After the general optimism that characterized the global economy during much of the 1990s and 2000s, many countries around the globe are currently experiencing the deepest recession since the 1930s, and there are no clear indications of how and when their economies will recover. Short-term interest rates have reached unprecedentedly low levels, sovereign defaults are no longer just a theoretical possibility that can safely be ignored, and stock markets seem more volatile than ever.

From our perspective as econometricians, the natural question to ask is, why did none of us see any of this coming? However, after more than fifty years of econometric research, the track record of economic forecasting calls for some modesty. I therefore ask the perhaps less ambitious but arguably more relevant question, what can we do to enhance the quality of our economic forecasts in the future? The contribution of this thesis is to provide new approaches to answering this last question. More specifically, I study novel methods to take advantage of the "haystack" of information (technically, and more prosaically, the "data-rich environment") that is at the disposal of a modern forecaster.

Econometricians from a few decades ago would be envious of the wealth of data that is available to forecasters nowadays. The focus of econometrics has traditionally been on how to squeeze as much information as possible out of the small data sets that were available. After the huge advances in computational power and in storage capacities in more recent times, it is now possible to obtain and use very large data sets, at low costs and with small time investments.

While it is obvious that a large data set contains more (or at least, not less) information than any subset of it, it is generally also more difficult to extract the useful information; a problem that is known as the *curse of dimensionality*. Searching through bigger and bigger haystacks may increase the probability that there is a needle somewhere, but it will not make it much easier to find one. Thus, a need for systematic methods of using large data sets for forecasting purposes arose.

Seminal contributions in this field of data-rich economic forecasting were made by Forni et al. (2000) and Stock and Watson (2002). Forni et al. (2000) derived a set of theoretical results for estimating the factors that are common to a large panel of time series, and they apply their tools to summarize a set of macroeconomic variables from individual European countries into a single European economic indicator. Stock and Watson (2002) applied principal component

analysis, a classical statistical tool, to summarize a large panel of macroeconomic variables into a small number (three or four) of representative factors. These factors are then used to forecast future developments in key macroeconomic indicators (production, income, sales, employment, inflation), and the resulting predictions were much more accurate than those obtained without making use of this data-rich environment. Their work has sparked a rich literature, to which Chapters 2 and 3 of this thesis contribute. It is argued that principal component analysis has some important shortcomings: it constructs factors without taking into account (a) what they will be used to forecast, (b) what these factors could actually mean, and (c) that economic time series may contain large outliers, which have an unduly large impact on the estimated factors, hindering their predictive ability. Chapter 2 examines various existing alternatives to principal component analysis that take point (a) into account, while in Chapter 3 a new alternative to handle points (b) and (c) is proposed.

Another tradition that the econometric profession has inherited from the era of limited computational capacities is its focus on linear models. Traditionally, these models were the method of choice because they are easy to estimate and easy to forecast with. Moreover, their results are easy to interpret: "if interest rates rise by 1%, we expect stock indices to rise by $x\%$". However, such models may be an oversimplification of reality: should we really expect that this number x is the same regardless of the current level of interest rates? Of stock prices? Of the variability of stock prices? And, is this the same x as ten years ago? To accommodate such features, nonlinear modelling is required.

Several relatively simple nonlinear models have received attention from econometricians during the last two decades. Among the most popular ones are regime-switching models (see Teräsvirta, 2006) and neural networks (see White, 2006). However, such methods are often found to improve the accuracy of forecasts only marginally. What is more, they are only suitable in situations with a small number of predictors, as their number of parameters grows rapidly with the number of variables. Chapters 4 and 5 of this thesis are devoted to an alternative technique from the machine learning literature, called *kernel ridge regression*, that can estimate flexible nonlinear relations in the presence of many predictor variables. This method is extended to allow for time-series forecasting in Chapter 4, and it is applied as an alternative to the already mentioned Stock and Watson (2002) framework. In Chapter 5, I derive some theoretical results concerning model selection in kernel ridge regression, which lead to guidelines for applying this method to practical problems.

1.2 Outline

This thesis consists of four self-contained chapters on novel techniques for economic forecasting. In all chapters, the focus is on methods that exploit the information in large data sets effectively. Each of these methods is compared to established techniques and in general, improvements in forecast quality are obtained, with reductions in mean squared prediction errors up to 30%. The methods that are introduced in this thesis are used to forecast the yields on U.S. Treasury Bills and Bonds in Chapter 2, housing prices in Chapter 3, and real macroeconomic aggregates measuring production, income, sales, and employment in Chapters 3 and 4. Chapters 3, 4 and 5 also contain simulation studies to investigate the use of these new tools in various contexts.

Broadly speaking, Chapters 2 and 3 deal with linear models, while nonlinear models are discussed in Chapters 4 and 5. In the universe of linear methods for data-rich environments, the state-of-the-art is to use simple, classical principal component analysis. Several shortcomings of this method are pointed out, and we advocate the use of more advanced techniques to overcome them. In Chapter 2, it is shown that the recent revival in the econometric literature of the partial least squares technique (Wold, 1966) is well-deserved. Moreover, Chapter 3 proposes a new factor construction method that leads to robust and interpretable factors, which is shown to perform favorably also when used for forecasting.

In Chapter 2, which is based on Exterkate, van Dijk, Heij, and Groenen (2011b), we compare various ways of extracting macroeconomic information from a data-rich environment for forecasting the yield curve using the Nelson-Siegel model. Five issues in extracting factors from a large panel of macro variables are addressed, namely, selection of a subset of the available information, incorporation of the forecast objective in constructing factors, specification of a multivariate forecast objective, data grouping before constructing factors, and selection of the number of factors in a data-driven way. Our empirical results show that each of these features helps to improve forecast accuracy, especially for the shortest and longest maturities. Factor-augmented methods perform well in relatively volatile periods, including the crisis period in 2008-9, when simpler models do not suffice. The macroeconomic information is exploited best by partial least squares methods, with principal component methods ranking second best. Reductions of mean squared prediction errors of 20-30% are attained, compared to the Nelson-Siegel model without macro factors.

Chapter 3 is based on Croux and Exterkate (2011). We criticize existing factor construction methods, which are widely used to summarize a large panel of variables by means of a relatively small number of representative factors, because they lack two important properties: robustness to outliers, and sparsity, that is, having relatively few nonzero factor loadings. We propose a novel factor construction procedure that enjoys both of these properties. Compared to more traditional factor construction methods, we find that this procedure leads to better interpretable factors and to a favorable forecasting performance, both in a Monte Carlo experiment and in two empirical applications to large data sets, one from macroeconomics and one from microeconomics.

In contrast with the linear universe, the world of nonlinear forecasting methods with large predictor sets is at a much earlier stage of its development. In Chapters 4 and 5, we borrow the kernel methodology from the machine learning community, where its main use is in classifying information, for example the optical recognition of hand-written characters. This framework is extended to the context of time-series prediction in Chapter 4 and it turns out that in many cases, kernel ridge regression outperforms traditional linear forecasting methods. However, there are important open questions related to model selection in kernel ridge regression. Little seems to be known about how to choose a kernel, let alone about the associated tuning parameters. The theoretical investigations in Chapter 5 culminate in an easy-to-use guideline for model selection within this framework, and it is shown that use of this guideline results in selecting models that provide high-quality forecasts.

In particular, Chapter 4, based on Exterkate, Groenen, Heij, and van Dijk (2011a), puts forward kernel ridge regression as an approach for forecasting with many predictors that are related nonlinearly to the target variable. In kernel ridge regression, the observed predictor variables are mapped nonlinearly into a high-dimensional space, where estimation of the predictive regres-

sion model is based on a shrinkage estimator to avoid overfitting. We extend the kernel ridge regression methodology to enable its use for economic time-series forecasting, by including lags of the dependent variable or other individual variables as predictors, as typically desired in macroeconomic and financial applications. Monte Carlo simulations as well as an empirical application to various key measures of real economic activity confirm that kernel ridge regression can produce more accurate forecasts than traditional linear methods for dealing with many predictors based on principal component regression.

In Chapter 5, based on Exterkate (2011), several modelling issues in kernel ridge regression are investigated. In particular, we study the influence of the choice of kernel and of the setting of tuning parameters on forecast accuracy. Several popular kernels are reviewed, including polynomial kernels, the Gaussian kernel, and the Sinc kernel. We interpret the latter two kernels in terms of their smoothing properties, and we relate the tuning parameters associated to all these kernels to smoothness measures of the prediction function and to the signal-to-noise ratio. Based on these interpretations, guidelines are provided for selecting the tuning parameters from small grids using cross-validation. A Monte Carlo study confirms the practical usefulness of these rules of thumb. Finally, the flexible and smooth functional forms provided by the Gaussian and Sinc kernels makes them widely applicable, and we recommend their use instead of the popular polynomial kernels in general settings, in which no information on the data-generating process is available.

All chapters of this thesis provide partial answers to the question posed in the introduction, what can we do to enhance the quality of our economic forecasts in the future? In each chapter, it is shown that the proposed methods for constructing factors (Chapters 2 and 3) or estimating high-dimensional nonlinear models (Chapters 4 and 5) help in producing more accurate forecasts, leading to reductions up to 30% in mean squared prediction errors in several cases. The obvious (and currently open) question is, what would happen if we could somehow put the ideas underlying these methods together in an integrated framework? Such a framework would enable a forecaster to estimate flexible nonlinear predictive models with a large number of predictors, while maintaining robustness to outlying observations and interpretability of the model. Such a framework may not be available any time soon, but attempts to move toward it will undoubtedly open up interesting avenues for future research.

In this thesis, the focus is mainly on macroeconomic applications, plus one microeconomic example in Chapter 3. The events of the last few years have shown how large the impact of financial developments has become, and how difficult they are to forecast. In the financial world, large data sets, nonlinearities, and large outliers also provide major challenges, arguably more than in other economic disciplines. Thus, it will be interesting to see if any of the techniques introduced in this thesis can be adapted for use in financial forecasting, and to which results they may lead.

Chapter 2

Forecasting the Yield Curve in a Data-Rich Environment using the Factor-Augmented Nelson-Siegel Model

This chapter is based on Exterkate et al. (2011b).

2.1 Introduction

Forecasting the yield curve is of much practical interest, not only for individual investors, but also for pension funds and monetary and fiscal policy makers. Despite its relevance, surprisingly little research effort has been spent on this issue until recently. This may be partly due to negative results obtained by early studies on yield curve forecasting, such as Duffee (2002). That study investigates the forecasting performance of affine term structure models, which postulate that yields evolve as affine functions of a limited number of latent risk factors (see Vašíček, 1977; Cox et al., 1985; Duffie and Kan, 1996; Dai and Singleton, 2000). Duffee (2002) dismisses this entire class of models for forecasting purposes by showing that the forecasts obtained from affine models are inferior to random walk (no-change) forecasts.

More positive results have emerged recently based on the framework of Nelson and Siegel (1987). Intended to describe cross-sectional aspects of the yield curve, the Nelson-Siegel model imposes a parsimonious three-factor structure on the links between yields on bonds with different maturities, where the factors can be interpreted as level, slope and curvature. While the original Nelson-Siegel model is static, Diebold and Li (2006) find that an extension with specifications of the factor dynamics renders a model providing forecasts that outperform the random walk and various other forecasting approaches, see also Christensen et al. (2011).

Both the Nelson-Siegel and affine models are essentially purely statistical models of the yield curve. At the same time, it is widely believed that yield curve dynamics are closely linked to macroeconomic developments, for various reasons. For example, central banks around the world use short-term interest rates as their main monetary policy instrument, and it is widely recognized that their actions respond to macroeconomic aggregates such as inflation and output; see Taylor (1993). Given that the central bank policy rate is closely linked to yields on bonds with short maturities, macroeconomic developments should (indirectly) affect the short

end of the yield curve. In fact, because longer-term interest rates can be regarded as a weighted average of expected future short-term rates, it is plausible that the entire yield curve responds to macroeconomic shocks. A link also exists in the reverse direction. Economic agents respond to changing interest rates by altering their investment plans and by adjusting their inflation expectations. Not surprisingly then, several recent studies have developed extensions of yield curve models that incorporate macro variables, in an attempt to capture their interaction, see Ang and Piazzesi (2003); Dewachter and Lyrio (2006); Hördahl et al. (2006); Rudebusch and Wu (2008). Diebold et al. (2006) propose a way to include macroeconomic factors in the Nelson-Siegel model, finding clear evidence that macro aggregates have a statistically significant effect on yields. The analysis in Diebold et al. (2006) (and in the other studies cited above) is purely based on in-sample fit, however, and does not consider out-of-sample forecasting.

In this chapter we examine several important aspects related to the inclusion of macroeconomic variables in the Nelson-Siegel model from a forecasting perspective. First, in today's data-rich environment, a natural question is which macro factors to include in the model. Diebold et al. (2006) use three specific variables, intended to represent the level of inflation, real economic activity, and monetary policy, respectively. Arguably, many more macro variables may influence the evolution of the yield curve. However, including a large number of individual variables leads to an abundance of parameters to be estimated, with the resulting estimation uncertainty negatively affecting the accuracy of out-of-sample forecasts. A natural possibility is to use factors extracted from a large panel of macro variables. Indeed, De Pooter et al. (2007) find that including a small number of principal components leads to an improvement in forecast accuracy, compared to the use of specific individual variables. In a comparative forecasting study by Favero et al. (2011), this setup compares favorably to various alternatives, including a factor-augmented affine model (Mönch, 2008) and various arbitrage-free models of the term structure. Here we examine this issue in more detail by comparing a number of data-based variable selection methods with several approaches to construct factors from a large panel of variables. The methods for selecting variables in a data-driven way are based on least angle regression (LARS) or multiresponse sparse regression (MRSR), as proposed by Efron et al. (2004) and Similä and Tikka (2006), respectively. The factor construction methods include Principal Component Analysis (PCA), next to a number of alternative approaches discussed below.

Second, we investigate whether it is useful to take the forecast objective explicitly into account when constructing the macro factors. When selecting specific individual variables it is natural to make use of this forecast objective, but this is not the case for methods that construct factors from a large panel of variables. In particular, PCA, which is by far the most popular factor construction method, renders the same factors regardless of which series we aim to forecast. We examine three alternative approaches that do take the series to be forecasted into consideration in the construction of the macro factors, namely Partial Least Squares (PLS; Wold, 1966), Principal Covariate Regression (PCovR; Heij et al., 2007), and PCA based on variables selected by thresholding rules (Bai and Ng, 2008). In the first two approaches, factors are constructed using all available macro variables, but with weights depending on their degree of comovement with the forecast objective. In the latter approach, principal components are taken only from those variables that are most correlated with the variable that we intend to predict. Hence, with this method we also address the issue whether or not it is desirable to include all available data in PCA, see also Boivin and Ng (2006).

Third, while there may be little doubt that the yield curve is linked to the macroeconomy, it is plausible that different characteristics of the yield curve are related to different macroeconomic aspects. For example, Diebold et al. (2006) show that the level factor in the Nelson-Siegel model seems closely related to inflation, while the slope factor is linked more to measures of economic activity. We examine this issue by using the variable selection and factor construction methods in two different ways. Specifically, we construct macro factors either for the three Nelson-Siegel factors jointly or for the level, slope and curvature factors separately.

Fourth, we explore whether it pays off to construct separate factors from different groups of related macro variables, instead of one large pool of all available variables. As inflation, real economic activity, and monetary policy appear to be the most relevant macroeconomic dimensions for the yield curve, it may be worthwhile to construct factors that are explicitly related to these characteristics. We implement this idea for both PCA and PCovR.

Fifth, in addition to the question which macro factors to include in a yield curve model, it is also relevant to ask how many factors to include. Here we compare the predictive accuracy of models with a fixed number of three factors to models with a varying number of macro factors based on historical forecasting performance.

We address the issues described above empirically, by examining the out-of-sample forecasting performance of the Factor-Augmented Nelson-Siegel (FANS) model for the US yield curve over the period from January 1994 until December 2009, for forecast horizons of one, three, six and twelve months ahead. We make use of a macroeconomic information set consisting of 132 individual variables. Our results show that macroeconomic information is particularly useful in volatile times, including the crisis period in 2008-9. The preselected variables suggested by Diebold et al. (2006) provide the best forecasts only for predicting medium-term yields (15 to 60 months) at the longest horizon considered here (12 months). Overall, the most accurate forecasts are obtained by using partial least squares; thus, explicitly considering the forecast objective when constructing factors. For long maturities (over 60 months), it also works well to form groups of related variables and then extract factors from these groups, preferably by using principal covariate regression. For shorter maturities, it is better to extract principal components or covariates from all available information. Methods that treat the three Nelson-Siegel factors jointly generally outperform methods that treat these factors separately. Finally, varying the number of macro factors based on recent historical performance leads to an additional improvement in forecast accuracy.

In the remainder of this chapter, Section 2.2 begins with a description of the Nelson-Siegel model and the extension by Diebold et al. (2006). Moreover, the techniques of least angle regression, multiresponse sparse regression, principal covariate regression, partial least squares, and hard and soft thresholding are discussed. Section 2.3 describes the data on U.S. zero-coupon yields and macroeconomic aggregates, as well as details of our forecasting procedure. Section 2.4 contains the empirical forecasting results, and Section 2.5 concludes.

2.2 Methodology

This section reviews the Nelson-Siegel model, its extension with macro factors, and the techniques that we employ to construct macroeconomic factors to be included in this model.

2.2.1 Nelson-Siegel model

Nelson and Siegel (1987) propose a parsimonious model for describing the yield curve. Using the representation as put forward in Diebold and Li (2006), the model is given by

$$y_t(\tau_i) = \beta_{1t} + \beta_{2t}\left(\frac{1-\exp(-\lambda_t\tau_i)}{\lambda_t\tau_i}\right) + \beta_{3t}\left(\frac{1-\exp(-\lambda_t\tau_i)}{\lambda_t\tau_i} - \exp(-\lambda_t\tau_i)\right), \quad (2.1)$$

where $y_t(\tau_i)$ is the yield at time t for a maturity of τ_i months. As discussed in Diebold and Li (2006), $\beta_{1t}, \beta_{2t},$ and β_{3t} can be interpreted as level, slope, and curvature factors, respectively. Further, λ_t determines the rate of decay of the loading for the slope factor β_{2t} and the maturity at which the loading for the curvature factor β_{3t} attains its maximum value.

We consider a fixed set of m maturities $(\tau_1, \tau_2, \ldots, \tau_m)$ and denote the corresponding vector of observed yields at time t by $y_t = (y_t(\tau_1), y_t(\tau_2), \ldots, y_t(\tau_m))'$ and the parameter vector by $\beta_t = (\beta_{1t}, \beta_{2t}, \beta_{3t})'$. Adding an error term, representing measurement error, for example, to Equation (2.1), we obtain

$$y_t = \Lambda_t\beta_t + \varepsilon_t, \quad (2.2)$$

where Λ_t depends on λ_t only.

Diebold and Li (2006) interpret Equation (2.2) as the measurement equation of a state space model. The state equations represent the dynamics of the factors β_t, which are assumed to evolve according to a first-order[1] vector autoregressive process (with mean μ):

$$\beta_t - \mu = A(\beta_{t-1} - \mu) + \eta_t. \quad (2.3)$$

The disturbances ε_t and η_t are assumed to be zero-mean white noise and to be mutually uncorrelated. The covariance matrix of ε_t is commonly assumed to be diagonal, see, for example, Diebold and Li (2006) and Diebold et al. (2006). The covariance matrix Q of η_t is left unrestricted in those studies. Assuming normally distributed error terms, maximum likelihood estimates and forecasts are obtained using the Kalman filter.

In most studies, λ_t is assumed constant, $\lambda_t = \lambda$, and its value is fixed by the researcher, as in Diebold and Li (2006). Alternatively, a constant λ can also be estimated along with the other model parameters using the Kalman filter, as in Diebold et al. (2006) and De Pooter et al. (2007). Koopman et al. (2010) propose ways of allowing for time-varying λ_t. In our empirical analysis, we will use a time-invariant λ, but we do estimate it along with the other model parameters.[2]

2.2.2 Factor-Augmented Nelson-Siegel model

Now assume that, at every time t, a large number of k macroeconomic variables is available, denoted by $x_t = (x_{1t}, \ldots, x_{kt})'$. For reasons of parsimony, we wish to summarize this large amount of information by a limited number of p factors, $f_t = (f_{1t}, \ldots, f_{pt})'$, with $p \ll k$. These factors can, for example, be obtained by preselection of a subset of variables from x_t, as in

[1] As Diebold and Li (2006) find that one lag is sufficient to describe the dynamics of β_t, no further lags are included in (2.3).

[2] That is, λ is kept fixed for the estimation period. As we discuss below, the Nelson-Siegel model is estimated using a ten-year rolling window, so that the estimated value of λ will actually vary from one window to the next.

Diebold et al. (2006). Alternatively, the factors can be extracted from x_t using principal component analysis as in De Pooter et al. (2007). We describe alternative methods for constructing these factors below, in Sections 2.2.3-2.2.7.

For now, assume that f_t is available and that all its elements are normalized to have mean zero. We follow Diebold et al. (2006) in their procedure for incorporating this information into the Nelson-Siegel framework. The observation equation (2.2) remains unchanged. In the state equation (2.3), f_t is appended to the state vector β_t and the dimensions of A, η_t, and Q are increased as appropriate. Hence, the macro factors affect the individual yields only via the Nelson-Siegel factors.

Introducing macroeconomic information into the model in this manner leads to a substantial increase in the number of parameters. For example, the inclusion of three macro factors increases the dimension of both A and Q from 3×3 to 6×6. As Q is symmetric, the factor-augmented model has $(36 - 9) + (21 - 6) = 42$ additional parameters compared to the basic Nelson-Siegel model in Equations (2.2) and (2.3). To avoid problems of overfitting and numerical difficulties associated with estimating such a large number of parameters, we impose two restrictions. First, the VAR transition matrix is restricted to have the following structure:

$$A = \left(\begin{array}{c|c} \text{diagonal} & \text{unrestricted} \\ \hline \text{zero} & \text{diagonal} \end{array} \right), \tag{2.4}$$

where the blocking corresponds to the partitioning of the state vector into β_t and f_t. In particular, this restriction implies that we do not model any yields-to-macro feedback, nor dynamic interaction among the Nelson-Siegel factors. Second, the covariance matrix Q is restricted to be diagonal. Diebold et al. (2006) tested and rejected both of these restrictions. However, as mentioned before, their analysis was strictly based on in-sample criteria. As our focus is on out-of-sample forecasting, the notion of parsimony is an important consideration. We have constructed yield forecasts with the model with and without these restrictions, finding that the restricted version outperforms the unrestricted model in almost all cases, whether macro information is included or not. The unrestricted model is superior only for forecasting short-term yields over short horizons. For all other cases, dropping the restrictions leads to a dramatic increase in mean squared prediction error of more than 40% on average, possibly due to overfitting. A similar result was obtained by Diebold and Li (2006) for the Nelson-Siegel model without macro factors. Therefore, in Section 2.4 we only report results obtained from the restricted model.

For convenience, the model considered in this study is shown below:

$$\begin{array}{rcl@{\qquad}rcl} y_t & = & \Lambda\beta_t + \varepsilon_t, & \varepsilon_t & \sim & \mathcal{NID}\left(0, \operatorname{diag}\left(\sigma_1^2, \ldots, \sigma_m^2\right)\right), \\ \begin{pmatrix} \beta_t - \mu \\ f_t \end{pmatrix} & = & A \begin{pmatrix} \beta_{t-1} - \mu \\ f_{t-1} \end{pmatrix} + \eta_t, & \eta_t & \sim & \mathcal{NID}\left(0, \operatorname{diag}\left(q_1, \ldots, q_{p+3}\right)\right), \end{array} \tag{2.5}$$

with Λ and A as defined above; in particular, A is restricted as in Equation (2.4). We shall refer to this model as the Factor-Augmented Nelson-Siegel (FANS) model.

Below, we describe methods to extract the factors f_t from the macroeconomic variables x_t. To facilitate the discussion, we define

$$B^+ = \begin{pmatrix} \beta_2' \\ \beta_3' \\ \vdots \\ \beta_T' \end{pmatrix}, \quad B = \begin{pmatrix} \beta_1' \\ \beta_2' \\ \vdots \\ \beta_{T-1}' \end{pmatrix}, \quad X = \begin{pmatrix} x_1' \\ x_2' \\ \vdots \\ x_{T-1}' \end{pmatrix}, \quad F = \begin{pmatrix} f_1' \\ f_2' \\ \vdots \\ f_{T-1}' \end{pmatrix},$$

where T is the length of the estimation sample. To rule out scale effects, every column of X (that is, the time series of observations on each variable separately) is normalized to have mean zero and unit variance over the estimation window.

In essence, all of the methods for constructing factors considered in this study boil down to choosing a $k \times p$ matrix W and defining $F = XW$. That is, the factors that we use in the FANS model are linear combinations of observed macro variables. The following sections describe methods for choosing these combinations. Common features are that they achieve a considerable dimension reduction (f_t contains far fewer elements than x_t; that is, $p \ll k$) and that f_{t-1}, together with β_{t-1}, has predictive power for β_t.

One more technical comment is in place. Each of the factor construction methods described below requires, in principle, full observability of B, whereas it is latent in our application. We overcome this problem by the following straightforward approach. First, we estimate the Nelson-Siegel model without macro factors, by maximum likelihood using the Kalman filter. The resulting estimates of β_t (in Kalman filtering terminology: the smoothed state vectors) form the matrices B and B^+, which are employed to build response variables in constructing the macro factors f_t. The FANS model is then reestimated using these factors. Although this procedure could be iterated, this is not pursued here for computational considerations.

2.2.3 Least angle regression

The Least Angle Regression (LARS) methodology (Efron et al., 2004) can be used to select a limited number of informative variables out of a large group for inclusion in a predictive regression model. The main idea is to "add" predictors to the model one at a time, starting with the predictor that correlates most with the target variable. This predictor's coefficient is increased from its starting value of zero, up to the point where the residual is equally correlated with the predictor chosen initially and a second predictor. This second predictor is added to the "most correlated" set, and the coefficients on both predictors in this set are now simultaneously increased in such a way as to keep the residual equally correlated with the two predictors. As soon as a third predictor shows equal correlation, it also enters the "most correlated" set, and so on, until either the residual is zero or all predictors have entered.

At any stage of this procedure, only the predictors in the "most correlated" set have nonzero coefficients. Hence, LARS can be used as a variable selection method by stopping the algorithm after a prespecified number of predictors have been selected. This procedure closely approximates the more well-known Lasso method proposed by Tibshirani (1996); see Efron et al. (2004) for a discussion of this similarity.

In our setting, we wish to select a small number of variables (columns) from X which, together with B, have predictive power for B^+. To this end, we employ a simple two-step procedure: we first estimate the auxiliary regression

$$B^+ = \iota\alpha' + B\Delta + E, \tag{2.6}$$

where ι is a vector of ones, α is a vector of constants, Δ is a matrix of regression coefficients, and E is a matrix of disturbance terms. The residuals R from this regression are used as response variables in LARS, in order to explain those features of the yield factors that remain after correcting for autoregressive effects.

As R is multivariate in our setup, the most obvious application of the LARS algorithm is to feed the columns of R into the algorithm one at a time. An alternative approach is "multiresponse sparse regression" (MRSR), proposed by Similä and Tikka (2006) as an extension of LARS that allows for a multivariate response variable.

To apply LARS in a multivariate setting, an extension of the correlation concept is required to make the condition "equally correlated" meaningful. Denote the fitted value of R, based on the first m regressors chosen, by $\hat{R}_m$ (with $\hat{R}_0 = 0$), and denote the j-th column of X by $x_{(j)}$. Following Similä and Tikka (2006), the role of correlations in the description of the univariate case above is now played by $\left|\left|\left(R - \hat{R}_m\right)' x_{(j)}\right|\right|$, where $||v||$ represents the L_2 vector norm $\left(\sum_i v_i^2\right)^{1/2}$. No other changes to the procedure are needed. An efficient algorithm to find the order in which variables are added is presented in Similä and Tikka (2006).

We use both the univariate and the multivariate variant of the LARS algorithm to select a predefined number of explanatory variables from a large panel of macro data. The selected predictors are used as the macro factors f_t in the FANS model.

2.2.4 Principal component regression and principal covariate regression

Heij et al. (2007) propose principal covariate regression (PCovR) as an alternative to principal component regression (PCR). In PCR, principal components are first extracted from a group of predictors and then used as regressors. To obtain p principal components, the k predictors in the matrix X are "summarized" in $p \ll k$ factors by minimizing $||X - XUV||$ over the $k \times p$ matrix U and the $p \times k$ matrix V. The desired factors are the columns of XU. The method then proceeds as a standard least-squares regression of the (univariate) dependent variable z on a constant and XU. That is, the objective function $||z - \alpha\iota - XU\gamma||$ is minimized over the scalar α and the vector γ.

Heij et al. (2007) argue that the failure to take the prediction objective into account when constructing the factors is a drawback of PCR. To overcome this problem, they combine the two steps of PCR into one objective function: in the same notation as above, they minimize

$$w \left|\left|z - \alpha\iota - XU\gamma\right|\right|^2 / \left|\left|z\right|\right|^2 + (1 - w) \left|\left|X - XUV\right|\right|^2 / \left|\left|X\right|\right|^2, \tag{2.7}$$

where $w \in [0, 1]$ is a tuning parameter that governs the relative weight placed on each of the two objectives. Thus, the aims of good prediction and adequate use of the data are balanced in the PCovR objective (2.7); setting w at a higher value means that more weight is placed on predicting z relative to summarizing X, whereas choosing $w = 0$ amounts to standard PCR.

An obvious multivariate extension of Objective (2.7) to our problem is to minimize

$$w \left|\left|R - XU\Gamma\right|\right|^2 / \left|\left|R\right|\right|^2 + (1 - w) \left|\left|X - XUV\right|\right|^2 / \left|\left|X\right|\right|^2, \tag{2.8}$$

where Γ is now a matrix and R is, as before, the matrix of residuals from the regression of B^+ on B and a constant. Direct minimization of Objective (2.8) can be done using two singular

value decompositions, as outlined in Heij et al. (2007). To operationalize this procedure, we need to specify a value for w. The high-dimensional nature of X leads to overfitting if w is chosen too large; see Heij et al. (2006) for a discussion. The upper bound that they propose corresponds to about $w \leq 0.1$ for our problem. In order to make PCovR sufficiently different from standard PCR, we do not want to set w too small either; therefore, we fix $w = 0.1$.

2.2.5 Partial least squares

Wold (1966) proposed partial least squares (PLS) as a technique for finding orthogonal linear combinations of the variables in X that have predictive power for a univariate response, say, z. That is, in contrast to PCR, the PLS factors are constructed to explain the variance of z rather than X. Following Garthwaite (1994), we describe PLS in terms of sequential regressions.

Assume that X and z have mean zero. We first regress z on each column $x_{(j)}$ of X, yielding fitted vectors $\hat{z}_{(j)}$. The first factor is then constructed as $f_1 = \sum_j w_j \hat{z}_{(j)}$, where the weights are proportional to the variances of the columns of X, $w_j = x'_{(j)} x_{(j)}$, such that $f_1 = \sum_j \operatorname{cov}\left(x_{(j)}, z\right) x_{(j)}$. Hence the first factor is a weighted average of the macro variables x_t, with weights depending on their covariance with the variable we aim to forecast. Next, replace both z and X by their residuals from regressions on f_1. The second factor f_2 is then found by applying the same procedure to the "new" z and X. Continuing in this manner, we can sequentially construct $f_3, \ldots, f_p$. This procedure can be copied almost verbatim for a multivariate response Z; details can be found in Garthwaite (1994).

To apply this method in our context, we use a similar two-step approach as for the LARS algorithm. That is, we first find the residuals R from a regression of B^+ on B and a constant. The three columns of R are then used, either one at a time or all together, as response vectors z in the PLS algorithm. The constructed factors are used together as f_t in the FANS model.

2.2.6 Hard thresholding

Bai and Ng (2008) propose hard thresholding as a simple method for variable selection. In their univariate setting, consider forecasting a univariate response z using its own lag z^- and the columns of X. To select the most relevant columns, estimate the linear model

$$z = \alpha\iota + \delta z^- + \gamma_j x_{(j)} + \varepsilon \tag{2.9}$$

with least squares, for all columns of X, $j = 1, 2, \ldots, k$. We then select those columns of X for which the t test rejects $\gamma_j = 0$, at a prespecified significance level. Following the suggestion in Bai and Ng (2008), to construct p predictors, we select p^* variables, with $p^* > p$. The first p principal components extracted from this subset of variables are then used as predictors.

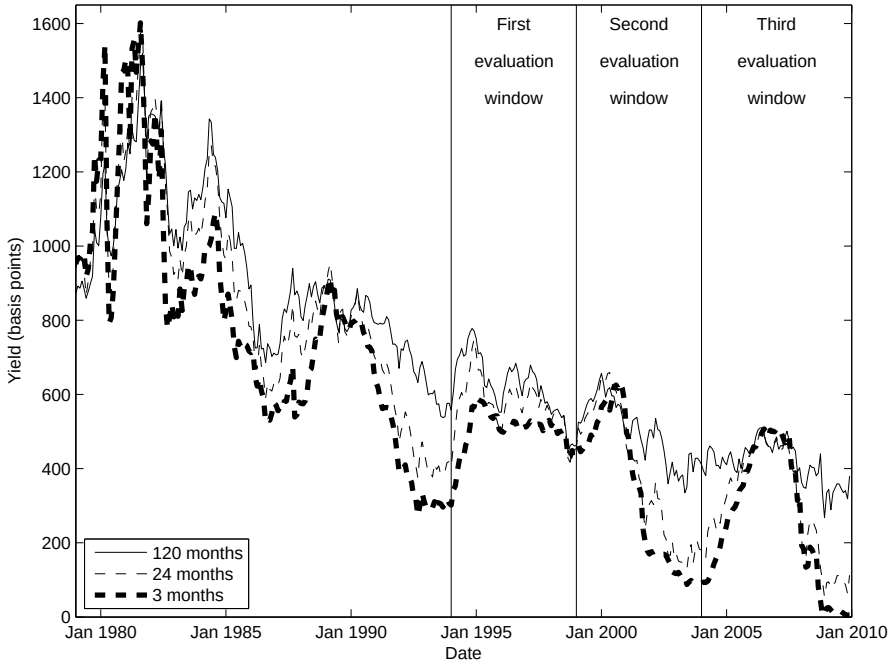
In our multivariate context, we estimate

$$B^+ = \iota\alpha' + B\Delta + x_{(j)}\gamma'_j + E \tag{2.10}$$

and compute the Wald statistic for the test $\gamma_j = 0$.[3] We select the p^* variables for which this

[3]Note that this Wald test involves a zero restriction on one parameter in each of multiple (in our case, three) interdependent univariate regression equations. The computation of the Wald statistic in this nonstandard situation is outlined in Roy (1957).

Figure 2.1: Time series plots of U.S. zero-coupon yields for three selected maturities.



hypothesis is rejected at 5% significance, and we use p principal components extracted from this set as macro factors.

2.2.7 Soft thresholding

Soft thresholding is also proposed by Bai and Ng (2008). They argue that a drawback of a hard thresholding rule is that it can easily lead to the selection of many similar predictors, because in deciding whether or not to include a certain predictor, the information contained in the other predictors is not considered.

They propose to use a sequential method, such as LARS or the Lasso, to select variables, and then to find the number of variables that minimizes an information criterion, such as the BIC. In our multivariate setting, we use MRSR as the selection method. The number of selected variables is chosen by minimizing the BIC in the system of linear equations

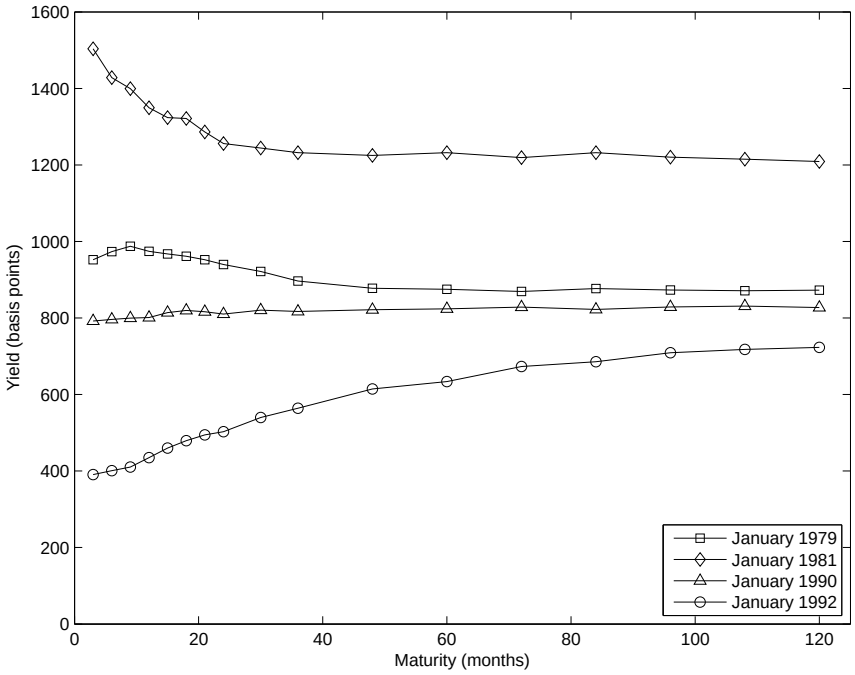
$$B^+ = \iota\alpha' + B\Delta + X_m\Gamma_m + E, \tag{2.11}$$

where X_m consists of the first m selected columns of X. The p macro factors to be used in the FANS model are again principal components extracted from the p^* variables that minimize the BIC.

2.3 Data and forecasting procedure

This section describes the yield and macroeconomic data sets that we use, as well as details of the estimation and forecasting procedures.

Figure 2.2: The yield curve in four selected months.



2.3.1 Data

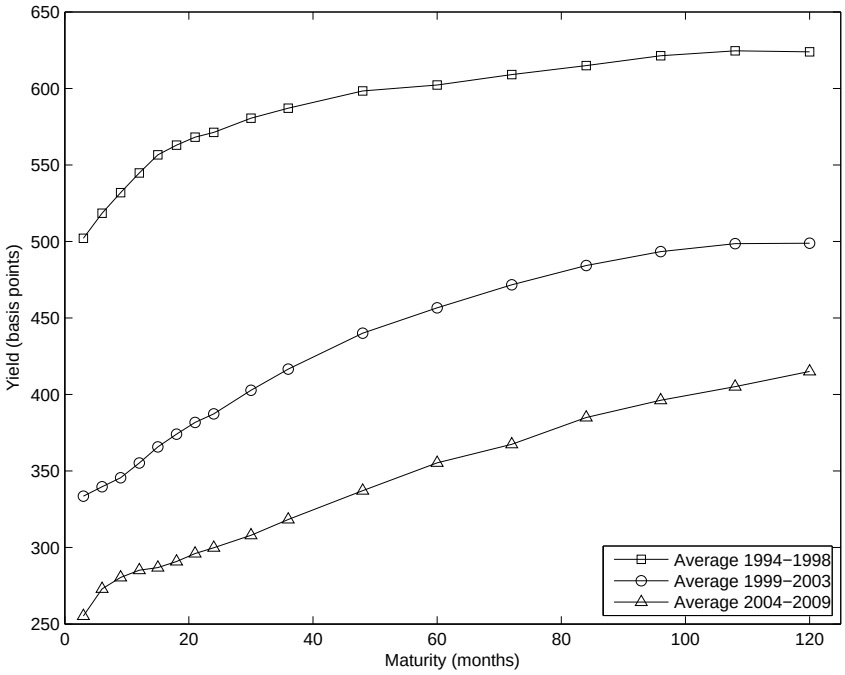
Our yield data are unsmoothed Fama and Bliss (1987) U.S. Treasury yields at the monthly frequency, updated using the Fama-Bliss algorithm and CRSP data on prices of individual Treasury securities to cover the period from January 1979 to December 2009, for maturities of 3, 6, 9, 12, 15, 18, 21, 24, 30, 36, 48, 60, 72, 84, 96, 108, and 120 months. [4] For more information on the yield construction methodology, we refer to Bliss (1997). Time series plots of a short-term (three months), a medium-term (two years), and a long-term (ten years) yield are shown in Figure 2.1. We observe that these yields vary over a wide range of values over the sample period. Moreover, although the yields for different maturities show a large degree of comovement, it can be seen that the spreads between the different yields also vary considerably over time.

The different possible shapes of the yield curve are illustrated in Figure 2.2, where we show all seventeen yields in four selected months. A "typical" yield curve is increasing and slightly concave, as is the case in January 1992. However, other shapes, such as humped (1979), decreasing (1981), and almost flat (1990), occur fairly regularly. The Nelson and Siegel (1987) model can accurately approximate all these shapes.

The macro factors that we include in the FANS model are extracted from a set of 132 monthly variables. These variables are categorized into eleven groups in an economically meaningful way, see Stock and Watson (2002). For example, one group contains price indices, one contains aggregates related to real output, and so on. A previous version of this data set, ending in December 2003, is described in Stock and Watson (2005). They also describe transformations to stationarity. We have extended their data set to cover the period until December 2009.

[4] This data set was also analyzed by Jungbacker et al. (2010) and Koopman et al. (2011). We thank Michel van der Wel for making these data available to us.

Figure 2.3: The average yield curve in the three evaluation subsamples.



2.3.2 Forecasting procedure

We estimate the parameters of the FANS model over a rolling window with a length of 120 months. At the end of each window (say, at time t), predictions are made for times $t + h$, with forecast horizon $h = 1, 3, 6,$ or 12. Thus, for forecast horizon h, the first prediction that we make is for month h of 1989. Note that, in contrast to Diebold and Li (2006), we produce iterated rather than direct forecasts for horizons longer than one month, as the FANS Model (2.5) is formulated in terms of one-step-ahead predictions.

We report forecasting results for the subsamples January 1994 until December 1998, January 1999 until December 2003, and January 2004 until December 2009. The first subsample closely corresponds to the forecasting period considered by Diebold and Li (2006); the last subsample includes the 2008-9 crisis period. (The forecasts for January 1989 until December 1993 are used only for selecting the number of macro factors, as we describe below.) As is clear from Figure 2.1, the yields exhibit a lower level and more volatility in the latter two of these three subsamples. Figure 2.3 shows that the average yield curves (defined as simple arithmetic averages) have roughly similar shapes in all subsamples, although the level decreases over time.

We denote the point forecast for the yield for maturity τ at time $t + h$, made at time t using method M, by $\hat{y}_{t+h|t}(M, \tau)$. Predictive accuracy is evaluated by the mean squared prediction error, defined as

$$\text{MSPE}(M, h, \tau) = \frac{1}{T} \sum_{t} \left[\hat{y}_{t+h|t}(M, \tau) - y_{t+h}(\tau) \right]^2, \tag{2.12}$$

where the summation runs over the period with length T over which we wish to evaluate the performance.

Table 2.1: Macro factor construction methods.

Abbreviation	Name	Description
Basic methods		
NO	No macro factors	The Nelson-Siegel model without any macro factors, as in Diebold and Li (2006)
DRA	Diebold, Rudebusch, Aruoba	Macro factors are capacity utilization, the federal funds rate, and the price deflator for personal consumption expenditures, as in Diebold et al. (2006)
LARSU	Univariate LARS	The first macro variable selected by LARS for each of the three yield factors β
PLSU	Univariate PLS	The first PLS factor constructed for each of the three yield factors
PCAG	PCA on groups	One principal component from each of the three groups "Real Output and Income", "Interest Rates and Spreads", and "Price Indices", from Stock and Watson (2002)
PCOVRG	PCovR on groups	One principal covariate from each of the same three groups as in PCAG
Multivariate methods		
LARSM	Multivariate LARS	The first three macro variables selected by MRSR for the three yield factors jointly
ST	Soft thresholding	Three factors constructed by the soft thresholding rule
HT	Hard thresholding	Three factors constructed by the hard thresholding rule
PLSM	Multivariate PLS	The first three PLS factors constructed for the three yield factors jointly
PCA	Principal component analysis	Three principal components from the full macro data set, as in De Pooter et al. (2007)
PCOVR	Principal covariate regression	Three principal covariates from the full macro data set

The forecasting methods that we consider differ in the procedure for constructing macroeconomic factors. Table 2.1 gives an overview. In all but the first of these methods, the number of macro factors is fixed at three to make our results comparable to those in Diebold et al. (2006) and De Pooter et al. (2007).

The groups of variables used in the PCAG and PCOVRG methods are three of the eleven groups identified by Stock and Watson (2002), chosen such that each group contains one of the variables used by Diebold et al. (2006). This choice is intended to make our results comparable to theirs.

We consider the first six methods listed in Table 2.1 as our basic methods. The second set of six are multivariate methods, each selecting three macro factors. A final set of six methods contains variants of each of these multivariate methods, in which the number of factors is not fixed at three. Instead, we estimate the FANS model with either 1, 2, 3, or 4 macro factors, after which the actual number of factors used in forecasting is selected based on past performance; that is, the number of factors leading to the smallest MSPE over the most recent 60 months is used. Note that this procedure may lead to different numbers of factors being selected for different forecast horizons. The abbreviations used to refer to these methods are equal to those for their three-factor counterparts, with a B appended for "Best number of factors".

Table 2.2: Mean squared prediction errors, 1994-1998.

Maturity		3				24				120			
Horizon		1	3	6	12	1	3	6	12	1	3	6	12
NO		281	1068	3031	8641	760	3376	7093	14102	715	2870	6756	13354
Basic methods													
DRA	(0)	1.68	1.32	1.27	1.77	1.06	1.18	1.38	1.96	1.01	1.09	1.21	1.65
LARSU	(0)	3.65	2.49	1.88	1.74	1.11	1.05	1.10	1.46	1.21	1.08	1.15	1.46
PLSU	(2)	2.34	2.02	1.67	1.49	1.31	1.17	1.03	0.88	1.43	1.17	1.00	0.92
PCAG	(6)	1.51	**0.95**	0.95	1.09	1.03	**0.98**	0.99	1.06	1.10	0.99	0.96	1.00
PCOVRG	(1)	1.26	0.99	1.14	1.62	1.02	1.08	1.26	1.59	1.03	1.08	1.19	1.37
Multivariate methods													
LARSM	(0)	2.37	1.90	1.44	1.46	1.13	1.11	1.16	1.66	1.12	1.03	1.09	1.51
ST	(0)	2.20	1.48	1.16	1.01	1.17	1.16	1.12	1.11	1.23	1.21	1.15	1.14
HT	(0)	2.49	2.33	1.85	1.48	1.11	1.21	1.23	1.19	1.13	1.21	1.26	1.27
PLSM	(3)	2.16	1.61	1.29	0.93	1.26	1.06	0.98	0.85	1.42	1.18	1.04	1.02
PCA	(6)	1.72	1.30	1.03	1.07	1.00	0.98	0.96	0.98	1.01	**0.97**	0.98	1.07
PCOVR	(0)	1.40	1.16	1.22	1.41	1.04	1.06	1.14	1.21	1.01	1.02	1.08	1.13
Selected number of factors													
LARSMB	(0)	2.44	2.18	1.64	1.48	1.16	1.17	1.22	1.59	1.10	1.09	1.10	1.43
STB	(5)	1.67	1.30	**0.88**	0.94	0.99	1.02	0.92	1.03	1.04	1.04	0.95	1.06
HTB	(0)	1.68	1.34	1.15	1.08	1.03	1.02	1.03	1.07	1.04	1.01	1.05	1.11
PLSMB	(5)	1.94	1.43	1.01	**0.73**	1.36	1.08	**0.81**	**0.60**	1.42	1.07	**0.77**	**0.71**
PCAB	(1)	1.74	1.32	1.12	1.13	**0.97**	1.01	1.04	1.07	1.00	1.01	1.03	1.09
PCOVRB	(0)	1.32	1.25	1.16	1.16	1.02	1.07	1.14	1.15	1.02	1.02	1.06	1.06

Notes: This table reports MSPE results for forecasts made for the period January 1994-December 1998. The first row reports the MSPE (in squared basis points), as defined in Equation (2.12), for the method NO, see Table 2.1. The remaining rows show relative values MSPE (M, h, τ) /MSPE (NO, h, τ) for all other methods described in Table 2.1. For each maturity and horizon (both measured in months), the lowest relative MSPE is printed in bold. If a column contains no bold entry, none of the methods outperforms the NO benchmark. The numbers in parentheses indicate in how many of the twelve cases the respective methods outperform this benchmark.

2.4 Forecasting results

In this section, we evaluate the forecasting performance of the methods of factor construction listed in Table 2.1. As our focus is not on the parameter estimates, a short discussion on that topic is deferred to Appendix 2.A.

2.4.1 Forecast accuracy

We divide the seventeen maturities that we study into three groups: short (one year or less), medium (one to five years), and long (more than five years). For each of these groups, we report mean squared prediction errors only for one representative maturity (three months, two years, and ten years, respectively). The results are qualitatively similar for all maturities within the three groups; results for the other maturities are available upon request. The values of MSPE (M, h, τ), as defined in Equation (2.12), are shown in Tables 2.2-2.4.

Table 2.3: Mean squared prediction errors, 1999-2003.

Maturity		3				24				120			
Horizon		1	3	6	12	1	3	6	12	1	3	6	12
NO		713	3663	11291	34360	1027	4065	10706	28126	942	2522	5571	12169
Basic methods													
DRA	(6)	1.35	1.13	1.01	0.97	0.94	0.86	**0.74**	**0.51**	1.06	1.12	1.03	0.79
LARSU	(8)	0.86	0.67	0.69	0.79	0.97	0.88	0.88	0.93	1.18	1.09	1.04	1.04
PLSU	(6)	1.14	0.77	0.78	0.84	0.97	0.92	1.00	1.00	1.15	1.10	1.14	1.06
PCAG	(4)	0.87	0.84	0.88	0.97	1.01	1.01	1.02	1.04	1.08	1.11	1.08	1.05
PCOVRG	(12)	0.88	0.64	0.59	0.72	0.94	0.83	0.78	0.76	**0.96**	**0.84**	**0.76**	**0.64**
Multivariate methods													
LARSM	(8)	0.80	0.68	0.71	0.74	0.97	0.91	0.89	0.93	1.14	1.09	1.06	1.09
ST	(5)	0.86	0.88	0.93	0.96	1.07	0.98	1.00	1.02	1.13	1.03	1.05	1.09
HT	(9)	0.89	0.70	0.72	0.77	1.04	0.90	0.85	0.80	1.16	1.02	0.87	0.80
PLSM	(11)	0.81	0.67	0.71	0.80	**0.81**	**0.81**	0.80	0.85	1.10	0.99	0.88	0.89
PCA	(10)	0.82	0.58	0.56	0.65	0.90	0.88	0.85	0.81	1.02	1.03	0.97	0.88
PCOVR	(9)	0.97	0.56	**0.52**	0.66	0.98	0.87	0.80	0.77	1.14	1.13	1.03	0.85
Selected number of factors													
LARSMB	(7)	0.84	0.68	0.71	0.75	1.10	0.91	0.89	0.93	1.21	1.07	1.07	1.09
STB	(7)	0.82	0.82	0.88	0.90	0.97	0.96	0.97	1.01	1.06	1.01	1.01	1.08
HTB	(8)	0.94	0.90	0.88	0.78	1.13	1.08	0.96	0.87	1.07	1.02	0.90	0.90
PLSMB	(11)	0.88	0.66	0.69	0.72	0.94	0.85	0.83	0.82	1.10	0.97	0.90	0.89
PCAB	(12)	**0.78**	**0.55**	0.55	**0.63**	0.90	0.83	0.77	0.75	0.99	0.95	0.81	0.75
PCOVRB	(9)	0.91	0.56	0.59	0.76	0.91	0.85	0.84	0.85	1.06	1.08	1.05	0.94

Notes: This table reports MSPE results for forecasts made for the period January 1999-December 2003.

For all methods, the general pattern can be summarized as follows. Forecasts are more accurate if the horizon is shorter, as can be seen from the rows labeled "NO" in Tables 2.2-2.4. Comparing these three rows, we can also see that the forecasts are in general the most accurate in the first, less volatile period. As Diebold and Li (2006) reported before, simple forecasts made without any macro information are already of reasonable quality. Consider, for example, the MSPE for maturity 3 months at a forecast horizon of 1 month in Table 2.2, which is equal to 281. This means that when forecasting the short-term yield one month ahead, the standard error is $\sqrt{281} \approx 16.8$ basis points. In fact, one-month-ahead prediction errors are around 20 or 30 basis points across all maturities and for all subperiods, except for the shortest maturity in the most recent period.

Considering Table 2.2 further, it becomes clear that including macro information does not improve much on the no-macro benchmark in the first subperiod, except at the longest forecast horizon. The yield curve was not very volatile in 1994-1998, see Figure 2.1, and the Nelson-Siegel model without any macro factors appears to suffice as a forecasting tool in this period.

In 1999-2003 (Table 2.3), however, gains in predictive accuracy range from 6% to as much as 49% for the basic methods. For both the long and the short ends of the yield curve, the best-performing basic method is generally PCOVRG. This method even beats the benchmark in all twelve situations listed in Table 2.3. For medium-term yields (15 to 60 months) the DRA

Table 2.4: Mean squared prediction errors, 2004-2009.

Maturity		3				24				120			
Horizon		1	3	6	12	1	3	6	12	1	3	6	12
NO		2744	6512	14376	35876	927	4257	9699	22608	791	1866	3381	6035
Basic methods													
DRA	(9)	0.94	0.95	0.99	0.92	0.97	0.88	0.85	0.90	1.09	1.03	0.91	1.03
LARSU	(6)	1.07	1.12	1.05	0.90	0.95	0.99	0.95	0.89	1.03	1.03	1.00	0.97
PLSU	(8)	0.74	0.78	0.74	0.66	0.90	0.86	0.75	0.69	1.15	1.04	1.02	1.01
PCAG	(10)	0.88	0.85	0.83	0.79	0.90	0.98	0.95	0.86	**0.96**	1.04	1.03	1.00
PCOVRG	(8)	0.91	0.93	0.91	0.81	0.95	1.00	0.95	0.87	1.15	1.14	1.09	1.06
Multivariate methods													
LARSM	(10)	0.99	0.92	0.86	0.85	0.92	0.90	0.87	0.87	1.12	1.06	0.99	0.99
ST	(9)	1.01	0.97	0.90	0.85	0.96	0.93	0.89	0.86	1.03	1.02	0.99	0.97
HT	(11)	0.81	0.83	0.78	0.67	0.93	0.99	0.99	0.72	1.06	**0.99**	0.95	0.80
PLSM	(10)	**0.62**	**0.61**	**0.69**	0.78	0.87	0.75	**0.74**	0.79	1.37	1.10	0.90	0.86
PCA	(8)	0.68	0.65	0.76	0.87	0.91	0.96	0.97	0.93	1.05	1.13	1.11	1.03
PCOVR	(8)	0.92	0.91	0.89	0.91	1.00	1.00	0.95	0.93	1.20	1.13	1.18	1.31
Selected number of factors													
LARSMB	(8)	1.10	0.98	0.94	0.87	1.08	0.93	0.90	0.90	1.07	1.07	0.96	0.97
STB	(10)	0.98	0.98	0.91	0.87	0.93	0.92	0.91	0.89	1.04	1.02	0.98	0.94
HTB	(10)	0.87	0.80	0.72	**0.61**	0.89	0.98	0.95	**0.67**	1.02	1.01	0.93	**0.76**
PLSMB	(9)	0.88	0.68	0.75	0.77	1.03	**0.72**	0.74	0.80	1.21	1.16	**0.89**	0.87
PCAB	(8)	0.67	0.66	0.77	0.80	**0.87**	0.97	0.96	0.88	1.10	1.16	1.10	1.01
PCOVRB	(7)	0.88	0.88	0.90	0.86	0.99	1.04	0.95	0.93	1.10	1.17	1.21	1.37

Notes: This table reports MSPE results for forecasts made for the period January 2004-December 2009.

method, based on the variables selected by Diebold et al. (2006), performs very well. Note that PCAG, PCOVRG and DRA have in common that a major part of the macro information is excluded from the model: PCAG and PCOVRG use only three out of the eleven groups identified by Stock and Watson (2002) (amounting to 58 out of 132 variables), whereas DRA uses only 3 out of 132 variables. On the other hand, the LARSU and PLSU procedures have all macro variables as inputs; their performance is not impressive, suggesting that these methods have difficulties in coping with the abundance of macro variables.

If we consider the multivariate methods with the number of factors fixed at three in the middle six rows of Table 2.3, we find that it is possible to include all macro information without obstructing forecast accuracy. PLSM, PCA and PCOVR perform better than their basic counterparts in many cases. The most notable example is the forecast for the short yield over a 6-month horizon: PCOVR results in a mean squared forecast error that is 12% (7 percentage points) smaller than PCOVRG, and 48% smaller than the no-macro benchmark.

Table 2.4 shows that the good performance of PLSM (and, to a lesser extent, PCA and PCOVR) continues in 2004-2009, especially for the short yield, where gains of over 30% are attained. We note that, as in the previous subperiod, LARSU and PLSU perform relatively poorly, and that among the multivariate methods, constructing factors is usually better than selecting variables.

Table 2.5: Average "best" number of factors, 1994-2009.

Horizon	1	3	6	12
LARSMB	1.96	2.75	2.47	2.10
STB	1.53	2.90	2.41	1.99
HTB	1.61	1.84	1.99	2.46
PLSMB	1.95	2.09	2.09	1.95
PCAB	1.96	2.84	2.93	2.86
PCOVRB	1.72	1.66	1.51	1.92

Notes: This table reports the number of factors selected by each method, averaged over the 192 estimation windows ending in 1994-2009.

Two of these multivariate methods have univariate counterparts: we may compare LARSM to LARSU and PLSM to PLSU. In general, the multivariate methods outperform their univariate counterparts. Thus, it is profitable to consider the yield factors β as a group rather than as three distinct variables. (Incidently, for PLS, Garthwaite (1994) argues that "in most situations, the univariate method is likely to give the better prediction equations." This does not seem to be the case in our application.)

We observe from the last six rows of Tables 2.2-2.4 that data-driven selection of the number of macro factors leads to further improvements (compared to fixing this number at three) in some cases, most notably for thresholding (STB and HTB) and principal component analysis (PCAB). For the other three methods, the predictive accuracy does not change much. As is shown in Table 2.5, fewer than three factors are generally required, suggesting that the first two factors constructed by these methods already summarize most relevant information. This result illustrates the importance of considering the forecast objective in factor construction: as the first two factors have sufficient predictive power, it is often better to neglect the third factor.

2.4.2 Significance tests

As a further evaluation of the results presented above, we formally test the null hypothesis of equal predictive accuracy for each pair of forecasting models using the Diebold and Mariano (1995) test.[5] For each pair of models, for each maturity and for each forecast horizon, we compare the series of forecast errors, but for the sake of brevity we only report the percentage of cases in which the null hypothesis is rejected.

Table 2.6 lists these rejection percentages for one-sided tests at 2.5% significance. The last number in the first row indicates that the no-macro benchmark is significantly outperformed in only 7% of the tests. However, this percentage varies considerably over the macro factor construction methods: the other numbers in the first row reveal that PLSM and PLSMB significantly outperform the benchmark in 34% and 60% of the tests, respectively. In fact, as the last number in the column headed "PLSMB" indicates, this method produces significantly better forecasts in as much as 27% of the cases, averaged over all seventeen competing methods. This result corroborates our finding that PLSMB often produces the most accurate forecasts.

[5] Although the no-macro benchmark model is nested in all other forecasting models that we compare, our use of a rolling estimation window allows us to use the Diebold-Mariano test also in this case; see Giacomini and White (2006).

Table 2.6: Outcomes of significance tests for pairwise comparisons of the forecast accuracy of alternative methods.

	NO	DRA	LARSU	PLSU	PCAG	PCOVRG	LARSM	ST	HT	PLSM	PCA	PCOVR	LARSMB	STB	HTB	PLSMB	PCAB	PCOVRB	Average
NO	×	.	.	.	.	3	.	.	.	34	9	.	.	.	.	60	9	.	7
DRA	4	×	.	.	.	1	.	.	.	4	6	.	.	.	.	.	7	1	1
LARSU	7	3	×	3	6	9	15	3	9	50	13	9	4	6	22	46	13	9	13
PLSU	1	.	.	×	1	.	.	.	.	9	21	.	.	.	1	16	19	.	4
PCAG	.	.	.	.	×	.	.	.	.	3	4	.	.	.	.	54	6	.	4
PCOVRG	.	.	.	.	.	×	.	.	.	1	6	.	.	.	.	12	6	.	1
LARSM	10	.	.	.	.	6	×	.	1	15	7	.	.	4	9	19	16	4	5
ST	25	.	.	1	6	13	.	×	3	38	24	3	.	46	31	69	26	10	17
HT	1	.	.	.	.	.	.	.	×	16	10	.	.	.	1	9	7	.	3
PLSM	4	1	.	.	1	3	.	.	1	×	4	1	.	1	4	.	4	3	2
PCA	.	.	.	.	.	.	.	.	.	.	×	.	.	.	.	18	.	.	1
PCOVR	6	.	.	.	.	1	.	.	.	1	21	×	.	.	.	22	37	4	5
LARSMB	16	13	.	1	16	21	21	.	4	32	25	7	×	25	21	31	26	22	17
STB	3	.	.	.	.	.	.	.	.	6	4	.	.	×	3	44	4	.	4
HTB	.	.	.	.	.	.	.	.	.	16	4	.	.	.	×	22	7	.	3
PLSMB	7	1	.	.	3	9	.	.	1	10	25	1	.	4	4	×	25	4	6
PCAB	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	4	×	.	0
PCOVRB	3	.	.	.	.	.	.	.	.	3	9	.	.	.	.	28	13	×	3
Average	5	1	0	0	2	4	2	0	1	14	11	1	0	5	6	27	13	3	×

Notes: This table reports rejection percentages of the Diebold-Mariano test, for all seventeen maturities, all four horizons, and all forecasts (1994-2009). For each cell in this table, $17 \times 4 = 68$ tests were performed, and 192 pairs of forecasts are compared in each test. We report the percentage of tests that indicate that the column-indexing method yields significantly better forecasts than the row-indexing method (at 2.5% significance, one-sided). For example, 60% of the tests comparing NO and PLSMB forecasts indicate that PLSMB forecasts are the better of the two, while 7% yield the opposite conclusion. Zeros have been replaced by dots for ease of legibility. The bottom row and the last column of the table report the average of the other entries in the respective column or row; for example, NO forecasts are found to be significantly better than other forecasts in 5% of the tests, and significantly worse in 7%.

Another conclusion we drew from Tables 2.2-2.4 was that selecting variables often performs worse than constructing factors. This result is echoed by the large numbers in the rows for LARS and thresholding methods in Table 2.6, and by the small numbers in the corresponding columns. Similarly, this table provides evidence for preferring LARSM to LARSU, PLSM to PLSU, and methods with selected numbers of factors (the methods with abbreviations ending with "B") to methods with the number of factors fixed at three.

In order to assess the differences in results between maturities, between forecast horizons, and over time, we present additional results in Appendix 2.B. To keep the exposition focused, we selected seven of the eighteen forecast methods for these tables. These include the methods found previously in the literature (NO, DRA, PCA), a variant with a selected number of factors

Table 2.7: Average weight of groups of macroeconomic variables, 1994-2009.

Group name	#	LARSU	PLSU	LARSM	ST	HT	PLSM	PCA	PCOVR
Real Output & Income	20	0.10	0.07	0.08	0.06	0.07	0.07	0.11	0.12
Employment & Hours	30	**0.26**	**0.52**	**0.32**	**0.29**	0.20	0.36	0.18	0.18
Housing	10	0.04	0.00	0.03	0.04	0.22	0.00	0.14	0.13
Orders & Inventories	10	0.20	0.03	0.21	0.21	0.08	0.02	0.06	0.06
Money & Credit	11	0.07	0.05	0.06	0.07	0.06	0.07	0.04	0.04
Stock Prices	4	0.03	0.00	0.06	0.05	0.02	0.00	0.01	0.01
Interest Rates & Spreads	17	0.17	0.00	0.18	0.18	**0.26**	0.00	0.19	0.19
Exchange Rates	5	0.05	0.00	0.01	0.03	0.02	0.01	0.01	0.01
Price Indices	21	0.05	0.30	0.02	0.04	0.07	**0.44**	**0.26**	**0.25**
Wages	3	0.00	0.02	0.00	0.00	0.00	0.03	0.00	0.00
Consumer Expectations	1	0.03	0.00	0.03	0.03	0.00	0.00	0.00	0.00

Notes: This table shows the relative importance of each of the eleven groups of macroeconomic variables in the factor construction methods. The numbers in the table are sums of squared weights, averaged over the 192 estimation windows ending in 1994-2009, and normalized to sum to one for each method. For each method, the largest weight is printed in bold. The column headed "#" lists the total number of variables per group.

(PCAB), and the partial least squares methods (PLSU, PLSM, PLSMB) that we have found to perform well. Comparing the results for different maturity groups in Table 2.B.1, we observe that for short horizons (up to a year) many methods perform better than both the no-macro benchmark and the selected variables from Diebold et al. (2006), with PLSM producing the best forecasts. For longer maturities, PLSMB clearly performs best, while other methods have more difficulties.

Table 2.B.2 shows that for long horizons (six or twelve months), PLSMB is the only method that consistently outperforms most others. Most remarkable is that the PLSMB forecasts are significantly better than the benchmark NO forecasts for all seventeen maturities at these horizons. For shorter horizons, PCA and PCAB are more useful. Finally, Table 2.B.3 illustrates that the no-macro benchmark is never beaten in 1994-1998, while PLSM and PLSMB provide much more accurate forecasts in the later subperiods. In the lower right panel of this table, we also report results for the crisis period (2008-2009) alone. Interestingly, all three variants of PLS provided much more accurate forecasts than any of the other methods for this period.

2.4.3 Further results

Table 2.7 indicates the relative importance that each method assigns to each group of macroeconomic variables. It shows the average squared weight of all variables in each group, normalized to sum to one for each method. For example, the first number in this table means that 10% of the variation in the LARSU factors comes from variables connected to real output and income. We observe that all methods assign relatively large weights to the "Employment and Hours" group; however, the best-performing methods (PLSM, PCA, PCOVR) put more weight on price indices, in line with Rudebusch and Wu (2008) and Koopman et al. (2011).

Given the relatively poor performance of both thresholding methods (ST and HT), it is perhaps surprising that the weights they assign to the eleven groups of variables are not very different from those assigned by other methods. This similarity means that the thresholding methods do not select variables from the wrong groups; rather, they pick "the wrong variables

from the right groups". As it is hard for thresholding methods to differentiate between variables with large in-sample correlations, it seems preferable to use methods that summarize all information in the data set, instead of selecting a smaller number of variables.

To gain more insight in the variability of the predictive performance over time, we plot a five-year rolling mean squared forecast error in Figure 2.4. Each point shows the mean squared forecast error measured over the past 60 months. We plot the errors only for two methods, one maturity, and one forecast horizon, but the picture is similar for the other cases. The best predictability of yields is realized around the turn of the century and in 2006-7, while forecast accuracy worsened with the advent of the financial crisis in 2008, especially for the no-macro benchmark.

Figure 2.5 illustrates the relative performance of univariate PLS, multivariate PLS, and PLS with data-driven selection of the number of factors. (Again, the picture for other methods is roughly similar.) The multivariate method almost always dominates the univariate method, and selection of the number of factors leads to additional gains. As we can see by comparing Figures 2.4 and 2.5, adding macroeconomic information helps most when forecasting without such information is difficult, particularly in 2003-5 and in 2009.

2.5 Conclusion

This study investigates various ways of incorporating macroeconomic information in the Nelson-Siegel framework for forecasting the yield curve. By comparing the predictive performance of these techniques with methods found in the recent literature, we find substantial evidence that the proposed alternative methods perform better in important cases. It is not easy to improve upon forecasts made without any macro information in stable times, like the years 1994-1998. When there is little volatility in yields, the dynamic Nelson-Siegel model without macro factors is sufficient for forecasting, although gains of more than 20% can still be attained in several cases. On the other hand, when volatility is relatively high (as in the years 1999-2009), macroeconomic variables are of substantive help in forecasting the yield curve. Gains of around 30% are attainable in this case. The largest gains in forecast accuracy are found in the 2008-9 crisis period.

The gains obtained by including macroeconomic information depend crucially on the way in which macro information is incorporated in the model. Selecting variables, as in Diebold et al. (2006), is useful only for forecasting yields of medium-term maturities (between one and five years). In all other cases, it is beneficial to extract information from a larger pool of data: from predefined groups of variables (using 58 variables in this study) for longer maturities, or from all available information (132 variables) for shorter maturities.

Principal component analysis (PCA), as advocated by De Pooter et al. (2007), yields acceptable results, as it generally improves forecast accuracy by around 20%, relative to the no-macro benchmark. Still, partial least squares (PLS) improves on PCA under most circumstances, often gaining another 5-10%. Other factor construction methods, based on selecting variables (using least angle regression, or a thresholding rule), are dominated by methods that use all available information, such as PLS.

Concerning the specific issues related to the use of macro factors as discussed in the introduction, we find several interesting conclusions. First, including factors extracted from a large

Figure 2.4: Rolling means of squared forecast errors for forecast horizon $h = 3$, maturity $\tau = 24$ months.

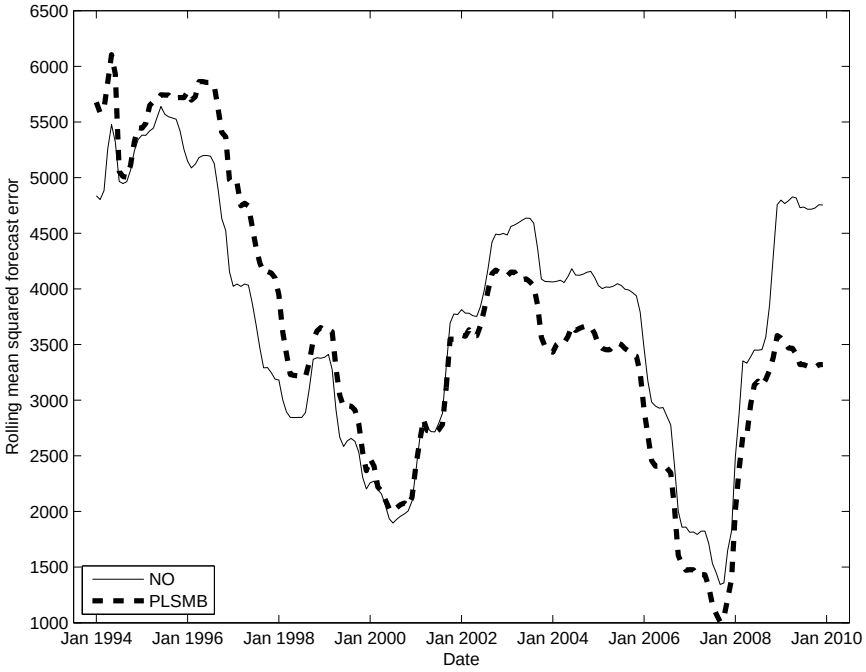


Figure 2.5: Rolling means of squared forecast errors for forecast horizon $h = 3$, maturity $\tau = 24$ months, relative to the NO benchmark.

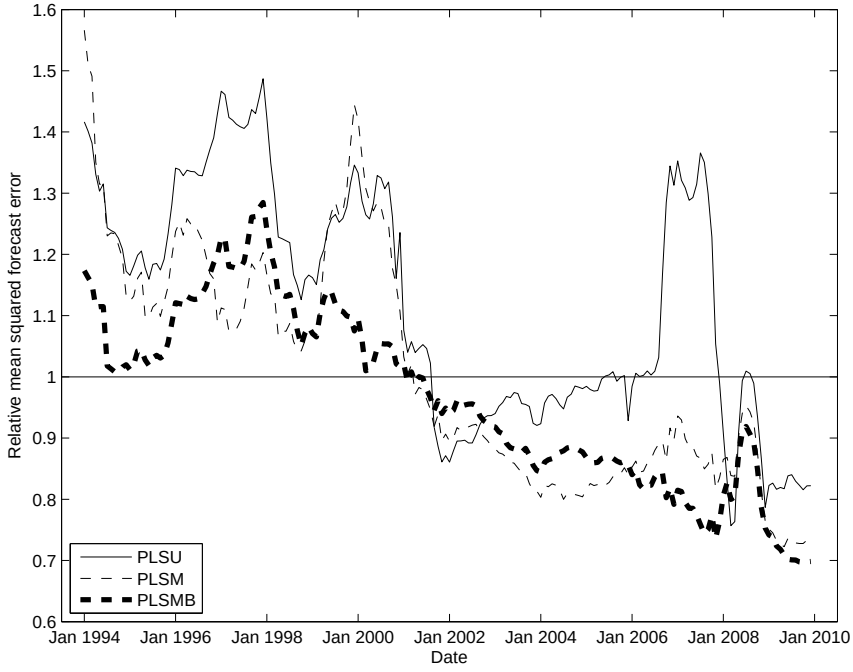
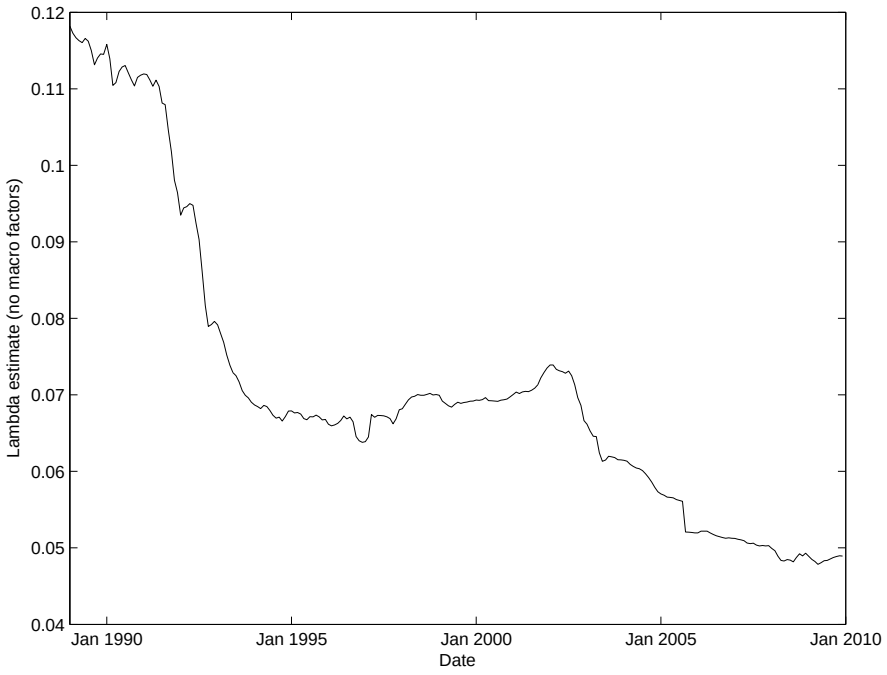


Figure 2.A.1: Estimates of the parameter λ in the FANS model.



panel of macro variables generally renders more accurate forecasts than selecting specific individual variables. Second, it is better to use the target variable in constructing macro factors than to ignore it. Third, it is better to construct a single set of factors for all three yield factors jointly than to treat each yield factor separately. Fourth, for long-term yields it is better to construct factors from groups of related macro variables, instead of one large pool of all available variables, but the opposite holds for short- and medium-term yields. Fifth, selecting the number of macro factors based on past predictive performance improves forecast accuracy for some methods.

To conclude, incorporating macroeconomic information appears to be most useful when yields are most volatile, and this information is best exploited by PLS. We expect that the factor construction rules studied here may also prove useful in other models and contexts. An interesting example is the use of augmented affine yield curve models, in the style of Mönch (2008) and the recent factor approach of Duffee (2011).

2.A Parameter estimates

This appendix illustrates some of the estimates of the parameters of the FANS model. Figure 2.A.1 shows a time series plot of the estimates of λ. All estimates are dated in the last month of the estimation window. We only show results for the model without macro factors; for the other models, the estimates are very similar. Typical values of λ found in the literature, such as Diebold and Li (2006), Diebold et al. (2006), and De Pooter et al. (2007), are around 0.06 to 0.08. Our estimates lie, on average, around these values. However, the λ estimates show great variability over time, casting some doubt on studies where λ is kept fixed over long time periods.

For the estimates of μ, our results are similar to those found in the papers cited above. The mean of the level parameter (the hypothetical yield of infinitely long maturity) is estimated to be around 600 basis points. The mean slope parameter is around -200, corresponding to an upward-sloping yield curve. Finally, the mean of the curvature parameter is small but positive, which indicates that the yield curve is slightly concave.

For the other parameters in the FANS model (see Equation (2.5)), our estimates are comparable to those reported by Diebold et al. (2006). The standard deviations σ_i are largest for the shortest maturities: typical values are around 15 basis points for short maturities, 2 to 4 for medium-term yields, and again slightly higher for longer maturities (over five years): roughly between 3 and 10 basis points. The estimated transition matrix A is always close to nonstationarity, with large own-lag coefficients for the level and slope parameters, illustrating the high persistence of yield curves. The block of A governing the macro-to-yields feedback relations contains values that are not negligible compared to the diagonal elements, confirming our economic intuition about such links and the result that these parameters are significantly nonzero, obtained by Diebold et al. (2006). Finally, the estimates of the state variance matrix Q generally show a pattern with much larger shocks to the curvature factor than to the other factors.

As an illustrative example, we provide the parameter estimates found by the PLSM method over the last estimation window, from January 2000 until December 2009. The vector of maturities τ is also shown, for ease of interpretation of the vector of standard deviations σ. Note that all elements of A and Q that are shown as "0" were actually imposed to be zero, as discussed in Section 2.2.2.

$$\lambda = 0.05, \qquad \mu = \begin{pmatrix} 518.87, & -237.92, & 119.91 \end{pmatrix}',$$

$$\begin{aligned} \tau = (\ & 3, & 6, & 9, & 12, & 15, & 18, & 21, & 24, & 30, \\ & & 36, & 48, & 60, & 72, & 84, & 96, & 108, & 120 \ )', \end{aligned}$$

$$\begin{aligned} \sigma = (\ & 34.38, & 15.72, & 9.85, & 5.49, & 3.76, & 2.44, & 2.40, & 3.66, & 2.82, \\ & & 3.18, & 5.55, & 3.39, & 5.71, & 4.40, & 1.07, & 4.57, & 9.09 \ )', \end{aligned}$$

$$A = \begin{pmatrix} .80 & 0 & 0 & .19 & .12 & .12 \\ 0 & .96 & 0 & -.20 & .01 & -.20 \\ 0 & 0 & .98 & -.22 & -.14 & -.13 \\ 0 & 0 & 0 & .04 & 0 & 0 \\ 0 & 0 & 0 & 0 & .73 & 0 \\ 0 & 0 & 0 & 0 & 0 & .24 \end{pmatrix},$$

$$Q = \begin{pmatrix} 478 & 0 & 0 & 0 & 0 & 0 \\ & 1090 & 0 & 0 & 0 & 0 \\ & & 8001 & 0 & 0 & 0 \\ & & & 9989 & 0 & 0 \\ & & & & 4572 & 0 \\ & & & & & 9427 \end{pmatrix}.$$

2.B Results of additional Diebold-Mariano tests

This appendix presents the results of Diebold-Mariano tests for equal forecast accuracy, split by maturity, by forecast horizon, and by subsample. These results are discussed in Section 2.4.2.

Table 2.B.1: Outcomes of significance tests for pairwise comparisons of the forecast accuracy of alternative methods, by maturity group.

	short (3 to 12 months)							*medium (15 to 60 months)*							*long (72 to 120 months)*						
	NO	DRA	PCA	PCAB	PLSU	PLSM	PLSMB	NO	DRA	PCA	PCAB	PLSU	PLSM	PLSMB	NO	DRA	PCA	PCAB	PLSU	PLSM	PLSMB
NO	×	.	31	25	.	88	75	×	.	3	6	.	28	59	×	.	.	.	.	.	50
DRA	.	×	19	19	.	19	.	3	×	3	6	.	.	.	10	×	.	.	.	.	.
PCA	.	.	×	.	.	.	.	.	.	×	.	.	.	9	.	.	×	.	.	.	45
PCAB	.	.	.	×	.	.	.	.	.	.	×	.	.	3	.	.	.	×	.	.	10
PLSU	.	.	38	38	×	38	.	.	.	13	16	×	.	6	5	.	20	10	×	.	45
PLSM	.	.	.	.	.	×	.	.	.	.	.	.	×	.	15	5	15	15	.	×	.
PLSMB	.	.	25	25	.	25	×	3	.	25	25	.	9	×	20	5	25	25	.	.	×

Notes: Each panel of this table is structured like Table 2.6. The data set contains four short, eight medium, and five long maturities. Thus, the percentages in this table are based on 16 (left panel), 32 (middle panel), or 20 tests (right panel).

Table 2.B.2: Outcomes of significance tests for pairwise comparisons of the forecast accuracy of alternative methods, by forecast horizon.

	one month ahead							*three months ahead*						
	NO	DRA	PCA	PCAB	PLSU	PLSM	PLSMB	NO	DRA	PCA	PCAB	PLSU	PLSM	PLSMB
NO	×	.	24	29	.	12	.	×	.	6	6	.	35	41
DRA	18	×	18	24	.	12	.	.	×	6	6	.	6	.
PCA	.	.	×	.	.	.	.	.	.	×	.	.	.	.
PCAB	.	.	.	×	.	.	.	.	.	.	×	.	.	.
PLSU	6	.	71	65	×	24	.	.	.	12	12	×	12	.
PLSM	18	6	18	18	.	×	.	.	.	.	.	.	×	.
PLSMB	29	6	100	100	.	41	×	.	.	.	.	.	.	×

	six months ahead							*twelve months ahead*						
	NO	DRA	PCA	PCAB	PLSU	PLSM	PLSMB	NO	DRA	PCA	PCAB	PLSU	PLSM	PLSMB
NO	×	.	6	.	.	47	100	×	.	.	.	.	41	100
DRA	.	×	.	.	.	.	.	.	×	.	.	.	.	.
PCA	.	.	×	.	.	.	47	.	.	×	.	.	.	24
PCAB	.	.	.	×	.	.	18	.	.	.	×	.	.	.
PLSU	.	.	.	.	×	.	41	.	.	.	.	×	.	24
PLSM	.	.	.	.	.	×	.	.	.	.	.	.	×	.
PLSMB	.	.	.	.	.	.	×	.	.	.	.	.	.	×

Notes: Each panel of this table is structured like Table 2.6. Each percentage in this table is based on 17 tests.

Table 2.B.3: Outcomes of significance tests for pairwise comparisons of the forecast accuracy of alternative methods, by subsample.

	first subsample (1994-1998)							*second subsample (1999-2003)*						
	NO	DRA	PCA	PCAB	PLSU	PLSM	PLSMB	NO	DRA	PCA	PCAB	PLSU	PLSM	PLSMB
NO	×	.	.	.	.	.	.	×	18	4	50	.	62	56
DRA	1	×	.	.	3	.	16	.	×	4	6	.	3	1
PCA	1	.	×	.	.	.	25	.	10	×	1	.	.	.
PCAB	1	.	.	×	.	.	28	.	7	.	×	.	.	.
PLSU	22	12	34	29	×	10	25	.	18	22	43	×	10	25
PLSM	15	6	18	25	.	×	.	.	7	.	1	.	×	.
PLSMB	28	7	24	24	.	.	×	.	.	1	1	.	.	×

	third subsample (2004-2009)							*crisis period (2008-2009)*						
	NO	DRA	PCA	PCAB	PLSU	PLSM	PLSMB	NO	DRA	PCA	PCAB	PLSU	PLSM	PLSMB
NO	×	.	3	3	6	35	38	×	7	4	3	44	50	49
DRA	.	×	3	3	.	4	.	.	×	3	3	4	18	10
PCA	.	.	×	.	.	10	9	.	4	×	.	16	16	10
PCAB	.	.	.	×	.	7	9	.	1	.	×	18	4	6
PLSU	.	.	3	1	×	3	.	.	.	.	.	×	.	.
PLSM	1	.	3	.	.	×	.	.	.	.	.	19	×	.
PLSMB	.	.	.	.	.	3	×	.	.	3	3	21	21	×

Notes: Each panel of this table is structured like Table 2.6. Each percentage in this table is based on 68 tests. The sample sizes for each test are 60 (both top panels), 72 (lower left panel), or 24 (lower right panel).

Chapter 3

Sparse and Robust Factor Modelling

This chapter is based on Croux and Exterkate (2011).

3.1 Introduction

Empirical researchers in a wide variety of fields face the problem of summarizing large data sets by a small number of representative factors, which can then be used for either descriptive or predictive purposes. In particular, the econometrics literature of the last decade contains successful applications of factor models to forecasting macroeconomic time series (Stock and Watson, 2002; Bai and Ng, 2008) and excess returns in stock and bond markets (Ludvigson and Ng, 2007, 2009).

Principal component analysis (PCA) is the classical tool for extracting such factors. In recent years, however, two major drawbacks of PCA have received attention. First, PCA lacks robustness to outliers. Even a very small proportion of data contamination results in inaccurate factors. This problem has been alleviated by explicitly downweighting such observations (Croux and Haesbroeck, 2000; Pison et al., 2003), by employing more robust loss functions than the usual sum of squares (De la Torre and Black, 2001), or by a combination of both approaches (Croux et al., 2003; Maronna and Yohai, 2008).

Second, in standard PCA all variables generally load on all extracted factors; that is, every original variable is represented as a linear combination of all factors. This feature leads to difficulties in giving an interpretation to the factors, as well as to a loss of degrees of freedom and thus to unnecessarily large estimation uncertainties. Penalized variants of standard PCA to overcome this problem have recently been developed by Jolliffe et al. (2003) and Witten et al. (2009), among others.

In this chapter, we propose a factor construction method that unifies both approaches, yielding robust factors with sparse loadings. Our procedure is a combination of the robust estimation methods from Maronna and Yohai (2008) and the penalization technique introduced by Witten et al. (2009). We provide a relatively simple alternating algorithm to solve the resulting optimization problem, and we document the good interpretability and forecasting properties of our method in a Monte Carlo study and in two empirical applications. Our first application concerns forecasting key U.S. macroeconomic variables, as in Stock and Watson (2002). The other application is microeconomic: we analyze the Boston housing data set from Harrison and

Rubinfeld (1978). The results show that ignoring the presence of outlying observations, which are often overlooked in empirical econometric studies, has important consequences for forecast accuracy.

To the best of our knowledge, our proposed method is the first to combine robustness and sparsity in the context of factor modelling. Moreover, while factors models are common in the macroeconomic forecasting literature, robustness issues are typically only considered in small sets of predictors (Fagiolo et al., 2008; Dehon et al., 2009; Bańbura et al., 2010). Sparsity is not commonly studied either, although a related approach using reduced-rank vector autoregressions was recently found to improve macroeconomic forecasts by Carriero et al. (2011).

The remainder of this article is structured as follows. We describe the methodology in Section 3.2 and test it in a simulation study in Section 3.3. Empirical applications to macroeconomic forecasting and to the Boston housing data set follow in Sections 3.4 and 3.5, respectively, and Section 3.6 concludes.

3.2 Methodology

3.2.1 Robust matrix approximation

We consider the problem of approximating an $n \times p$ matrix X by a rank-q matrix $\hat{X} = FA'$, where F has dimensions $n \times q$ and A is $p \times q$. The standard way to proceed is to apply principal component analysis (PCA), in which F and A are estimated by minimizing

$$Q_{L_2}(F, A; X) = \frac{1}{2n} \sum_{j=1}^{p} \sum_{i=1}^{n} (x_{ij} - f_i' a_j)^2, \tag{3.1}$$

where f_i and a_j denote rows of F and A, respectively. Although it is well-known that Q_{L_2} can be minimized using the singular value decomposition of X, we note that an alternating approach (due to Wold, 1966) is also possible. Given initial estimated of F and A, we iterate until convergence:

- Solve (3.1) for A by solving p ordinary least-squares (OLS) problems: the jth row is $a_j = (F'F)^{-1} F'x_j$, where x_j denotes the jth column of X.
- Solve (3.1) for F by solving n OLS problems: the ith row is $f_i = (A'A)^{-1} A'x_i$, where x_i denotes the ith row of X.

As all least-squares procedures, PCA is very sensitive to outlying observations (Maronna et al., 2006). A more robust alternative to (3.1) is to replace the sums of squared deviations by sums of absolute deviations; that is, to minimize

$$Q_{L_1}(F, A; X) = \frac{1}{2n} \sum_{j=1}^{p} \sum_{i=1}^{n} |x_{ij} - f_i' a_j|. \tag{3.2}$$

This L_1 minimization problem can be solved using a similar alternating algorithm as in the L_2 case, replacing OLS regressions by least absolute deviations (LAD) regressions. This procedure was advocated by Croux et al. (2003), who labelled it Robust Alternating L_1 Regressions (RAR).

Maronna and Yohai (2008) propose to replace the squared or absolute deviations by an even more robust error measure, the Tukey biweight loss function $\rho(r) = \min\left\{1, \left(1-(r/c)^2\right)^3\right\}$. This loss function is bounded, which makes it very robust to large outliers. The constant c is fixed at 3.4437, so that 85% efficiency at the normal distribution is attained. Because the Tukey loss function downweights large residuals, it is essential that the columns are appropriately scaled to decide what "large" means. Thus, for every variable j, let $\hat{\sigma}_j$ denote an estimate of the scale of the residuals $x_{ij} - f_i' a_j$, for $i = 1, 2, \ldots, n$. Then, Maronna and Yohai (2008) propose to minimize

$$Q_{\text{Tukey}}(F, A; X) = \frac{1}{2n} \sum_{j=1}^{p} \hat{\sigma}_j^2 \sum_{i=1}^{n} \rho\left(\frac{x_{ij} - f_i' a_j}{\hat{\sigma}_j}\right). \tag{3.3}$$

As a robust scale estimate, they consider the median absolute deviation

$$\hat{\sigma}_j = 1.4826 \underset{i}{\text{median}}\left\{|x_{ij} - f_i' a_j|\right\}. \tag{3.4}$$

If we would set $\rho(r) = r^2$, Criterion (3.3) would reduce to the PCA criterion (3.1). In order to be able to apply the alternating algorithm to minimize (3.3) for the Tukey loss function as well, we rewrite it as a weighted least squares (WLS) problem. Defining weights

$$w_{ij} = \left(\frac{x_{ij} - f_i' a_j}{\hat{\sigma}_j}\right)^{-2} \rho\left(\frac{x_{ij} - f_i' a_j}{\hat{\sigma}_j}\right), \tag{3.5}$$

the objective in equation (3.3) can be rewritten as

$$Q_{\text{Tukey}}(F, A; X) = \frac{1}{2n} \sum_{j=1}^{p} \sum_{i=1}^{n} w_{ij} \left(x_{ij} - f_i' a_j\right)^2. \tag{3.6}$$

This means that, given initial estimates of F, A, and the residual scales $\hat{\sigma}_j$, we can solve (3.3) by iterating the following scheme until convergence:

- Solve (3.6) for A by solving p WLS problems: the jth row is $a_j = \left(F' D_j F\right)^{-1} F' D_j x_j$, where D_j is a diagonal matrix containing $w_{1j}, w_{2j}, \ldots, w_{nj}$.
- Update $\hat{\sigma}_j$ for $j = 1, 2 \ldots, p$ using (3.4) and, hence, all weights w_{ij} using (3.5).
- Solve (3.6) for F by solving n WLS problems: the ith row is $f_i = \left(A' D_i A\right)^{-1} A' D_i x_i$, where D_i is a diagonal matrix containing $w_{i1}, w_{i2}, \ldots, w_{ip}$.
- Update the scale estimates $\hat{\sigma}_j$ and the weights w_{ij} again.

We shall consider all three different criteria introduced above. All columns of X are standardized before the estimation procedure. For the L_2 criterion (3.1) we standardize all columns to mean zero and variance one; for the L_1 criterion (3.2), to median zero and mean absolute deviation one; and for the Tukey criterion (3.3), to median zero and median absolute deviation one. Initial estimates for F and A are obtained as described by Maronna and Yohai (2008).

3.2.2 A sparsity condition

In factor-model terminology, the columns of F represent factors and A is the loading matrix. In order to improve the interpretability of the estimated factors, it may be desirable to impose a sparsity condition on the loading matrix; that is, to limit the number of nonzero factor loadings. In addition to improving interpretability, another interesting effect of such a condition is reducing the estimation uncertainty, which is an important consideration for forecasting. In the spirit of Witten et al. (2009), we implement this sparsity condition by adding an L_1 penalty to (3.1), (3.2), or (3.3): for some positive scalar λ, we aim to minimize

$$Q\left(F, A; X\right) + \lambda \sum_{j=1}^{p} \sum_{k=1}^{q} \left|a_{jk}\right|, \tag{3.7}$$

where Q denotes either Q_{L_2}, Q_{L_1}, or Q_{Tukey}. As it stands, objective (3.7) does not attain a minimum value. Although the linear subspace spanned by the columns of F is identified, we observe that for any candidate minimum point $\left(\hat{F}, \hat{A}\right)$, the equivalent factorization $\left(c\hat{F}, \frac{1}{c}\hat{A}\right)$ leads to a smaller objective value for any $c > 1$. To remove this unwanted feature, we restrict the magnitude of F by adding another penalty term to (3.7). As our purpose is not to impose sparsity on F, this additional term will be an L_2 penalty: we minimize

$$Q\left(F, A; X\right) + \lambda \sum_{j=1}^{p} \sum_{k=1}^{q} \left|a_{jk}\right| + \nu \sum_{i=1}^{n} \sum_{k=1}^{q} f_{ik}^2. \tag{3.8}$$

Finally, we note that Problem (3.8) is overparameterized: if the factorization $\left(\hat{F}, \hat{A}\right)$ solves (3.8) for the penalty parameters (λ^*, ν^*), then the equivalent factorization $\left(c\hat{F}, \frac{1}{c}\hat{A}\right)$ is a solution for $\left(c\lambda^*, \frac{\nu^*}{c^2}\right)$ for any $c > 0$. Therefore, we lose no generality in fixing either λ or ν at a specific positive value. We set $\nu = 1/\left(2n\right)$, so that only λ measures the degree of sparsity.

The alternating procedures in Section 3.2.1 can be adapted for problem (3.8). First, given F and (in the Tukey case) the weights w_{ij}, finding the jth row of A amounts to minimizing

$$Q\left(F, A; X\right) + \lambda \sum_{k=1}^{q} \left|a_{jk}\right|. \tag{3.9}$$

For the L_2 and Tukey criterion functions, we recognize (3.9) as a Lasso problem (Tibshirani, 1996), with regressand $\left(\sqrt{w_{ij}}\right) x_{ij}$ and regressors $\left(\sqrt{w_{ij}}\right) f_i$. Efficient algorithms to solve this problem are known; see Friedman et al. (2010). For the L_1 criterion, minimizing (3.9) is a LAD-Lasso problem (Wang et al., 2007).

Second, given A (and the weights), finding the ith row of F is equivalent to minimizing

$$Q\left(F, A; X\right) + \frac{1}{2n} \sum_{k=1}^{q} f_{ik}^2. \tag{3.10}$$

The ridge regression problem (3.10) can be solved analytically for the L_2 and Tukey criteria, resulting in

$$f_i = \left(A' D_i A + I\right)^{-1} A' D_i x_i, \tag{3.11}$$

where we set $D_i = I$ in the L_2 case. For the L_1 criterion, we use a standard numerical minimization routine.

3.2.3 Tuning parameters

The sparse and robust factor extraction procedure that we developed in Sections 3.2.1 and 3.2.2 is characterized by two tuning parameters; the number of factors (q) and the penalty parameter (λ). To specify values for q and λ, we minimize the Bayesian Information Criterion

$$BIC_{q,\lambda} = 2\sum_{j=1}^{p} \log \hat{\sigma}_{j;q,\lambda} + \mathrm{df}_{q,\lambda} \cdot \frac{\log n}{n}. \tag{3.12}$$

As argued by Zou et al. (2007), the "degrees of freedom" $\mathrm{df}_{q,\lambda}$ can be approximated by the number of nonzero entries in the estimated A. Further, we approximate the determinant of the residual covariance matrix by the product of scale estimates $\hat{\sigma}_j^2$, which are median absolute deviations (3.4) when using the Q_{Tukey} criterion, mean absolute deviations when using Q_{L_1}, and standard deviations when using Q_{L_2}. This amounts to discarding all covariances between columns of the residual matrix. We feel that this is a reasonable choice, as most of the correlation structure in X should be captured by the factors. Moreover, this procedure circumvents the nontrivial task of robustly estimating covariances.

3.3 Monte Carlo simulation

To evaluate the potential of the sparse robust factor extraction procedure described in Section 3.2, we assess its performance through a Monte Carlo study. As $n \approx p$ is typical for situations to which factor modelling is applied, we simulate data sets with $n = p = 100$. The number of latent factors will be $q = 2$.

We generate data from a factor model $X = FA' + E$. Here, the matrix A contains the factor loadings, and we impose that its true structure is sparse. The loading matrix has 100 rows and two columns:

$$A = \begin{pmatrix} 10 \text{ rows} & (+1, & +1) \\ 10 \text{ rows} & (+1, & -1) \\ 10 \text{ rows} & (-1, & +1) \\ 10 \text{ rows} & (-1, & -1) \\ 60 \text{ rows} & (\ \ 0, & \ \ 0) \end{pmatrix}. \tag{3.13}$$

For the 100×2 matrix of latent factors F and the 100×100 matrix of noise E, we consider the following four data-generating processes:

- *Normal:* the entries of F and E are independent draws from the $N(0, 1)$ distribution.
- *Heavy tails:* the entries of F are drawn from the $N(0, 1)$ distribution, those of E from Student's t distribution with two degrees of freedom.
- *Vertical outliers:* like the "Normal" DGP, but a random selection of 10% of the entries of E are replaced by the value 20.
- *Bad leverage rows:* like the "Normal" DGP, but a random selection of 10% of the rows of F are replaced by $(+20, +40)$, and the corresponding rows of E are replaced by $(-20, -40)\, A'$.

Table 3.1: Estimated structure of the loading matrix in the Monte Carlo simulations.

		Number of rows				Number of rows	
		correct	correct			correct	correct
DGP	Criterion	zero	nonzero	DGP	Criterion	zero	nonzero
Normal	L_2, $\lambda = 0$	0	40	Vertical outliers	L_2, $\lambda = 0$	0	40
	L_2, $\lambda > 0$	8.781	40		L_2, $\lambda > 0$	11.957	34.872
	L_1, $\lambda = 0$	0	40		L_1, $\lambda = 0$	0	40
	L_1, $\lambda > 0$	27.326	40		L_1, $\lambda > 0$	37.977	40
	Tukey, $\lambda = 0$	0	40		Tukey, $\lambda = 0$	0	40
	Tukey, $\lambda > 0$	6.377	40		Tukey, $\lambda > 0$	6.995	40
Heavy tails	L_2, $\lambda = 0$	0	40	Bad leverage rows	L_2, $\lambda = 0$	0	40
	L_2, $\lambda > 0$	11.314	39.860		L_2, $\lambda > 0$	5.266	40
	L_1, $\lambda = 0$	0	40		L_1, $\lambda = 0$	0	40
	L_1, $\lambda > 0$	29.902	40		L_1, $\lambda > 0$	30.791	40
	Tukey, $\lambda = 0$	0	40		Tukey, $\lambda = 0$	0	40
	Tukey, $\lambda > 0$	5.710	40		Tukey, $\lambda > 0$	14.603	40

Notes: This table reports average results over 1000 replications of each of the data-generating processes described in the text. The numbers indicate how many of the rows of the loading matrix A were correctly estimated to be zero/nonzero; the true loading matrix contains 60 zero and 40 nonzero rows.

Note the difference between the final two DGPs. If an observation is a vertical outlier, the latent factors behave normally but the observed variable is contaminated. In a bad leverage row, the factors behave abnormally but the observed variables are not informative about this fact.

In Tables 3.1 and 3.2 we report average results over 1000 simulation runs for each of these DGPs. We consider the L_2, L_1, and Tukey loss functions. For each of these, we report results using both the unpenalized criteria (3.1)-(3.3) and the penalized criterion (3.8). In the latter case, the penalty parameter λ is selected by minimizing the BIC (3.12) over the grid $\{0.0001, 0.001, 0.01, 0.1, 1\}$. We treat the true number of factors ($q = 2$) as known.

Table 3.1 reports on the structure of the estimated loading matrix A. Specifically, it shows how many of the 60 zero rows and 40 nonzero rows of the true A were correctly identified as zero or nonzero. From these results, it is clear that unpenalized estimation methods cannot succeed in exactly estimating zero loadings. The results for all penalized methods, on the other hand, are quite good: the penalized L_1 criterion correctly estimates more than half of the zero rows. Moreover, except for the penalized L_2 criterion, there are no false zero rows in the estimated loading matrix; thus, all variables that load on the factors are correctly identified.

An important application of factor models is forecasting a variable y, which is assumed to be driven by (a subset of) the same factors that drive X; say, $y = F\beta + \eta$, where η is noise. After $\hat{F}$ is obtained as above, we would estimate β using a form of regression (either ordinary least squares or a more robust variant) on the observations for which y_i is known, and then construct a forecast $\hat{y}_i = \hat{f}_i'\hat{\beta}$ for the remaining observations.

Instead of forecasting a specific linear combination of the factors, we consider the problem of forecasting *any* linear combination of the factors. The quality of such forecasts is assessed by computing the angle between the two-dimensional linear subspaces of $\mathbb{R}^{100}$ spanned by F and $\hat{F}$: the smaller this angle is, the more suitable $\hat{F}$ is for forecasting variables of the form $F\beta$.

Table 3.2: Summary statistics for the Monte Carlo simulations.

		Approximation of X			Angle
DGP	Criterion	RMSE	MnAE	MdAE	$(F, \hat{F})$
Normal	L_2, $\lambda = 0$	**0.975**	0.778	0.658	0.225
	L_2, $\lambda > 0$	0.979	0.781	0.660	**0.219**
	L_1, $\lambda = 0$	0.991	**0.770**	**0.640**	0.259
	L_1, $\lambda > 0$	0.995	0.778	0.650	0.256
	Tukey, $\lambda = 0$	0.981	0.778	0.653	0.233
	Tukey, $\lambda > 0$	0.984	0.780	0.655	0.228
Heavy tails	L_2, $\lambda = 0$	**3.423**	1.478	0.915	0.435
	L_2, $\lambda > 0$	3.436	1.453	0.886	0.412
	L_1, $\lambda = 0$	3.487	**1.383**	**0.793**	0.295
	L_1, $\lambda > 0$	3.493	1.395	0.804	**0.291**
	Tukey, $\lambda = 0$	3.480	1.396	0.816	0.326
	Tukey, $\lambda > 0$	3.483	1.396	0.813	0.311
Vertical outliers	L_2, $\lambda = 0$	**5.873**	3.638	2.182	1.314
	L_2, $\lambda > 0$	5.898	3.599	2.147	1.332
	L_1, $\lambda = 0$	6.316	**2.681**	**0.747**	**0.286**
	L_1, $\lambda > 0$	6.325	2.697	0.763	0.288
	Tukey, $\lambda = 0$	6.312	2.692	0.762	0.300
	Tukey, $\lambda > 0$	6.315	2.694	0.764	0.291
Bad leverage rows	L_2, $\lambda = 0$	1.169	0.867	0.671	1.264
	L_2, $\lambda > 0$	1.185	0.880	0.697	1.289
	L_1, $\lambda = 0$	0.944	**0.701**	**0.562**	0.344
	L_1, $\lambda > 0$	0.948	0.706	0.569	0.338
	Tukey, $\lambda = 0$	**0.936**	0.708	0.575	0.325
	Tukey, $\lambda > 0$	0.938	0.713	0.577	**0.320**

Notes: This table reports average results over 1000 replications of each of the data-generating processes described in the text. In the group of columns headed "Approximation of X", X is compared to $\hat{X} = \hat{F}\hat{A}'$; the root mean squared error and the mean and median absolute error are reported. In the rightmost column, we report the angle between the linear subspaces spanned by the columns of F and $\hat{F}$, in radians; for the "Bad leverage rows" DGP, the bad leverage rows are removed for this computation. For each DGP, the smallest RMSE, MeanAE, MedianAE and angle are printed in boldface.

The average values of this angle, again over 1000 simulation runs, are reported in the rightmost column of Table 3.2. Here, the value of using a penalized criterion function becomes apparent: in almost all cases, the angle between the true and estimated factors is smaller if a nonzero penalty is present. For the normal DGP, the different criterion functions yield similar results. For the other three DGPs, in which outliers are present, the L_2 factor estimates are much less accurate than those obtained using more robust criterion functions. An extreme example is the "bad leverage rows" DGP, for which angles between 1.2 and 1.3 radians are observed. As a right angle measures $\pi/2 \approx 1.571$ radians, it is clear that the L_2 criterion is severely misguided by the bad leverage rows. The Tukey criterion performs remarkably well in this case.

We also report results for the approximation of the data matrix X in Table 3.2, expressed as the root mean squared error (RMSE), mean absolute error (MnAE), and median absolute error

(MdAE). First, we notice that the in-sample approximation of X is most accurate without a penalty term, and using the L_2 or L_1 loss function, depending on whether the approximation quality is measured in squared or absolute errors. This result was to be expected, as the corresponding objective minimizes this error. The only exception to this rule is the "bad leverage rows" DGP, where the L_2 algorithm apparently failed to converge. We also note that little accuracy is lost when a positive penalty term λ is applied, and that the differences between loss functions in RMSEs are minor. Measured in mean or median absolute errors, the differences between the results from using the Tukey or L_1 loss are still small, but L_2 performs markedly worse in all DGPs except the normal.

3.4 Application: Macroeconomic forecasting

3.4.1 Data and forecast model

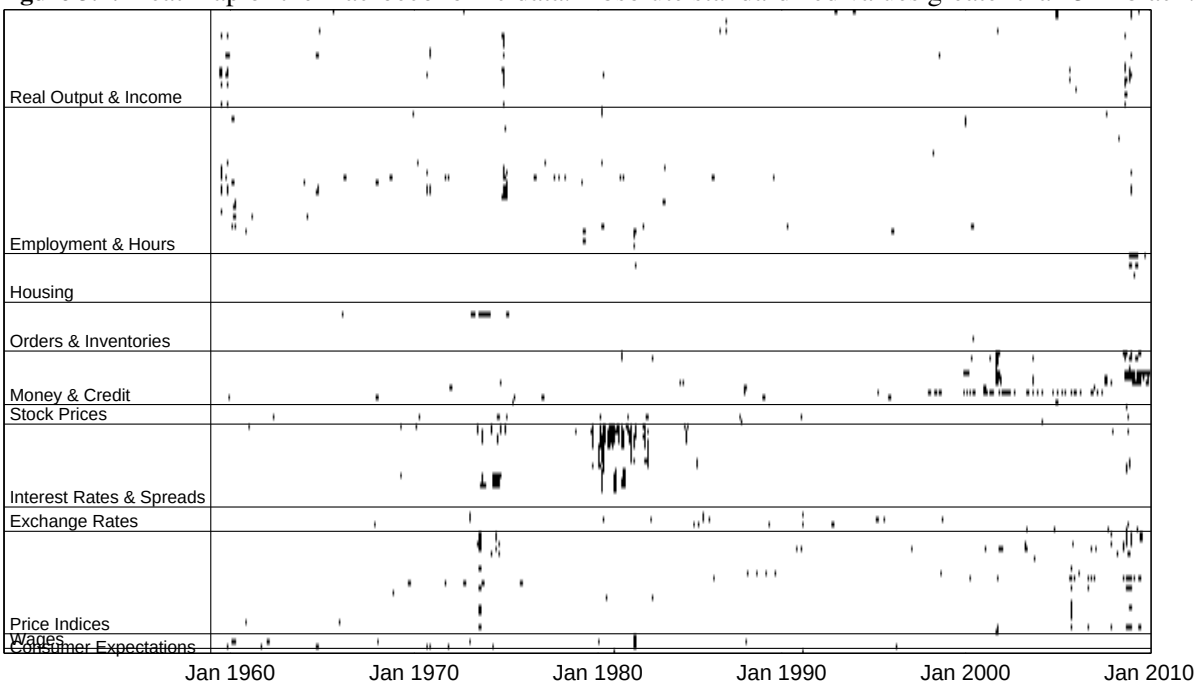
To evaluate the forecast performance of sparsely and robustly estimated factor models in an empirical application, we consider forecasting of four key macroeconomic variables. The data set consists of monthly observations on 132 U.S. macroeconomic variables, including various measures of production, consumption, income, sales, employment, monetary aggregates, prices, interest rates, and exchange rates. All series have been transformed to stationarity by taking logarithms and/or differences, as described in Stock and Watson (2002). They also define a partitioning of the data set into economically meaningful groups of related variables. We use an updated version of their data set, covering the period from January 1959 until (and including) January 2010, taken from Exterkate et al. (2011a). Some of the 132 time series start later than January 1959, while a few other variables have been discontinued before the end of the sample period. For each month under consideration, observations on at most five variables are missing.

A heat map of this data set is shown in Figure 3.1. For this figure, all time series were standardized to have median zero and median absolute deviation one. Each entry of the resulting matrix is shown in either black or white, depending on whether the standardized value is greater or smaller than five in absolute value. Time runs along the horizontal axis, and the different time series are organized in groups of related variables shown along the vertical axis. Despite the efforts to transform the data to near normality, a relatively large number of outliers shows up in various time series, mainly in interest rates series during the monetarist experiment in 1979-82, and in money and credit series in the recessions of 2000-01 and (especially) 2008-09. For this reason, we analyze these data using the robust methods outlined in Section 3.2.

We focus on forecasting four key measures of real economic activity: Industrial Production, Personal Income, Manufacturing & Trade Sales, and Employment. (The acronyms by which Stock and Watson (2002) refer to these series are `ip`, `gmyxpq`, `msmtq`, and `lhnag`, respectively.) For each of these variables, we produce out-of-sample forecasts for the annualized h-month percentage growth rate, which are computed as $y_{t+h}^h = (1200/h) \ln(v_{t+h}/v_t)$, where v_t is the untransformed observation on the level of each variable in month t. We consider growth rate forecasts for $h = 1, 3, 6$ and 12 months.

The most widely used approach to forecasting in this setup is the diffusion index (DI) approach of Stock and Watson (2002), who document its good performance for forecasting these four macroeconomic variables. The DI methodology extends the standard principal component

Figure 3.1: Heat map of the macroeconomic data. Absolute standardized values greater than 5 in black.



regression by including autoregressive lags as well as lags of the principal components in the forecast equation. Specifically, using ℓ_y autoregressive lags and ℓ_f lags of q factors, at time t, this "extended" principal-components method produces the forecast

$$\hat{y}^h_{t+h|t} = \hat{\alpha} + \sum_{s=0}^{\ell_y - 1} \hat{\beta}_s y^1_{t-s} + \sum_{s=0}^{\ell_f - 1} \sum_{k=1}^{q} \hat{\gamma}_{ks} \hat{f}_{k,t-s}. \tag{3.14}$$

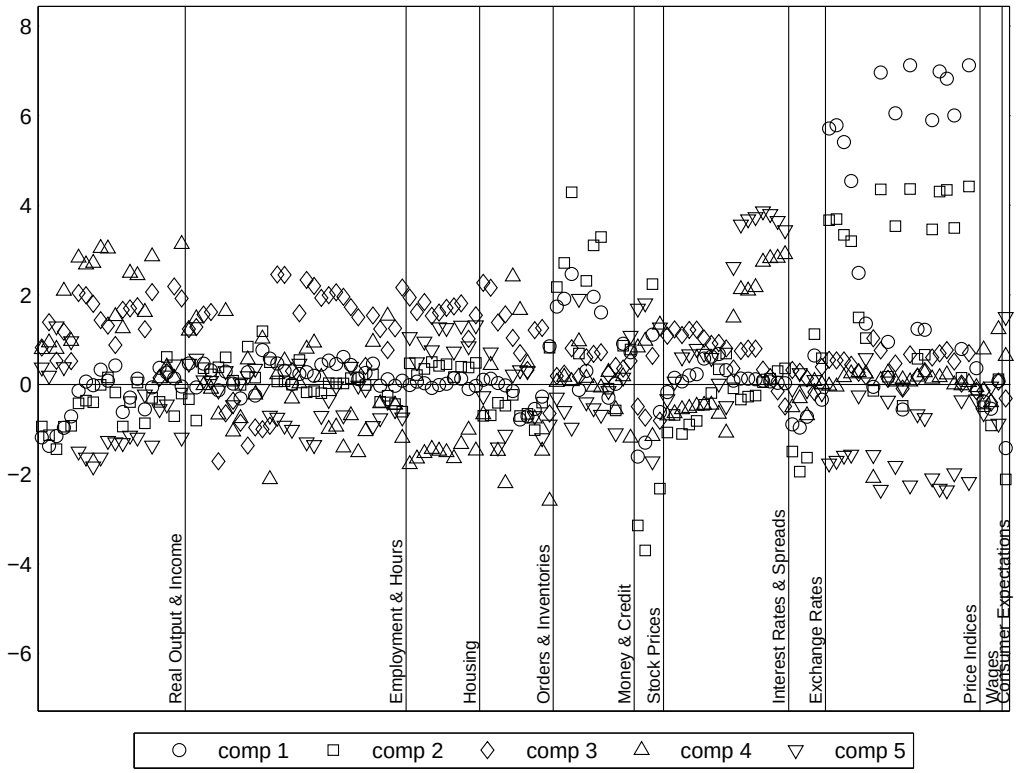
The lags of the dependent variable in Equation (3.14) are one-month growth rates, irrespective of the forecast horizon h, because using h-month growth rates for $h > 1$ would lead to highly correlated regressors. In Stock and Watson (2002), the factors $\hat{f}_{kt}$ are standard principal components extracted from all 132 predictor variables, and $\hat{\alpha}$, $\hat{\beta}_s$ and $\hat{\gamma}_{ks}$ are OLS estimates.

In this study, we retain the forecast equation (3.14), but we change the estimation methods for the factors $\hat{f}_{kt}$ and the regression coefficients. In addition to standard principal components, which corresponds to the L_2 criterion (3.1), we use the L_1 and Tukey variants of this criterion to estimate the factors. Moreover, we also estimate factors using the penalized criterion (3.8) for these three loss functions. After the $\hat{f}_{kt}$ have been obtained, we estimate the coefficient vector $\left(\alpha, \beta_0, \ldots, \beta_{\ell_y - 1}, \gamma_{10}, \ldots, \gamma_{q0}, \gamma_{11}, \ldots, \gamma_{q,\ell_f - 1}\right)'$ in (3.14) using either OLS, L_1 regression, or Tukey regression; the same loss function used to extract the factors. As the number of parameters is relatively small, we do not consider penalized regression estimation in this equation.

In each case, the lag lengths ℓ_y and ℓ_f, the number of factors q, and (if applicable) the penalty parameter λ are selected by minimizing the Bayesian Information Criterion (BIC). As our primary concern in this exercise is forecasting, we do not use expression (3.12) for the BIC, which measures how well the factors $\hat{F}$ fit X. Instead, we minimize

$$BIC_{\ell_y, \ell_f, q, \lambda} = 2 \log \hat{\sigma}_{\ell_y, \ell_f, q, \lambda} + (1 + \ell_y + \ell_f \cdot q) \frac{\log n}{n}, \tag{3.15}$$

Figure 3.2: Nonzero factor loadings for the macroeconomic data, L_2 criterion, $\lambda = 0$.



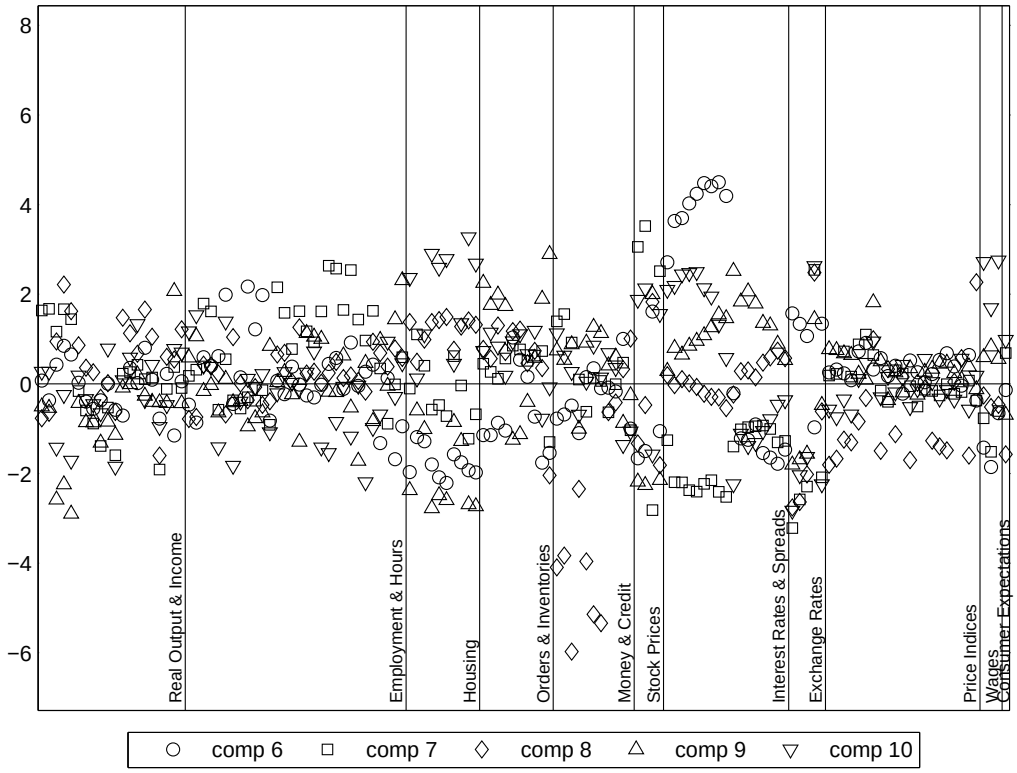
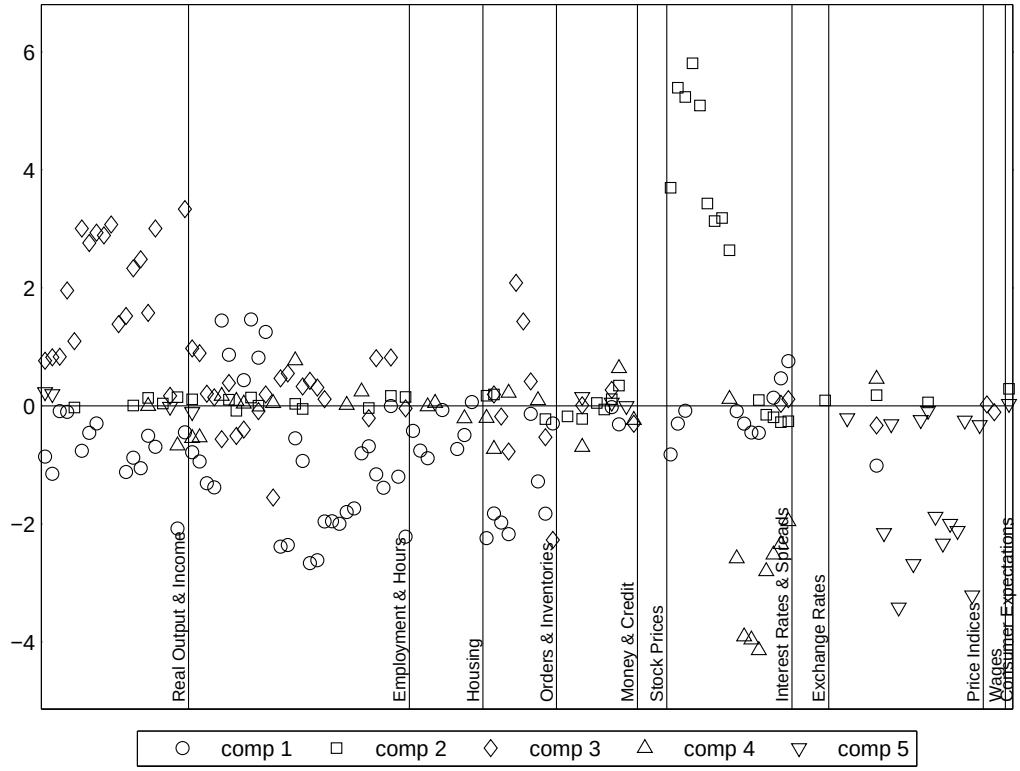


Figure 3.3: Nonzero factor loadings for the macroeconomic data, Tukey criterion, $\lambda = 0.1$.



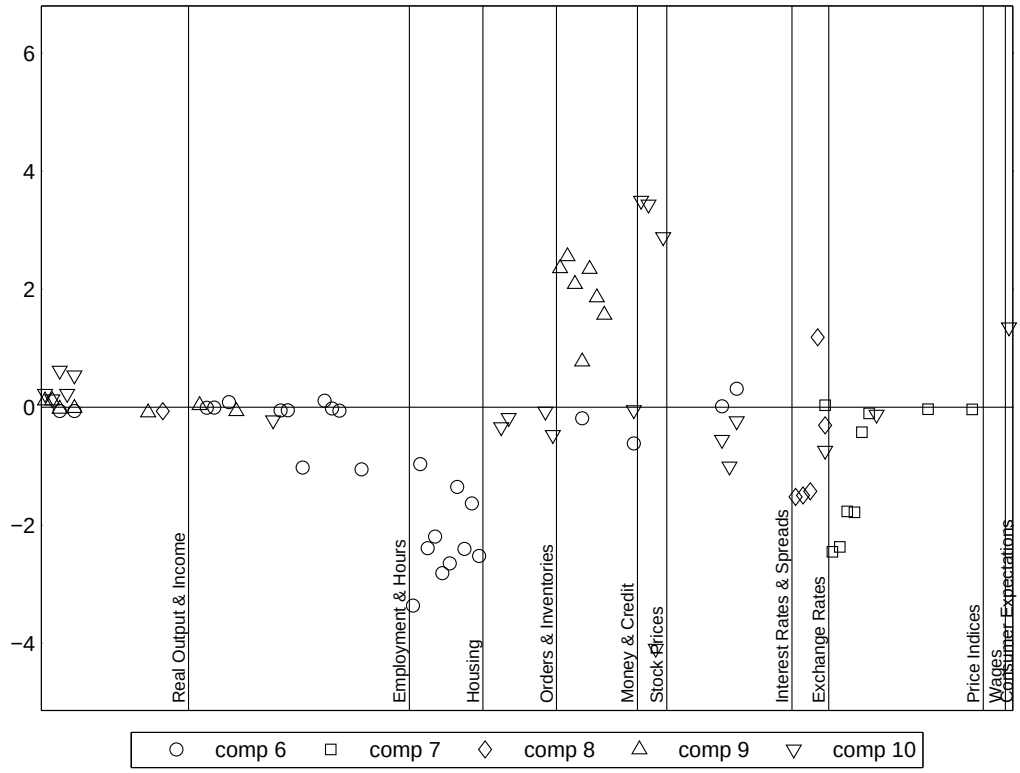


Table 3.3: Summary statistics for the in-sample fit in the macroeconomic data set.

	Nonzero	Approximation quality				Nonzero	Approximation quality		
Criterion	loadings	RMSE	MnAE	MdAE	Criterion	loadings	RMSE	MnAE	MdAE
L_2, $\lambda = 0$	1320	1.068	0.663	0.454	L_2, $\lambda = 0.1$	753	**1.061**	0.656	0.447
L_1, $\lambda = 0$	1320	1.246	**0.616**	**0.364**	L_1, $\lambda = 0.1$	842	1.258	0.622	0.365
Tukey, $\lambda = 0$	1320	1.081	0.626	0.422	Tukey, $\lambda = 0.1$	296	1.213	0.643	0.424

Notes: This table reports the number of nonzero entries in the estimated 132×10 loading matrix $\hat{A}$, as well as the root mean squared error and mean and median absolute error for the approximation $X \approx \hat{F}\hat{A}'$, after standardizing all variables to median zero and median absolute deviation one.

where $(1 + \ell_y + \ell_f \cdot q)$ is the number of parameters in Equation (3.14), and where $\hat{\sigma}_{\ell_y,\ell_f,q,\lambda}$ is an estimate of the scale of the residuals $y^h_{t+h} - \hat{y}^h_{t+h|t}$. As in Section 3.2.1, this scale estimate is either the standard deviation, the mean absolute deviation, or the median absolute deviation, depending on which loss function is used.

As Stock and Watson (2002) find that allowing for multiple lags of the factors does not substantially improve the forecasting performance, we fix $\ell_f = 1$. For the other parameters, we allow $0 \leq \ell_y \leq 6$, $0 \leq q \leq 4$, and $\log_{10}\lambda \in \{-4, -3, -2, -1, 0\}$. Note that $\ell_y = 0$ and $q = 0$ correspond to using no autoregressive information and no information from factors, respectively.

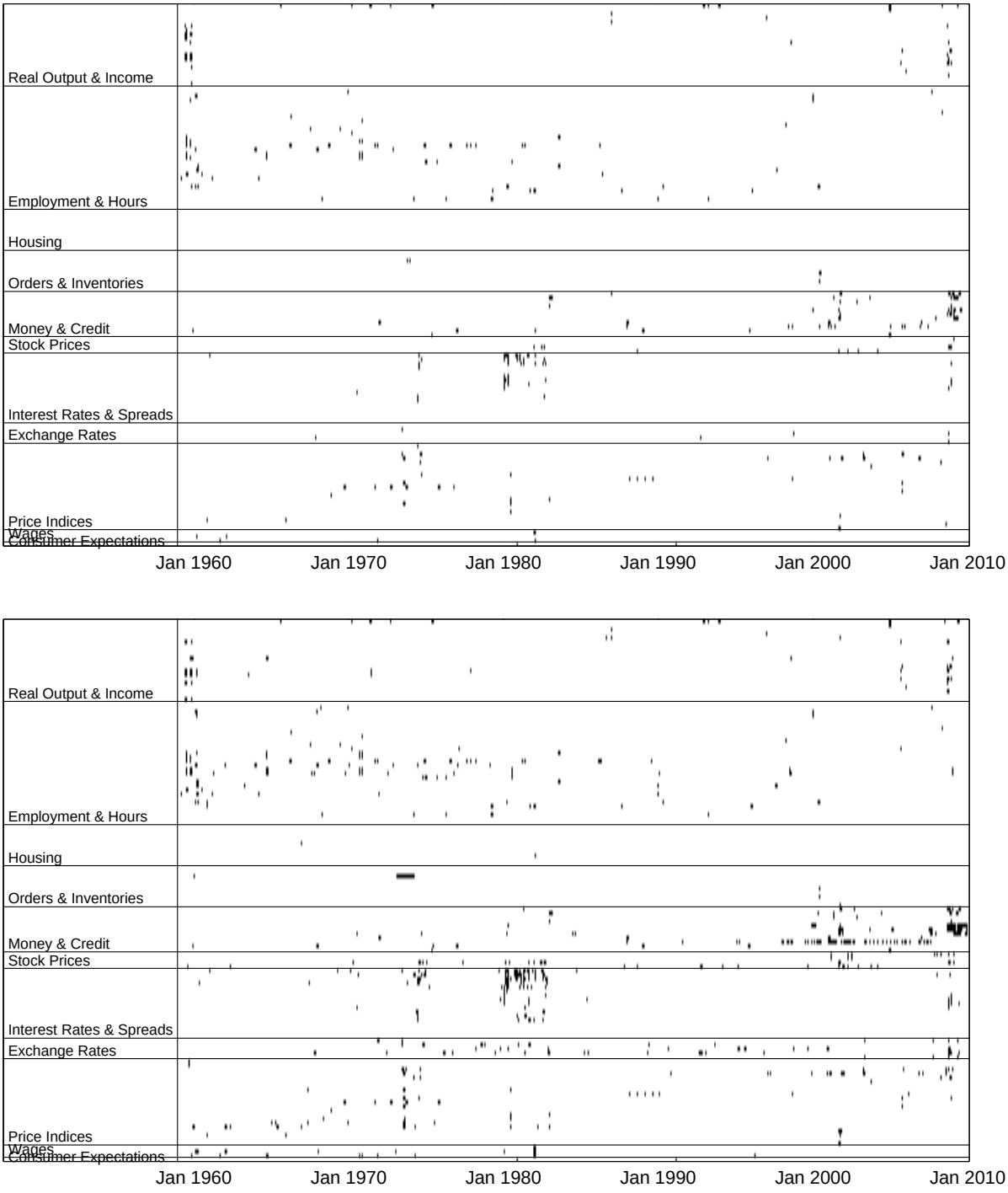
All models are estimated on rolling windows with a fixed length of 120 months, such that the first forecast is produced for the growth rate during the first h months of 1970. For each window, the tuning parameter values are re-selected and the regression coefficients are re-estimated. That is, all of the tuning parameters (ℓ_y, q, λ) are allowed to differ over time and across methods.

3.4.2 In-sample fit

Before turning to forecasting, we first consider the ability of estimated factor models to summarize the data set. To this end, we extracted $q = 10$ factors using each of the three different loss functions. We selected the penalization parameter λ by minimizing the BIC (3.12); in all three cases, $\lambda = 0.1$ was selected. From Table 3.3 we note that, as expected, using the L_2 criterion leads to the smallest mean squared error $\left|\left|X - \hat{X}\right|\right|_2^2$, while using the L_1 criterion leads to the smallest mean and median absolute error. For all error measures, the Tukey criterion yields results in between these extremes. As for the simulated data in Section 3.3, we observe that setting a positive penalty term does not substantially influence the in-sample goodness of fit.

This table also clearly shows the sparsity effect of choosing $\lambda > 0$, leading to as few as 296 (out of 1320) nonzero factor loadings for the Tukey criterion. Figures 3.2 and 3.3 show how this property aids in the interpretation of the factors. In these figures, the variable number is on the horizontal axis, with groups of variables separated by vertical lines. The factor loading is on the vertical axis, and exact zero loadings were omitted for legibility. The factor loadings obtained by standard PCA (Figure 3.2) are quite difficult to interpret. (Stock and Watson (2002) resort to computing pairwise correlations between constructed factors and original variables to alleviate this problem.) On the other hand, Figure 3.3 allows for a reasonable interpretation

Figure 3.4: Heat maps of the residuals for the macroeconomic data. Top: L_2 criterion, $\lambda = 0$. Bottom: Tukey, $\lambda = 0.1$.



of all ten factors extracted using the penalized Tukey criterion. For example, the pattern of nonzero loadings on the first component (circles in the top panel of Figure 3.3) suggests that this component is mostly associated with employment-related series. Continuing in this manner, we can assign labels to all ten factors as follows:

Table 3.4: Forecasting results for the macroeconomic data: Industrial Production and Personal Income.

Horizon	Criterion	RMSE	MnAE	MdAE	Horizon	Criterion	RMSE	MnAE	MdAE
	Industrial Production					Personal Income			
$h = 1$	$L_2, \lambda = 0$	8.258	5.917	4.395	$h = 1$	$L_2, \lambda = 0$	5.723	3.703	2.716
	$L_2, \lambda > 0$	8.368	5.961	4.357		$L_2, \lambda > 0$	5.932	3.706	2.786
	$L_1, \lambda = 0$	**7.889**	**5.717**	**4.161**		$L_1, \lambda = 0$	5.416	3.550	2.628
	$L_1, \lambda > 0$	8.023	5.742	4.238		$L_1, \lambda > 0$	5.430	3.563	2.587
	Tukey, $\lambda = 0$	7.944	5.720	4.322		Tukey, $\lambda = 0$	**5.390**	**3.505**	2.642
	Tukey, $\lambda > 0$	7.969	5.768	4.422		Tukey, $\lambda > 0$	5.414	3.537	**2.563**
$h = 3$	$L_2, \lambda = 0$	5.811	4.352	3.350	$h = 3$	$L_2, \lambda = 0$	3.369	2.521	1.945
	$L_2, \lambda > 0$	5.834	4.347	3.338		$L_2, \lambda > 0$	3.387	2.539	2.038
	$L_1, \lambda = 0$	5.792	4.305	3.455		$L_1, \lambda = 0$	3.403	2.541	**1.923**
	$L_1, \lambda > 0$	**5.750**	**4.300**	3.347		$L_1, \lambda > 0$	**3.364**	**2.513**	1.981
	Tukey, $\lambda = 0$	5.927	4.346	**3.171**		Tukey, $\lambda = 0$	3.515	2.575	1.997
	Tukey, $\lambda > 0$	5.927	4.351	3.243		Tukey, $\lambda > 0$	3.415	2.547	2.101
$h = 6$	$L_2, \lambda = 0$	4.933	3.682	2.760	$h = 6$	$L_2, \lambda = 0$	**2.775**	2.141	1.689
	$L_2, \lambda > 0$	4.875	**3.617**	2.756		$L_2, \lambda > 0$	2.792	2.148	1.728
	$L_1, \lambda = 0$	**4.867**	3.758	3.080		$L_1, \lambda = 0$	2.880	2.100	1.598
	$L_1, \lambda > 0$	4.925	3.802	3.115		$L_1, \lambda > 0$	2.841	**2.081**	**1.545**
	Tukey, $\lambda = 0$	5.281	3.820	**2.672**		Tukey, $\lambda = 0$	3.025	2.209	1.625
	Tukey, $\lambda > 0$	4.965	3.684	2.673		Tukey, $\lambda > 0$	3.011	2.235	1.697
$h = 12$	$L_2, \lambda = 0$	3.825	**2.769**	2.051	$h = 12$	$L_2, \lambda = 0$	2.486	1.957	1.557
	$L_2, \lambda > 0$	**3.821**	2.775	2.165		$L_2, \lambda > 0$	**2.447**	1.937	1.557
	$L_1, \lambda = 0$	4.073	3.002	2.265		$L_1, \lambda = 0$	2.537	1.935	1.465
	$L_1, \lambda > 0$	3.996	2.947	2.243		$L_1, \lambda > 0$	2.487	**1.920**	**1.455**
	Tukey, $\lambda = 0$	4.001	2.862	**2.043**		Tukey, $\lambda = 0$	2.566	1.994	1.551
	Tukey, $\lambda > 0$	3.999	2.889	2.125		Tukey, $\lambda > 0$	2.534	1.951	1.546

Notes: This table reports the root mean squared forecast error and mean and median absolute forecast error for the macroeconomic forecast example. For each series, the smallest RMSE, MnAE, and MdAE are printed in boldface.

1. employment;
2. interest rates;
3. production;
4. interest rate spreads;
5. consumer price inflation;
6. housing;
7. producer price inflation;
8. exchange rates;
9. monetary policy; and
10. stock prices.

Recalling the large number of outliers in the data, as visualized in the heat map in Figure 3.1, it is of interest to repeat this outlier detection exercise for the residuals after ten factors have been extracted. The corresponding heat maps are shown in Figure 3.4. We observe that the L_2 factor extraction procedure is severely influenced by the outlying observations identified in Figure 3.1: many of the outliers are no longer present in the residuals, which means that the extracted factors fit these observations well. This result continues to hold if a positive penalty λ is selected. On the other hand, the residuals from the Tukey criterion exhibit a similar outlier pattern to the original data. In this criterion, large outliers are downweighted, so they have less impact on the factor estimates. Similar results are obtained using the L_1 criterion (not shown).

Table 3.5: Forecasting results for the macro data set: Manufacturing & Trade Sales and Employment.

Horizon	Criterion	RMSE	MnAE	MdAE	Horizon	Criterion	RMSE	MnAE	MdAE
	Manufacturing & Trade Sales					Employment			
$h = 1$	$L_2, \lambda = 0$	**11.463**	**8.680**	7.040	$h = 1$	$L_2, \lambda = 0$	**2.980**	2.227	**1.708**
	$L_2, \lambda > 0$	11.540	8.774	6.990		$L_2, \lambda > 0$	3.045	2.277	1.779
	$L_1, \lambda = 0$	11.779	8.963	7.246		$L_1, \lambda = 0$	2.991	**2.226**	1.710
	$L_1, \lambda > 0$	11.819	9.021	7.449		$L_1, \lambda > 0$	2.983	2.229	1.771
	Tukey, $\lambda = 0$	12.072	9.028	6.795		Tukey, $\lambda = 0$	3.072	2.307	1.778
	Tukey, $\lambda > 0$	12.108	9.066	**6.669**		Tukey, $\lambda > 0$	3.071	2.293	1.761
$h = 3$	$L_2, \lambda = 0$	6.205	4.689	3.648	$h = 3$	$L_2, \lambda = 0$	1.765	1.322	**0.984**
	$L_2, \lambda > 0$	6.363	4.781	3.747		$L_2, \lambda > 0$	1.773	1.336	1.025
	$L_1, \lambda = 0$	6.201	4.719	3.787		$L_1, \lambda = 0$	**1.733**	**1.296**	0.987
	$L_1, \lambda > 0$	**6.074**	**4.660**	3.705		$L_1, \lambda > 0$	1.757	1.323	1.015
	Tukey, $\lambda = 0$	6.297	4.763	**3.625**		Tukey, $\lambda = 0$	1.770	1.343	1.044
	Tukey, $\lambda > 0$	6.345	4.802	3.672		Tukey, $\lambda > 0$	1.780	1.338	1.038
$h = 6$	$L_2, \lambda = 0$	**4.663**	**3.406**	2.509	$h = 6$	$L_2, \lambda = 0$	**1.422**	**1.076**	**0.820**
	$L_2, \lambda > 0$	4.757	3.448	2.567		$L_2, \lambda > 0$	1.435	1.093	0.827
	$L_1, \lambda = 0$	5.127	3.695	2.605		$L_1, \lambda = 0$	1.456	1.108	0.837
	$L_1, \lambda > 0$	4.920	3.603	2.728		$L_1, \lambda > 0$	1.444	1.107	0.845
	Tukey, $\lambda = 0$	4.922	3.538	**2.367**		Tukey, $\lambda = 0$	1.524	1.143	0.823
	Tukey, $\lambda > 0$	4.868	3.494	2.467		Tukey, $\lambda > 0$	1.525	1.137	0.839
$h = 12$	$L_2, \lambda = 0$	3.664	2.613	**1.931**	$h = 12$	$L_2, \lambda = 0$	1.235	0.932	0.685
	$L_2, \lambda > 0$	**3.557**	**2.607**	2.016		$L_2, \lambda > 0$	**1.194**	**0.904**	**0.671**
	$L_1, \lambda = 0$	3.740	2.734	2.110		$L_1, \lambda = 0$	1.228	0.913	0.710
	$L_1, \lambda > 0$	3.714	2.731	2.156		$L_1, \lambda > 0$	1.238	0.914	0.711
	Tukey, $\lambda = 0$	3.630	2.679	2.121		Tukey, $\lambda = 0$	1.294	0.979	0.747
	Tukey, $\lambda > 0$	3.579	2.656	2.013		Tukey, $\lambda > 0$	1.290	0.978	0.737

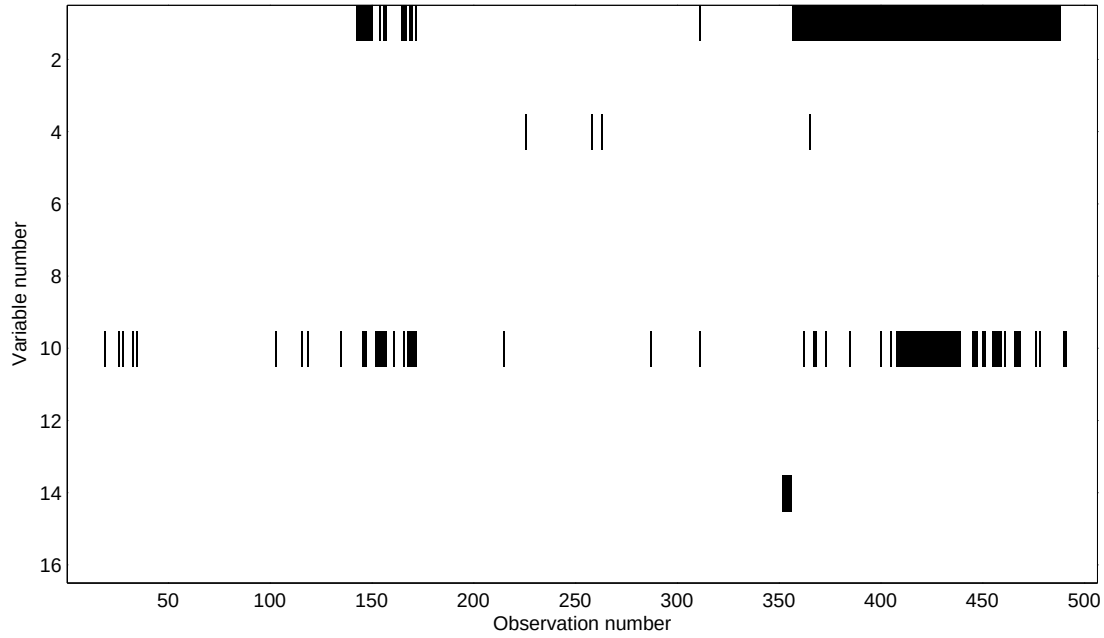
3.4.3 Forecasting results

Next, we focus on forecasting four key macroeconomic series, as described above. The results are summarized in Tables 3.4 and 3.5. For Industrial Production and Personal Income (Table 3.4), we find that our sparse and robust methods often outperform the benchmark of standard PCA. For horizons shorter than a year, the more robust Tukey and L_1 criterions generally lead to better forecasts than the standard L_2 criterion, irrespective of which measure we use to evaluate the performance. Thus, the lack of robustness in PCA that we observed negatively affects the forecasting performance, and more robust criterion functions remedy this situation.

For the easier task of forecasting annual growth rates ($h = 12$), we find that the L_2 criterion does lead to adequate forecasts. In this case, however, imposing a sparsity constraint improves the forecast quality: a relatively simple task is best performed using relatively simple models.

The results for the other two series, Manufacturing & Trade Sales and Employment, are shown in Table 3.5. It is very hard to improve on standard PCA forecasts for these series, a finding that was also documented by Exterkate et al. (2011a). Nevertheless, our result that sparse modelling leads to better forecasts for annual growth rates also applies here.

Figure 3.5: Heat map of the Boston housing data. Absolute standardized values greater than 5 in black.



3.5 Application: Boston Housing data

3.5.1 Data and forecast model

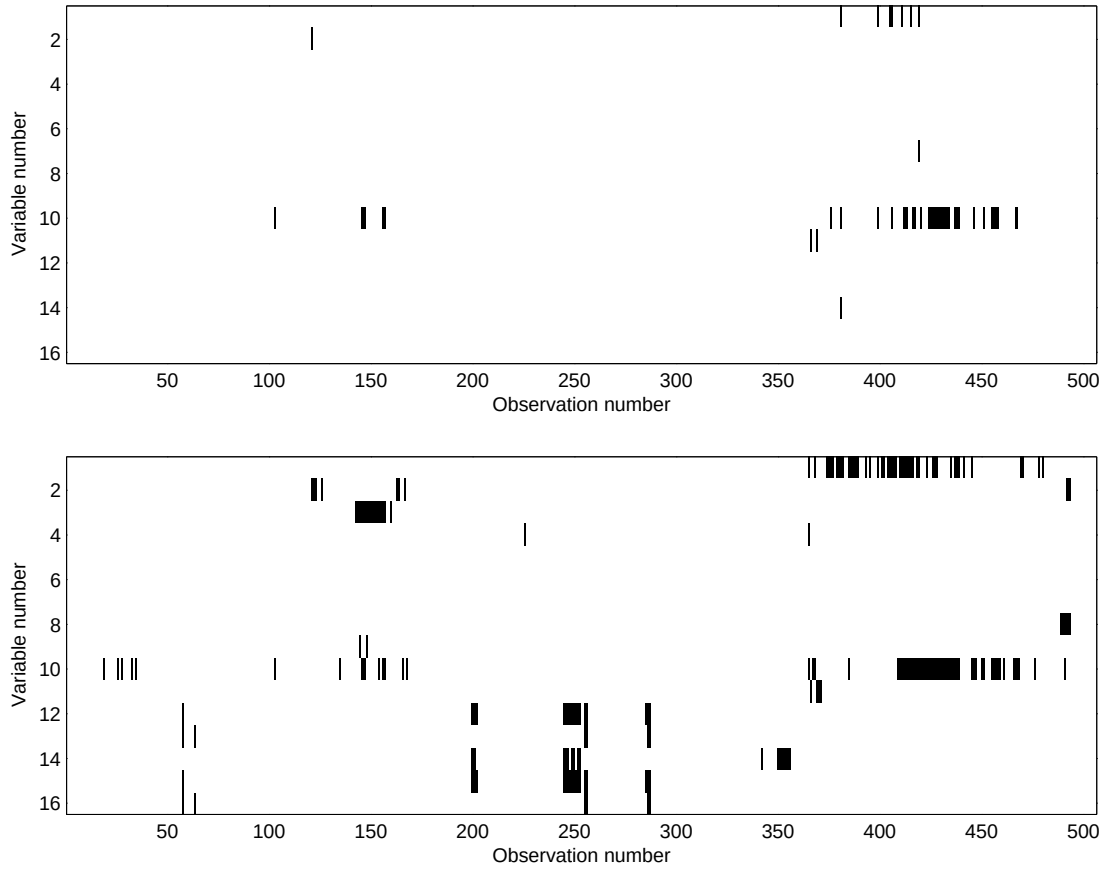
The Boston Housing data set, originating with Harrison and Rubinfeld (1978), has been extensively analyzed in the robust statistics literature. We use the corrected version of the data set by Pace and Gilley (1997). The data set contains various characteristics of houses, demographics, air pollution, and geographical details on 506 census tracts in or nearby Boston. The objective is to relate the median house price to the other characteristics, and our model will be inspired by the one in Pace and Gilley (1997):

$$\log \text{ Price} = \begin{pmatrix} 1, & \text{ZN}, \text{ CHAS}, \text{ CRIM}, \text{ INDUS}, \text{ NOX}^2, \text{ RM}^2, \text{ AGE}, \ldots \\ & \log \text{ DIS}, \log \text{ RAD}, \text{ TAX}, \text{ PTRATIO}, \text{ B}, \log \text{ LSTAT}, \ldots \\ & \text{LON}, \text{ LAT}, \text{ LON} \times \text{LAT}, \text{ LON}^2, \text{ LAT}^2 \end{pmatrix} \beta + \varepsilon, \tag{3.16}$$

where the regressors denote the proportion of area zoned with large lots (ZN), a dummy for a location contiguous to the Charles River (CHAS), the crime rate (CRIM), the proportion of nonretail business areas (INDUS), levels of nitrogen oxides (NOX), the average number of rooms (RM), the proportion of structures built before 1940 (AGE), weighted distances to the employment centers (DIS), an index of accessibility (RAD), the property tax rate (TAX), the pupil/teachter ration (PTRATIO), the black population proportion (B), the lower status population proportion (LSTAT), and the geographical longitude (LON) and latitude (LAT).

Before applying the methods of Section 3.2 to this data set, we remove two variables for which the median absolute deviation is zero; namely, the proportion of large lots (ZN) and the

Figure 3.6: Heat maps of the residuals for the Boston housing data. Top: L_2 criterion, $\lambda = 0$. Bottom: Tukey, $\lambda = 0.010$.



Charles River dummy (CHAS). As we scale all variables by dividing by their median absolute deviations before extracting factors using the Tukey criterion, we cannot handle these variables in our algorithm. As a compromise, we estimate the model

$$\log \text{ Price} = \alpha + \beta_1 \text{ ZN} + \beta_2 \text{ CHAS} + F\gamma + \varepsilon, \tag{3.17}$$

where we extract the factors F from the remaining right-hand-side variables in Equation (3.16).

A heat map of the data is shown in Figure 3.5, with the variables ordered as in Equation (3.16), starting with CRIM. Thus, the variables containing many outlying observations can be identified as the crime rate (CRIM, variable 1) and the proportion of black population (B, variable 10). The groups of observations at which these outliers occur correspond to locations in the cities of Cambridge (around observation 150) and Boston (around observation 430).

Our forecasting procedure is as follows. We first extract the factors F from the full data set. Then, we estimate Model (3.17) on a random selection of 80% of the 506 observations, and we decide on the number of factors, the value of λ, and whether or not to include ZN and/or CHAS in the model by minimizing a Bayesian Information Criterion similar to the one in Equation (3.15). The selected model is then used to forecast the prices for the 20% of the observations that were left out in the estimation, and we repeat the procedure five times, ensuring that each observation is being predicted exactly once.

Table 3.6: Summary statistics for the in-sample fit in the Boston housing data set.

	Approximation quality				Approximation quality		
Criterion	RMSE	MnAE	MdAE	Criterion	RMSE	MnAE	MdAE
L_2, $\lambda = 0$, $q = 5$	4.683	1.159	0.186	L_2, $\lambda = 0.001$, $q = 5$	4.687	1.153	0.185
L_1, $\lambda = 0$, $q = 5$	6.219	0.915	**0.039**	L_1, $\lambda = 0.100$, $q = 5$	6.600	0.985	0.056
Tukey, $\lambda = 0$, $q = 5$	**1.619**	**0.391**	0.204	Tukey, $\lambda = 0.010$, $q = 5$	5.496	0.748	0.172

Notes: This table reports the selected numbers of factors and penalization parameters, as well as the root mean squared error and mean and median absolute error, after standardizing all variables to median zero and median absolute deviation one. The smallest errors are printed in boldface.

Table 3.7: Forecasting results for the Boston housing data set.

Criterion	RMSE	MnAE	MdAE	Criterion	RMSE	MnAE	MdAE
L_2, $\lambda = 0$	0.224	0.148	0.100	L_2, $\lambda > 0$	**0.217**	**0.142**	**0.097**
L_1, $\lambda = 0$	0.240	0.156	0.097	L_1, $\lambda > 0$	0.241	0.157	0.100
Tukey, $\lambda = 0$	0.269	0.191	0.130	Tukey, $\lambda > 0$	0.233	0.152	0.100

Notes: This table reports the root mean squared error and mean and median absolute error in forecasting the logarithm of the median house price. The smallest errors are printed in boldface.

3.5.2 In-sample fit

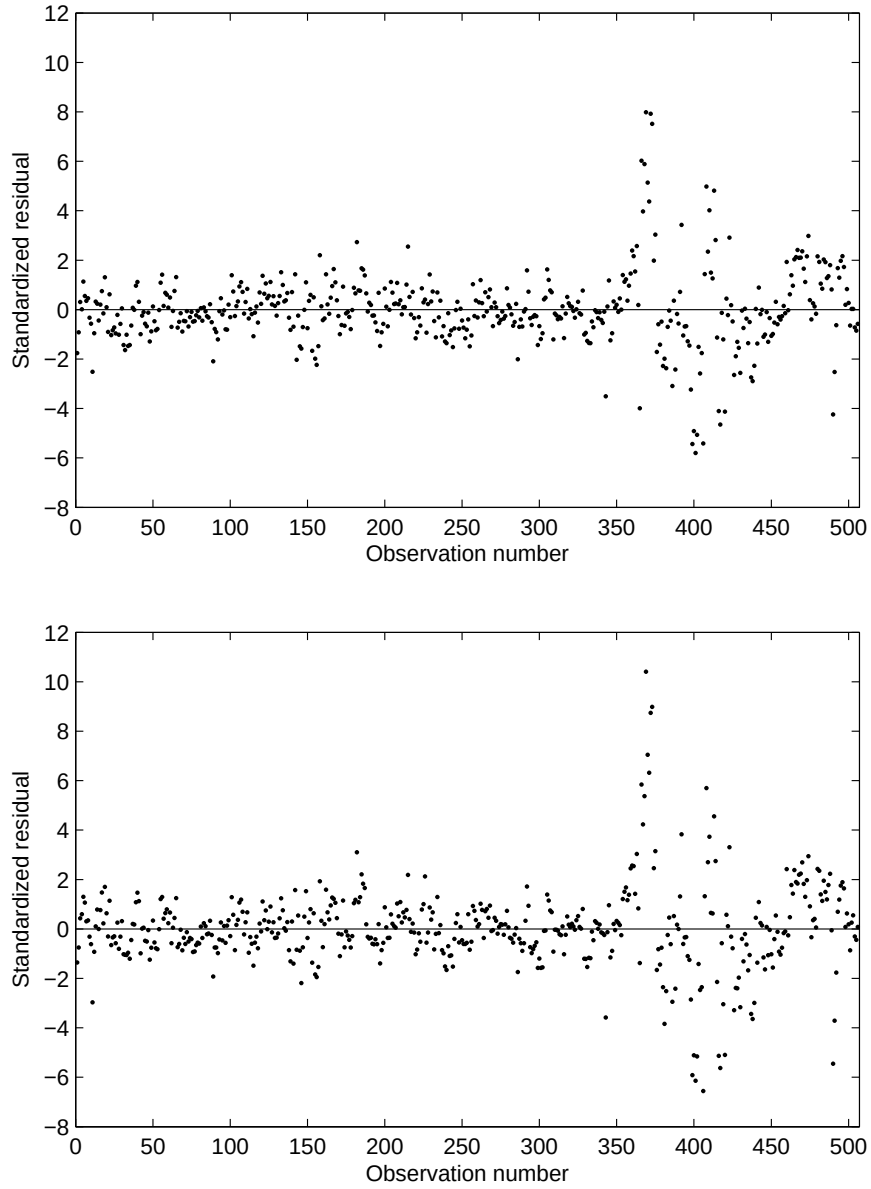
Again, we first consider the in-sample fit, selecting the number of factors and the penalization parameter by minimizing the BIC in Equation (3.12). Summary statistics are shown in Table 3.6. Note that in all cases, the maximum number of five components is selected. Given that the data set contains only sixteen variables, it seems undesirable to extract even more factors. As in the other data sets, we find that allowing for a positive penalization parameter λ does not substantially worsen the in-sample fit — except, in this case, for the Tukey criterion, which performs extremely well with $\lambda = 0$.

Heat maps of the residuals are shown in Figure 3.6. The heat map for standard PCA residuals (top panel) indicates that many of the outlying observations that were identified from Figure 3.5 are fitted by the factor structure. This is the well-known effect of least-squares methods being sensitive to large outliers. On the other hand, most of the outliers are still present in the residuals from penalized Tukey factor extraction. In fact, new groups of outliers are detected, especially in the variables numbered 12 (geographical longitude), 14 (longitude times latitude), and 15 (longitude squared). It turns out that the outliers correspond to locations relatively far west of Boston. The Tukey criterion tries to fit *most* of the data, rather than *all* of the data.

3.5.3 Forecasting results

Table 3.7 summarizes the forecasting results. We observe that even in this relatively small data set, penalized estimation leads to better forecasts. On the other hand, despite the large number of outliers identified in Figure 3.5, robust methods perform (somewhat) worse than L_2 estimation. Closer inspection of the data reveals that the house price has a high correlation with the two variables containing most outliers. Thus, in this case, it is undesirable to downweight the outlying observations, but the results are not too heavily affected by this fact.

Figure 3.7: Standardized residuals for the house price equation (3.17). Top: L_2 criterion, $\lambda = 0.1$, $q = 5$. Bottom: Tukey criterion, $\lambda = 0.1$, $q = 4$.



To illustrate this result, we re-estimated Equation (3.17) over the full sample, again selecting the number of factors and the value of λ by minimizing the BIC. The residuals, standardized to median absolute deviation one, are plotted in Figure 3.7 for the L_2 and Tukey factor estimates. These plots show that although both methods have difficulties fitting the house prices for the city of Boston (around observation 400), using the more robust Tukey factors leads to greater errors for these observations than using L_2-based factors. This result implies that these outliers can be considered "good leverage points": downweighting them adversely affects the fit of the model.

3.6 Conclusion

We propose a novel factor extraction method that unifies two recent strands in the factor modelling literature, robustness and sparsity. This method leads to a sparse factor loading matrix and to factors that are robust to outlying observations in the original data. We are the first to combine these two issues in the context of factor modelling, and we argue that both properties can be helpful in forecasting. A Monte Carlo study confirms this intuition: compared to standard principal component analysis, our proposed method gives a much closer approximation to the true factor space; hence, it is more suitable for forecasting purposes. This improvement is obtained at the cost of only small losses in in-sample fit.

We apply this method to two economic data sets. Our first application concerns macroeconomic forecasting using a large panel of predictors. We show that, compared to traditional principal component analysis, our proposed method leads to more interpretable factors. Moreover, we report favorable forecasting performance: for annual growth rates, imposing sparsity on the factor loadings leads to more accurate forecasts for all target variables considered. For shorter-term growth rates, robust estimation provides an additional advantage in forecasting U.S. Industrial Production and Personal Income. This result shows that our factor extraction method, which can be thought of as "multivariate data cleaning", is useful even after the standard univariate data cleaning that was performed by Stock and Watson (2002).

In the second economic application, we analyze the well-known Boston Housing data set. Even in this relatively small data set (sixteen predictor variables), we find that sparse estimation improves the quality of forecasts. We also argue that robust techniques can be expected to fare worse in this data set, as the outliers are actually "good leverage points"; however, their impact on the forecast accuracy turns out to be minimal.

We note that if prior knowledge on a sparse factor structure is available, it is of course possible to impose that certain elements of the loading matrix are zero and use more traditional factor extraction methods, as in Hallin and Liška (2011). This is the case in the macroeconomic data set analyzed in Section 3.4 of this chapter, in which the series are categorized into groups of related variables. However, the results in Section 3.4.2 show that even in this case, the selection of factor loadings that our methodology sets to zero in a data-driven way is similar to the selection that we would impose to be zero based on prior information.

To conclude, we find that sparse and robust estimation of factor models has a great potential for improving both the interpretability of the estimated factors and the accuracy of forecasts. Given its favorable performance in a macroeconomic forecasting study, an interesting generalization of our method would be to dynamic factor models (Forni et al., 2005), in which explicit assumptions about the evolution of the factors over time are made.

Chapter 4

Nonlinear Forecasting with Many Predictors using Kernel Ridge Regression

This chapter is based on Exterkate et al. (2011a).

4.1 Introduction

In current practice, forecasters in macroeconomics and finance face a trade-off between model complexity and forecast accuracy. Due to the uncertainty associated with model choice and parameter estimation, a highly complex predictive regression model based on many variables or intricate nonlinear structures is often found to produce less accurate forecasts than a simpler model that ignores major parts of the information that is at the researcher's disposal. Various methods for working with many predictors while circumventing this *curse of dimensionality* in a linear framework have been applied in the recent forecasting literature, as surveyed by Stock and Watson (2006). Most prominently, Stock and Watson (2002) advocate summarizing large panels of predictor variables into a small number of principal components, which are then used for forecasting purposes in a dynamic factor model. Alternative approaches include combining forecasts based on multiple models, each of which includes only a small number of variables (Faust and Wright, 2009; Wright, 2009; Aiolfi and Favero, 2005; Huang and Lee, 2010; Rapach et al., 2010), partial least squares (Groen and Kapetanios, 2008), and Bayesian regression (De Mol et al., 2008; Bańbura et al., 2010; Carriero et al., 2011). Stock and Watson (2009) find that for forecasting macroeconomic time series, the dynamic factor model approach is preferable to these alternatives; see also Ludvigson and Ng (2007, 2009) and Çakmaklı and van Dijk (2010) for successful applications in finance.

The possibility of nonlinear relations among macroeconomic and financial time series has also received ample attention during the last two decades. Among the most popular nonlinear forecast methods are regime-switching models and neural networks, see the surveys by Teräsvirta (2006) and White (2006), respectively, and the recent comprehensive overview by Kock and Teräsvirta (2011). Typically, these approaches are only suitable for a small number of predictors, and even then their ability to improve upon the predictive accuracy of conventional linear forecasting techniques seems limited, see Stock and Watson (1999); Medeiros et al. (2006), and Teräsvirta et al. (2005), among others.

In this chapter, we introduce a forecasting technique that can deal with high-dimensionality and nonlinearity simultaneously. The central ideas are to employ a flexible set of nonlinear prediction functions and to prevent overfitting by penalization, in a way that limits the computational complexity. In this approach, which is known as *kernel ridge regression*, the set of predictors is mapped into a high-dimensional (or even infinite-dimensional) space of nonlinear functions of the predictors. A forecast equation is estimated in this high-dimensional space, using a penalty (or shrinkage, or ridge) term to avoid overfitting. In this manner, kernel ridge regression does not suffer from the curse of dimensionality, which plagues alternative nonparametric approaches to allow for flexible types of nonlinearity (Pagan and Ullah, 1999). Computational tractability is achieved by choosing the kernel in a convenient way, so that calculations in the high-dimensional space actually are prevented. This approach avoids computational difficulties also encountered in standard linear ridge regression when the number of predictor variables is large relative to the number of time series observations. Taking all these elements together, kernel ridge regression provides an attractive framework for estimating nonlinear predictive relations in a data-rich environment.

The kernel methodology has been developed in the machine learning community, an area which often involves large data sets. The terminology originates from operator theory, as computations are performed in terms of the kernel of a positive integral operator, see Vapnik (1995). We use the term *kernel* in this sense, as it is the established term for this method in machine learning. This meaning should not be confused with other uses of the word, such as in kernel smoothing methods for local regression.

A typical application of kernel methods is classification, for example, in optical recognition of pixel-by-pixel scans of handwritten characters. Schölkopf et al. (1998) document outstanding performance of kernel methods for this classification task. Kernel ridge regression has been found to work well also in many other applications. Time-series applications are scarce and seem to be limited to deterministic (that is, non-stochastic) time series (Müller et al., 1997). Kernel ridge regression has, to our knowledge, not yet been applied in the context of macroeconomic or financial time-series forecasting.

In this chapter, we provide two methodological contributions to kernel ridge regression. First, we extend the approach to enable the use of models that include lags of the dependent variable or other individual variables as predictors, as is typically desired in such applications. Second, we derive an efficient cross-validation procedure for selecting the tuning parameters involved in kernel ridge regression, in particular, the shrinkage parameter that determines the strength of the penalization factor.

We provide Monte Carlo simulation evidence, demonstrating that kernel ridge regression delivers more accurate forecasts than conventional methods based on principal components in the presence of many predictors that are related nonlinearly with the target variable. These conventional methods include the extensions of the principal component regression methodology to accommodate nonlinearity as put forward by Bai and Ng (2008). The potential practical usefulness of kernel methods is confirmed in an empirical application to forecasting four key measures of U.S. macroeconomic activity over the period 1970-2009: Industrial Production, Personal Income, Manufacturing & Trade Sales, and Employment. We find that, when traditional methods perform poorly, kernel ridge regression yields substantial improvements. This result holds for Industrial Production and for Personal Income. Further, when traditional forecasts are of good quality, as is the case for the Sales and Employment series, kernel-based forecasts remain com-

petitive. We also find that kernel ridge regression is much less affected by the recent financial and economic crisis in 2008-9 than the traditional methods.

The remainder of this chapter is organized as follows. Section 4.2 describes the kernel methodology. The Monte Carlo simulation is presented in Section 4.3, and Section 4.4 discusses the empirical application. Conclusions are provided in Section 4.5. Details of the technical results are collected in Appendix 4.A.

4.2 Methodology

The technique of kernel ridge regression (KRR) is based on ordinary least squares (OLS) regression and ridge regression. Therefore, we begin this section with a brief review of these methods, highlighting their drawbacks in dealing with nonlinearity and high-dimensionality. Next, we show how kernel ridge regression overcomes these drawbacks by means of the so-called *kernel trick*. We also present the properties of some kernel functions that are popular because of their computational efficiency. As will become clear below, kernel ridge regression involves tuning parameters. We close this section with a description of a cross-validation procedure for selecting values for these parameters.

4.2.1 Preliminaries

Consider the following general setup for forecasting. At the end of period T we wish to forecast the value of a target variable y at a specific future date, denoted y_*, given an $N \times 1$ vector of predictors x_*. Historical observations for $t = 1, \ldots, T$ are available for all variables, collected in the $T \times 1$ vector y and the $T \times N$ matrix X. If we assume a linear prediction function $\hat{y}_* = x_*'\hat{\beta}$, we may obtain $\hat{\beta}$ by minimizing the OLS criterion $||y - X\beta||^2$, where $||\cdot||$ denotes the L_2 norm. Provided that X has rank N, the solution is $\hat{\beta} = (X'X)^{-1} X'y$, which leads to the forecast $\hat{y}_* = x_*' (X'X)^{-1} X'y$.

The OLS procedure presupposes that $N \leq T$, and in practice, $N \ll T$ is required to prevent overfitting problems. That is, if N is not small compared to T, we may obtain a good in-sample fit (indeed, if $N = T$, the in-sample fit will be perfect), but the out-of-sample prediction $\hat{y}_*$ is generally found to be of poor quality. A possible solution to this problem is shrinkage estimation or ridge regression, which aims to balance the goodness-of-fit and the magnitude of the coefficient vector β. The ridge criterion is given by $||y - X\beta||^2 + \lambda ||\beta||^2$, where the penalty parameter $\lambda > 0$ is to be specified by the user. As every element of the parameter vector β is equally penalized, the predictors in X should be scaled appropriately. In our applications, we studentize each column of X over the estimation sample, so that each predictor has zero mean and unit variance. The solution $\hat{\beta}$ that minimizes the ridge criterion is most easily found by defining the $(T + N) \times 1$ vector $u = \left(y', 0'_{N\times 1}\right)'$ and the $(T + N) \times N$ matrix $V = \left(X', \sqrt{\lambda}\, I_{N\times N}\right)'$, where $I_{N\times N}$ denotes the N-dimensional identity matrix. We may then write $||y - X\beta||^2 + \lambda ||\beta||^2 = ||u - V\beta||^2$. Minimizing this criterion by OLS yields $\hat{\beta} = (V'V)^{-1} V'u$, or, in terms of the original variables, $\hat{\beta} = (X'X + \lambda I)^{-1} X'y$ (where we omit the subscript $N \times N$ from I for notational convenience). The resulting forecast $\hat{y}_* = x_*' (X'X + \lambda I)^{-1} X'y$ can be computed also if the number of predictors N is larger than

the number of observations T. Nevertheless, if the number of regressors becomes very large, the calculation of the ridge forecast may present computational difficulties, as it involves inverting the possibly ill-conditioned $N \times N$ matrix $X'X + \lambda I$. In practice, this hampers the use of ridge regression when $N \gg T$, unless the shrinkage parameter λ is taken to be very large.

4.2.2 Kernel ridge regression and the kernel trick

Kernel ridge regression extends the general setup considered above to allow for nonlinear prediction functions $\hat{y}_* = f(x_*)$. At the same time, it provides a way to avoid the computational complications involved in producing the ridge forecast when the number of predictors becomes very large. As will become clear below, this is particularly relevant in the context of nonlinear forecasting. From now on, let $\varphi : \mathbb{R}^N \to \mathbb{R}^M$ be a (possibly nonlinear) mapping of the N observed predictor variables x resulting in M transformed predictor variables $z = \varphi(x)$. We assume that the prediction function is linear in z, say $\hat{y}_* = z_*'\hat{\gamma}$, where $z_* = \varphi(x_*)$. Collecting the transformed predictor variables in the $T \times M$ matrix Z with rows $z_t' = \varphi(x_t)'$, we may apply ridge regression to obtain $\hat{\gamma} = (Z'Z + \lambda I)^{-1} Z'y$, and hence,

$$\hat{y}_* = z_*' (Z'Z + \lambda I)^{-1} Z'y. \tag{4.1}$$

In macroeconomic and financial applications we often work with high-dimensional data, sometimes with the number of observed predictors N exceeding the number of time series observations T. Moreover, to allow for flexible forms of nonlinearity in the forecast equation, we need $M \gg N$. For example, if we approximate the unknown forecast function f by a dth order Taylor expansion, the mapping φ effectively transforms the $N \times 1$ vector x into the $M \times 1$ vector z containing powers and cross-products of its elements, with M proportional to N^d. Thus, M may become extremely large for realistic values of N and d. As the matrix $Z'Z$ has dimensions $M \times M$, this can cause computational difficulties in producing the nonlinear ridge forecast.

An efficient method to solve this curse of dimensionality problem is provided by the so-called kernel trick. Essentially this method is based on the idea that if the number of regressors M is much larger than the number of observations T, working with T-dimensional instead of M-dimensional objects can lead to notable computational savings. To appreciate the dimension reductions involved, we consider the macroeconomic application that will be discussed in Section 4.4. In this application, we estimate models with $N = 132$ predictors on an estimation sample containing $T = 120$ observations. One of the models includes a constant, all observed predictors, their squares, and the cross-products of all pairs of predictors, leading to a total of $M = (N+1)(N+2)/2 = 8911$ transformed predictor variables. The results that we describe in the remainder of this section allow working with a 120×120 matrix instead of the 8911×8911 matrix $Z'Z$.

This dimension reduction can be achieved by relatively straightforward algebraic manipulations of the expression of the nonlinear ridge forecast equation $\hat{y}_* = z_*'\hat{\gamma}$. First, we rewrite the ridge regression estimator $\hat{\gamma} = (Z'Z + \lambda I)^{-1} Z'y$ as $Z'Z\hat{\gamma} + \lambda\hat{\gamma} = Z'y$, or

$$\hat{\gamma} = \frac{1}{\lambda}(Z'y - Z'Z\hat{\gamma}) = \frac{1}{\lambda} Z'(y - Z\hat{\gamma}).$$

If we pre-multiply $Z'Z\hat{\gamma} + \lambda\hat{\gamma} = Z'y$ by the matrix Z, this gives $ZZ'Z\hat{\gamma} + \lambda Z\hat{\gamma} = ZZ'y$, or

$$Z\hat{\gamma} = (ZZ' + \lambda I)^{-1} ZZ'y.$$

Combining these two results, we find

$$\begin{aligned}\hat{y}_* &= z_*'\hat{\gamma} = \frac{1}{\lambda}z_*'Z'\left(y - Z\hat{\gamma}\right) = \frac{1}{\lambda}z_*'Z'\left(y - \left(ZZ' + \lambda I\right)^{-1}ZZ'y\right) \\ &= \frac{1}{\lambda}z_*'Z'\left(ZZ' + \lambda I\right)^{-1}\left(ZZ' + \lambda I - ZZ'\right)y = z_*'Z'\left(ZZ' + \lambda I\right)^{-1}y.\end{aligned}$$

If we define the $T \times T$ matrix $K = ZZ'$ and the $T \times 1$ vector $k_* = Zz_*$, this result can be written as

$$\hat{y}_* = k_*'\left(K + \lambda I\right)^{-1}y. \tag{4.2}$$

The forecast $\hat{y}_*$ in (4.2) is exactly identical to the one in (4.1). The major advantage of using (4.2) to determine the forecast $\hat{y}_*$ is that the inverse matrix in this equation has dimensions $T \times T$, so that the $M \times M$-dimensional computations in (4.1) are prevented.

To achieve computational savings over the straightforward application of ridge regression, it is crucial that K and k_* can be computed in a relatively simple way. The (s,t)-th element of $K = ZZ'$ equals $z_s'z_t = \varphi\left(x_s\right)'\varphi\left(x_t\right)$, and similarly, the t-th element of k_* equals $\varphi\left(x_t\right)'\varphi\left(x_*\right)$. This implies that the computational efficiency increases greatly if we choose a mapping φ for which the inner product $\kappa\left(a,b\right) = \varphi\left(a\right)'\varphi\left(b\right)$ can be computed quickly, that is, without computing $\varphi\left(a\right)$ and $\varphi\left(b\right)$ explicitly. In this context, κ is called the *kernel function* and K is the *kernel matrix*. This procedure for implicitly finding the optimal parameter vector $\hat{\gamma}$ in the "high" dimension M while working exclusively in the "low" dimension T is known as the *kernel trick* and is due to Boser et al. (1992).

As the above discussion shows, KRR is no different from ordinary ridge regression on transformations of the regressors, except for an algebraic trick to improve computational efficiency. The key to a successful application of this kernel trick is choosing a mapping φ that leads to an easy-to-compute kernel function κ, while, obviously, at the same time φ should be chosen such that the corresponding prediction function $\varphi\left(x_*\right)'\gamma$ provides a good approximation to the true but unknown nonlinear prediction function $f\left(x_*\right)$. Various such mappings are known, and a recent overview is given in Smola and Schölkopf (2004). The next section presents the most commonly used instances of these mappings.

In a time series context, we often prefer to include specific predictors in the forecast equation separately from the nonlinear mapping φ. In macroeconomic applications, these "preferred" predictors may include lags of the dependent variable to account for serial correlation. In financial applications such as predicting stock returns, these predictors may include valuation ratios such as the dividend yield or interest rate related variables; see Ludvigson and Ng (2007), Çakmaklı and van Dijk (2010), for example. In such cases the generalized forecast equation takes the form $\hat{y}_* = w_*'\hat{\beta} + z_*'\hat{\gamma}$, where the $P \times 1$ vector w_* contains the variables to be treated linearly. As the number of these additional predictors is limited and their effect is of particular interest, we do not penalize the parameters β and restrict the ridge penalization to γ. We show in Appendix 4.A.1 that the derivations that lead to (4.2) can be extended to include such unpenalized linear terms, resulting in the "extended" KRR forecast equation

$$\hat{y}_* = \begin{pmatrix} k_* \\ w_* \end{pmatrix}' \begin{pmatrix} K + \lambda I & W \\ W' & 0 \end{pmatrix}^{-1} \begin{pmatrix} y \\ 0 \end{pmatrix}, \tag{4.3}$$

where the $T \times P$ matrix W contains historical observations on the variables to be treated linearly. This is the forecast equation that will be used in the empirical application in Section 4.4.

4.2.3 Some common kernel functions

A first and obvious example is the identity mapping $\varphi(a) = a$, for which $\kappa(a, b) = a'b$. With this choice of κ, the kernel forecast $\hat{y}_* = k_*' (K + \lambda I)^{-1} y = x_*' X' (XX' + \lambda I)^{-1} y$ equals the linear ridge forecast $\hat{y}_* = x_*' (X'X + \lambda I)^{-1} X'y$, as can be seen by taking $Z = X$ and $z_* = x_*$ in the derivations leading to (4.2).

Next we consider a mapping such that $\varphi(a)$ contains a constant term, all original variables $a_1, a_2, \ldots, a_N$, and all squares and cross products of these variables. Some experimentation reveals that $\kappa(a, b)$ takes a particularly simple form if we multiply some elements of $\varphi(a)$ by the constant $\sqrt{2}$. That is, if we choose the mapping

$$\varphi(a) = \left(1, \sqrt{2}a_1, \sqrt{2}a_2, \ldots, \sqrt{2}a_N, a_1^2, a_2^2, \ldots, a_N^2, \sqrt{2}a_1a_2, \sqrt{2}a_1a_3, \ldots, \sqrt{2}a_{N-1}a_N\right)',$$

the corresponding kernel function is

$$\begin{aligned}
\kappa(a, b) &= \varphi(a)' \varphi(b) \\
&= 1 + 2(a_1b_1 + a_2b_2 + \ldots + a_Nb_N) + a_1^2b_1^2 + a_2^2b_2^2 + \ldots + a_N^2b_N^2 \\
&\quad + 2(a_1a_2b_1b_2 + a_1a_3b_1b_3 + \ldots + a_{N-1}a_Nb_{N-1}b_N) \\
&= 1 + 2(a_1b_1 + a_2b_2 + \ldots + a_Nb_N) + (a_1b_1 + a_2b_2 + \ldots + a_Nb_N)^2 \\
&= 1 + 2a'b + (a'b)^2 = (1 + a'b)^2.
\end{aligned}$$

With this specification of the kernel function, the computation of each of the $T(T+1)/2$ distinct elements of the kernel matrix K requires $2(N+1)$ additions and multiplications. In the absence of the indicated scaling, the vector of constant, first-order, and second-order terms contains $M = (N+1)(N+2)/2$ elements. The computation of each element of the kernel matrix would then require $2M = (N+1)(N+2)$ additions and multiplications. Thus, the proposed scaling reduces the amount of computations by a factor of $(N+2)/2$.

As noted by Poggio (1975), this result can be generalized to the kernel function

$$\kappa(a, b) = (1 + a'b)^d \qquad \text{for any integer } d \geq 1, \tag{4.4}$$

corresponding to a mapping for which $\varphi(a)$ consists of all polynomials in the elements of a of degree at most d. Observe that this class of so-called polynomial kernel functions encompasses not only the quadratic mapping, for $d = 2$, but also the identity mapping (and hence, standard linear ridge regression), for $d = 1$.

Because smart choices of φ enable us to avoid M-dimensional computations, the kernel methodology even allows letting $M \to \infty$. A common way to do this, dating back to Broomhead and Lowe (1988), is by using the Gaussian kernel function

$$\kappa(a, b) = \exp\left(-\frac{1}{2}||a - b||^2\right). \tag{4.5}$$

We show in Appendix 4.A.2 that the corresponding mapping $\varphi(a)$ contains as elements, for all degrees $d_1, d_2, \ldots, d_N \geq 0$, the "dampened" polynomials

$$e^{-a'a/2} \prod_{n=1}^{N} \frac{a_n^{d_n}}{\sqrt{d_n!}}.$$

In this study, we consider the polynomial kernels (4.4) of degrees $d = 1$ and 2, as well as the Gaussian kernel (4.5). To control for the relative importance of the terms in $\varphi(x)$, we replace each observation x by $(1/\sigma)\,x$ before computing κ, for some positive scaling factor σ. Such scaling affects the weight placed on different polynomial degrees, as it amounts to dividing linear terms in $\varphi(x)$ by σ, second-order terms by σ^2, and so forth. Although we are performing linear regression on $\varphi(x)$, such scaling is not without effect, as its regression coefficients in the forecast equation $\hat{y}_* = w_*'\hat{\beta} + \varphi(x_*)'\hat{\gamma}$ are all penalized equally by the ridge term in the criterion function $||y - W\beta - Z\gamma||^2 + \lambda\,||\gamma||^2$.

4.2.4 Selection of tuning parameters

The implementation of kernel ridge regression involves two tuning parameters, namely, the shrinkage parameter λ and the scaling parameter σ. Additionally, our empirical application in Section 4.4 to several macroeconomic time series involves the selection of lag lengths, which can also be seen as tuning parameters from a model selection perspective. This section addresses the question of how to select the values for these tuning parameters.

We determine the values of the tuning parameters by means of leave-one-out cross-validation, as this is a natural criterion for the purpose of out-of-sample forecasting. For given values of the tuning parameters, we estimate the model on the sample of size $T-1$ that remains when the observation for period t is removed. We then use this model to "forecast" the value of y_t that was left out. This is repeated T times, leaving out each observation for $t = 1, 2, \ldots, T$ once. Performing this cross-validation exercise for each of the candidate values of the tuning parameters, we select those values that lead to the smallest mean squared prediction error (MSPE) over these T forecasts. These values are then used to estimate the model on the full sample $t = 1, 2, \ldots, T$, from which we produce out-of-sample forecasts.

In the form stated above, this cross-validation procedure is computationally very expensive, as it requires estimating the model on T different samples for each possible setting of the tuning parameters. Cawley and Talbot (2008) propose a method that yields all leave-one-out prediction errors as a by-product of estimating (4.2) only once, that is, on the full sample. We derive a similar result, extended to allow for the additional linear terms in (4.3), in Appendix 4.A.3.

In the simulation study and in the empirical application below, we use this method to select both the lag lengths and the ridge parameter λ from a grid. We will use a rolling window of fixed length for estimation, and we reselect the values of the tuning parameters for each window. As it is difficult to find intuitively reasonable values for the shrinkage parameter, we employ a fairly wide grid, containing 45 candidate values:

$$\log_{10}(\lambda) \in \{-8, -5, -4.0, -3.8, -3.6, \ldots, 3.6, 3.8, 4.0, 5, 8\}.$$

The same cross-validation procedure could also be used to select the scaling parameter σ. Preliminary simulation evidence shows that it is difficult to identify λ and σ simultaneously from data, as a wide range of (λ, σ) combinations is found to lead to very similar fits and forecasts. Only little out-of-sample forecast quality is sacrificed if the scaling parameter σ is fixed a priori. Based on these exploratory results for polynomial kernels of degree up to three and for the Gaussian kernel, we rescale the data to have mean zero and variance one and then we use, as a practical rule of thumb, $\sigma = 2^d$ for the polynomial kernel of degree d and $\sigma = 10$

for the Gaussian kernel. The increasing value of σ for higher degree d reflects the fact that the number of regressors is proportional to N^d.

As a technical note on cross-validation, serial correlation in time-series data leads to dependence between the observations in the estimation sample and the observation that was left out. This dependence implies that the standard leave-one-out cross-validation procedure may not be fully adequate; see Racine (2000) for an extensive discussion and a modification to overcome these problems. Although the method outlined in Appendix 4.A.3 can easily be adapted to this modified form of cross-validation, the resulting implementation is computationally quite intensive. We find that the results from using this modified procedure are not appreciably different from those obtained with simple leave-one-out cross-validation (detailed results are available upon request). Therefore, we only report the results that are obtained using the latter method.

4.3 Monte Carlo simulation

To evaluate the potential of kernel ridge regression in a data-rich environment (that is, when many predictor variables are present), we assess its forecasting performance for a set of static factor models through a Monte Carlo study. We consider a setting with two latent factors f_{1t} and f_{2t}, which are taken to be uncorrelated standard normal variables. As predictor variables, $N = 100$ noisy linear combinations of these factors are generated by $x_{it} = \theta_{i1} f_{1t} + \theta_{i2} f_{2t} + \eta_{it}$, where the factor loadings θ_{ij}, $j = 1, 2$, are drawn from the standard normal distribution. The noise η_{it} is also normal with mean zero, while its variance is selected to control the fraction of the variance of each x_i variable explained by the factors, denoted by R_x^2. We consider two cases with R_x^2 equal to 0.4 or 0.8, which we label as "weak" and "strong" factor structure, respectively. The target variable y is constructed according to three different DGPs:

$$\text{Linear:} \qquad y_t = f_{1t} + f_{2t} + \varepsilon_t \tag{4.6}$$

$$\text{Squared:} \qquad y_t = f_{1t}^2 + f_{2t}^2 + \varepsilon_t \tag{4.7}$$

$$\text{Cross-product:} \quad y_t = f_{1t} f_{2t} + \varepsilon_t \tag{4.8}$$

Here ε_t is normal with mean zero and a variance selected to control R_y^2, the fraction of the variance of y_t that is explained by the factors. For R_y^2 we also consider the values 0.4 and 0.8, which we refer to as "weak" and "strong" predictive structure, respectively.

In each Monte Carlo replication, we generate time series of x_i, $i = 1, \ldots, N$, and y, each consisting of $T + 1$ observations. The first T observations are used for estimation, and a forecast for y_{T+1} is made based on x_{T+1}. All variables are standardized to have mean zero and variance one in the estimation sample. We set the sample size to $T = 120$, which corresponds to the length of the estimation window (ten years of monthly observations) used in the empirical application in Section 4.4. We present results based on 5000 replications.

In principle, OLS regression using the individual predictor variables can be applied in this setting, as the number of predictors is smaller than the number of observations in the estimation sample. It should not come as a surprise, though, that this procedure leads to a very poor forecasting performance, given the large amount of parameter estimation uncertainty when $N = 100$ and $T = 120$. We therefore do not report the OLS results for brevity. Instead, we consider four alternative prediction methods for comparison with KRR:

(i) the "mean" forecast, with $\hat{y}_{121} = (1/120) \sum_{t=1}^{120} y_t$;

(ii) principal component regression (PC), which amounts to OLS but with regressors $\hat{f}_t$ being the first k principal components of the predictor variables x;

(iii) "PC-squared" (PC^2), as suggested by Bai and Ng (2008), which corresponds to principal component regression with the squares of $\hat{f}_t$ as additional regressors; and

(iv) "Squared PC" (SPC), also proposed by Bai and Ng (2008), which is principal component regression but using the principal components of the original predictor variables x and their squares.

Bai and Ng (2008) also propose a quadratic principal component (QPC) regression variant, in which principal components are taken not only of the original variables and their squares (as in SPC), but also including their cross-products. They report high computational costs and poor forecasting performance for this technique, and our preliminary analysis confirms these results. For this reason, QPC is not considered in our study.

For KRR, the shrinkage parameter λ is selected from the grid defined in Section 4.2.4 using leave-one-out cross-validation. For each of the principal-components-based methods, we select the number of components k by minimizing the Bayesian Information Criterion (BIC), where $1 \leq k \leq 10$. Our reason for minimizing BIC instead of performing cross-validation for these methods is twofold. First, using BIC in principal components forecasting settings is common in the literature; see, for example, Stock and Watson (2002) and Bai and Ng (2008). Second, preliminary simulation evidence shows that using BIC leads to superior results compared to using cross-validation.

Table 4.1 shows mean squared prediction errors (MSPEs) relative to the variance of the series being predicted. Note that if the factor values $f_{1,T+1}$ and $f_{2,T+1}$ were known, these relative MSPEs would be close to $1 - R_y^2$, or 0.6 and 0.2 in the two scenarios of "weak" ($R_y^2 = 0.4$) and "strong" ($R_y^2 = 0.8$) predictive structure considered here. Standard PC shows good performance for the linear DGP, while PC^2 performs well for the squared DGP. Such results were to be expected, because the forecast equation in these methods corresponds exactly with these DGPs. Interestingly, the kernel methods are not much less accurate than these "optimal" methods, with the obvious exception of the Poly(1) (that is, linear) kernel in the squared DGP (for which standard PC also fares badly). This finding holds regardless of whether R_x^2 and R_y^2 are high or low, although the difference between PC or PC^2 and the best performing kernel method is smaller when the factor structure in the predictor variables is stronger (compare $R_x^2 = 0.4$ with $R_x^2 = 0.8$). Thus, we find that kernel methods can work well in standard factor model settings.

For the cross-product DGP, the SPC method from Bai and Ng (2008) and the Poly(2) kernel can both be expected to perform well. We observe that KRR provides the most accurate forecasts here, and that the gains are larger for lower R_x^2. Thus KRR performs well in this case, especially when the factor structure of the predictors is not very strong, as is often the case for empirical macroeconomic and financial data. The Gaussian kernel also performs reasonably well.

We conclude that the use of kernel methods in a factor context works quite well, especially for nonlinear relations and in situations where the observed predictors give relatively little information on the factors.

Table 4.1: Relative mean squared prediction errors for the factor models (4.6)-(4.8).

DGP	Linear				Squared				Cross-product			
	$R_y^2 = 0.4$		$R_y^2 = 0.8$		$R_y^2 = 0.4$		$R_y^2 = 0.8$		$R_y^2 = 0.4$		$R_y^2 = 0.8$	
$R_x^2 =$	0.4	0.8	0.4	0.8	0.4	0.8	0.4	0.8	0.4	0.8	0.4	0.8
Mean	1.00	1.00	1.02	1.02	1.03	1.03	1.07	1.07	1.04	1.04	1.10	1.10
PC	**0.62**	**0.61**	**0.23**	**0.20**	1.04	1.05	1.10	1.10	1.06	1.06	1.13	1.13
PC^2	0.63	0.62	0.23	0.21	**0.65**	**0.63**	**0.27**	**0.22**	0.90	0.89	0.73	0.72
SPC	0.63	0.63	0.24	0.22	0.70	0.64	0.35	0.22	0.79	0.65	0.52	0.23
Poly(1)	0.65	0.62	0.24	0.21	1.04	1.04	1.09	1.09	1.06	1.06	1.12	1.12
Poly(2)	0.68	0.64	0.27	0.22	0.70	0.64	0.32	0.23	**0.70**	**0.64**	**0.32**	**0.23**
Gauss	0.68	0.67	0.29	0.26	0.77	0.70	0.41	0.32	0.84	0.75	0.49	0.38

Notes: This table reports mean squared prediction errors (MSPEs) for models (4.6)-(4.8), averaged over 5000 forecasts, and relative to the variance of the series being predicted. The smallest relative MSPE for each DGP (column) is printed in boldface.

4.4 Macroeconomic forecasting

4.4.1 Data and forecast models

We evaluate the forecast performance of kernel ridge regression in an empirical application, involving a large panel of U.S. macroeconomic and financial variables. The data set consists of monthly observations on 132 variables, including various measures of production, consumption, income, sales, employment, monetary aggregates, prices, interest rates, and exchange rates. All series are transformed to stationarity by taking logarithms and/or differences, as described in Stock and Watson (2005). We have updated their data set, which starts in January 1959 and ends in December 2003, to cover the period up to (and including) January 2010. The cross-sectional dimension varies somewhat over time because of data availability: some time series start later than January 1959, while a few other variables have been discontinued before the end of our sample period. For each month under consideration, observations on at most five variables are missing.

We focus on forecasting four key measures of real economic activity: Industrial Production, Personal Income less Transfer Payments (referred to as Personal Income in the following), Manufacturing & Trade Sales, and Employment on Non-Agricultural Payrolls (referred to as Employment in the remainder of this section), as in Stock and Watson (2002), among others. For each of these variables, we produce out-of-sample forecasts for the annualized h-month percentage growth rate, computed as

$$y_{t+h}^h = \frac{1200}{h} \ln\left(\frac{v_{t+h}}{v_t}\right),$$

where v_t is the untransformed observation on the level of each variable in month t. To simplify notation, we denote the one-month growth rate as y_{t+1}. We consider growth rate forecasts for $h = 1, 3, 6$ and 12 months.

Kernel ridge regression is compared against several alternative forecasting approaches that are popular in current macroeconomic practice. As benchmarks we include the "mean" forecast (that is, the average growth over the estimation window); the "no-change" or random-walk (RW) forecast, and an autoregressive (AR) forecast (using lagged values of the one-month growth rates as predictors). In addition, as the primary competitor for kernel methods we consider the diffusion index (DI) approach of Stock and Watson (2002), who document its good performance for forecasting the same four macroeconomic variables as considered here. The DI methodology extends the standard principal component regression to a dynamic setting by including autoregressive lags as well as lags of the principal components in the forecast equation. Specifically, using p autoregressive lags and q lags of k factors, at time t, this "extended" principal-components method produces the forecast

$$\hat{y}^h_{t+h|t} = w_t'\hat{\beta} + \hat{f}_t'\hat{\gamma},$$

where $w_t = \left(1, y_t, y_{t-1} \ldots, y_{t-(p-1)}\right)'$ and $\hat{f}_t = \left(\hat{f}_{1,t}, \hat{f}_{2,t}, \ldots, \hat{f}_{k,t}, \hat{f}_{1,t-1}, \ldots, \hat{f}_{k,t-(q-1)}\right)'$. The lags of the dependent variable in w_t are one-month growth rates, irrespective of the forecast horizon h, because using h-month growth rates for $h > 1$ would lead to highly correlated regressors. The factors $\hat{f}$ are principal components extracted from all 132 predictor variables, and $\hat{\beta}$ and $\hat{\gamma}$ are OLS estimates. Aside from standard principal components (PC), we also consider its extensions PC^2 and SPC, discussed in Section 4.3. In each case, the lag lengths p and q and the number of factors k are selected by minimizing the Bayesian Information Criterion (BIC). This criterion is used instead of cross-validation for two reasons. We want our results to be comparable to those in Stock and Watson (2002) and Bai and Ng (2008), and preliminary experimentation with the PC methods has revealed that using the BIC leads to superior results. Following Stock and Watson (2002), we allow $0 \leq p \leq 6$ (where $p = 0$ means that $w_t = 1$), $1 \leq q \leq 3$, and $1 \leq k \leq 4$. Thus, the simplest model that can be selected uses no information on current or lagged values of the dependent variable, and information from the other predictors in the current month only, summarized by a single factor. Also in line with Stock and Watson (2002), we do not perform an exhaustive search across all possible combinations of the first four principal components and lag structures. Instead, we assume that factors are included sequentially in order of importance, while the number of lags is assumed to be the same for all included factors.

For KRR, the corresponding forecast equation is

$$\hat{y}^h_{t+h|t} = w_t'\hat{\beta} + \varphi\left(\left(x_t', x_{t-1}', \ldots, x_{t-(q-1)}'\right)'\right)'\hat{\gamma},$$

in the notation of Section 4.2.2, where w_t is as defined above and x_t contains all 132 predictors at time t. The parameter vectors $\hat{\beta}$ and $\hat{\gamma}$ are obtained by KRR, resulting in the forecast equation (4.3). The lag lengths p and q, as well as the kernel penalty parameter λ, are selected by leave-one-out cross-validation.

All models are estimated on rolling windows with a fixed length of 120 months, such that the first forecast is produced for the growth rate during the first h months of 1970. For each window, the tuning parameter values are re-selected and the regression coefficients are re-estimated. That is, all of the tuning parameters (p, q, k, λ) are allowed to differ over time and across methods.

Table 4.2: Relative mean squared prediction errors for the macroeconomic series.

Forecast	Industrial Production				Personal Income			
method	$h=1$	$h=3$	$h=6$	$h=12$	$h=1$	$h=3$	$h=6$	$h=12$
Mean	1.02	1.05	1.07	1.08	1.02	1.06	1.10	1.17
RW	1.27	1.08	1.34	1.64	1.60	1.36	1.14	1.35
AR	0.93	0.89	1.02	1.02	1.17	1.05	1.10	1.15
PC	0.81	0.71	**0.77**	0.63	1.04	**0.79**	0.90	0.90
PC^2	0.94	0.85	1.20	1.07	1.09	0.92	1.03	1.15
SPC	0.88	0.98	1.35	0.99	1.07	1.04	1.05	1.50
Poly(1)	**0.75**	**0.67**	0.85	**0.62**	0.92	0.82	0.93	1.12
Poly(2)	0.91	0.77	0.97	0.76	0.96	0.88	0.98	1.01
Gauss	0.80	0.80	0.78	0.67	**0.89**	0.81	**0.81**	**0.80**

Forecast	Manufacturing & Trade Sales				Employment			
method	$h=1$	$h=3$	$h=6$	$h=12$	$h=1$	$h=3$	$h=6$	$h=12$
Mean	1.01	1.03	1.05	1.08	0.98	0.96	0.97	0.97
RW	2.17	1.49	1.45	1.53	1.68	0.95	1.00	1.20
AR	1.01	1.02	1.10	1.08	0.96	0.85	0.90	0.96
PC	**0.89**	**0.80**	**0.77**	0.63	**0.76**	**0.56**	**0.48**	**0.48**
PC^2	0.94	0.97	1.13	1.06	0.76	0.61	0.69	0.60
SPC	0.99	1.18	1.59	1.02	0.81	0.81	0.90	0.72
Poly(1)	0.94	0.85	0.81	**0.56**	0.87	0.61	0.55	0.56
Poly(2)	0.97	1.03	1.20	0.83	0.82	0.71	0.64	0.84
Gauss	0.95	0.91	0.86	0.74	0.85	0.67	0.63	0.64

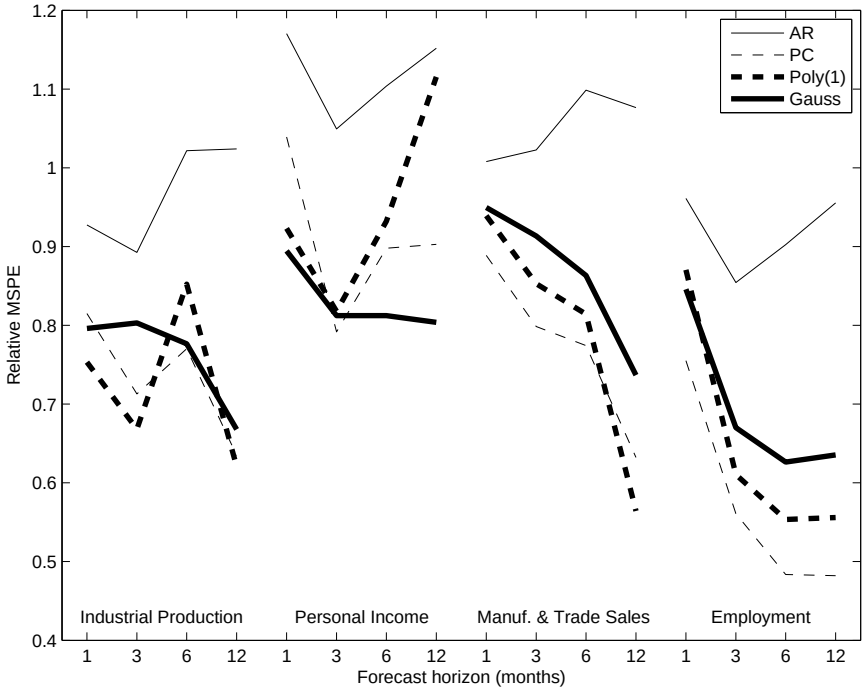
Notes: This table reports mean squared prediction errors (MSPEs) for four macroeconomic series, over the period 1970-2010, relative to the variance of the series being predicted. The smallest relative MSPE for each series (column) is printed in boldface.

4.4.2 Results

Table 4.2 shows the MSPEs for the period 1970-2010 for the three benchmark methods, three PC-based methods, and three kernel methods. Several conclusions can be drawn from these results. First, we observe that KRR provides more accurate forecasts than any of the three benchmarks (mean, random walk, and autoregression) for all target variables and all forecast horizons, with larger gains for longer horizons. This holds irrespective of the kernel function that is used, the only exceptions being that the second-order polynomial kernel produces worse forecasts for the three-month and six-month growth rates of Manufacturing & Trade Sales. In many cases the improvements in predictive accuracy are substantial, even compared to the AR forecast, which seems the best of the three benchmarks. For example, for 12-month growth rate forecasts, kernel ridge regression based on the Gaussian kernel achieves a reduction in MSPE of about 30% for all four variables (relative to the AR forecasts).

Second, if we compare the forecasts based on KRR and the linear PC-based approach, we find somewhat mixed results, but generally the kernel methods perform better. Kernel ridge

Figure 4.1: Relative mean squared prediction errors for four macroeconomic series, for selected methods.



forecasts are superior for Industrial Production and Personal Income for three of the four horizons considered. The improvements in relative MSPE range from 0.01 for Industrial Production at the 12-month horizon to 0.15 for Personal Income at the shortest horizon of one month. For Manufacturing & Trade Sales, kernels perform better at the longest horizon and slightly worse at the shorter horizons. Finally, for Employment, the PC-based forecasts are more accurate than kernel-based forecasts by about 10-15%.

Third, the KRR approach convincingly outperforms the PC^2 and SPC variants of the principal component regression framework. In fact, also the linear PC specification renders substantially more accurate forecasts than these two extensions in all cases. Apparently, the PC^2 and SPC methods cannot successfully cope with the possibly nonlinear relations between the target variables and the predictors in this application. (Bai and Ng (2008) report somewhat better forecast performance if SPC is applied to a selected subset of the predictors, rather than to the full predictor set. Also with this modification, SPC has difficulties outperforming simpler linear methods in their application.)

Fourth, among the kernel-based methods, the Poly(1) kernel and the Gaussian kernel generally perform best, achieving lower MSPEs than the Poly(2) kernel in all but a few cases. Furthermore, all MSPE / variance ratios in Table 4.2 are below one for these methods (except for the Poly(1) kernel for Personal Income at $h = 12$). Neither of the two consistently outperforms the other. Although Poly(1) performs better than the Gaussian kernel in some cases, the latter kernel method shows satisfactory results in all situations.

A subset of the results in Table 4.2 is reproduced graphically in Figure 4.1. This graph allows us to interpret the mixed results in the comparison of kernel-based and linear PC-based forecasts as follows. KRR (especially using the Gaussian kernel) shows roughly the same good performance for all four series. However, the quality of PC forecasts varies substantially among

Table 4.3: Estimated coefficients $\hat{\alpha}$ from the forecast combining regression (4.9).

Forecast	Industrial Production				Personal Income			
method	$h=1$	$h=3$	$h=6$	$h=12$	$h=1$	$h=3$	$h=6$	$h=12$
PC	0.83* (0.15)	0.80* (0.14)	0.72*† (0.13)	0.79* (0.11)	0.97* (0.26)	0.87* (0.12)	0.70*† (0.10)	0.70*† (0.09)
PC^2	0.48*† (0.15)	0.55*† (0.11)	0.42*† (0.12)	0.48*† (0.13)	0.71* (0.18)	0.66*† (0.12)	0.54*† (0.07)	0.50*† (0.10)
SPC	0.57*† (0.08)	0.43*† (0.11)	0.37*† (0.12)	0.51*† (0.08)	0.75* (0.21)	0.51*† (0.09)	0.52*† (0.10)	0.39*† (0.08)
Poly(1)	0.89* (0.11)	0.91* (0.14)	0.63*† (0.13)	0.81* (0.12)	1.07* (0.19)	1.01* (0.14)	0.63*† (0.12)	0.52*† (0.12)
Poly(2)	0.53*† (0.09)	0.77* (0.19)	0.55* (0.25)	0.85* (0.15)	1.01* (0.23)	0.78* (0.14)	0.63*† (0.15)	0.68* (0.17)
Gauss	1.23* (0.18)	0.74* (0.16)	0.89* (0.15)	0.99* (0.17)	1.29*† (0.14)	1.10* (0.15)	0.96* (0.15)	0.95* (0.14)

Forecast	Manufacturing & Trade Sales				Employment			
method	$h=1$	$h=3$	$h=6$	$h=12$	$h=1$	$h=3$	$h=6$	$h=12$
PC	0.83* (0.12)	0.86* (0.13)	0.87* (0.17)	0.91* (0.12)	1.02* (0.09)	0.93* (0.09)	0.92* (0.10)	1.04* (0.12)
PC^2	0.64*† (0.08)	0.55*† (0.11)	0.48*† (0.19)	0.51*† (0.15)	0.91* (0.06)	0.74*† (0.07)	0.62*† (0.10)	0.77* (0.14)
SPC	0.53*† (0.09)	0.39*† (0.14)	0.29† (0.15)	0.52*† (0.10)	0.74*† (0.07)	0.53*† (0.08)	0.50*† (0.09)	0.60*† (0.09)
Poly(1)	0.66*† (0.14)	0.83* (0.18)	0.80* (0.16)	0.97* (0.15)	0.68*† (0.12)	0.97* (0.14)	0.97* (0.11)	0.91* (0.14)
Poly(2)	0.61*† (0.12)	0.48*† (0.17)	0.40† (0.24)	0.89* (0.26)	0.90* (0.11)	0.74* (0.21)	0.77* (0.13)	0.66* (0.19)
Gauss	0.76* (0.13)	0.89* (0.17)	0.86* (0.14)	1.06* (0.22)	0.98* (0.15)	1.21* (0.18)	1.10* (0.13)	1.07* (0.18)

Notes: This table reports $\hat{\alpha}$, the weight placed on the candidate forecast in the forecast combining regression (4.9). HAC standard errors follow in parentheses. An asterisk ($*$) indicates rejection of the hypothesis $\alpha = 0$ and a dagger (†) indicates rejection of $\alpha = 1$, at 5% significance.

the series and is exceptionally high for the Employment series. Recall that in the Monte Carlo experiment in Section 4.3, we find the analogous result that kernel-based methods yield better relative performance, compared to PC-based methods, if the factor structure is relatively weak. That is, our results suggest that kernel ridge regression performs better than principal component regression unless the latter performs very well.

Following Stock and Watson (2002), we provide a further evaluation of our results by using the forecast combining regression

$$y_{t+h}^h = \alpha\, \hat{y}_{t+h|t}^h + (1-\alpha)\, \hat{y}_{t+h|t}^{h,\,\text{AR}} + u_{t+h}^h, \tag{4.9}$$

where y_{t+h}^h is the realized growth rate over the h-month period ending in month $t+h$, $\hat{y}_{t+h|t}^h$ is a candidate forecast from either the PC-based methods or from kernel ridge regression made at time t, and $\hat{y}_{t+h|t}^{h,\,\text{AR}}$ is the benchmark autoregressive forecast. Estimates of α are shown in Table 4.3, with heteroscedasticity and autocorrelation consistent (HAC) standard errors in parentheses. The null hypothesis that the AR forecast receives unit weight ($\alpha = 0$) is strongly rejected in almost all cases, which means that PC-based and kernel-based forecasts have significant additional predictive ability relative to this benchmark. Actually, the null hypothesis that the candidate forecast receives unit weight ($\alpha = 1$) cannot be rejected in many cases. Note that $\alpha = 1$ in fact means that the candidate forecast encompasses the AR forecast. This hypothesis is not rejected for PC-based methods in 17 out of 48 cases, and for kernel-based methods even in 37 out of 48 cases. This result confirms the conclusion drawn above that including macro factors improves forecasting performance relative to univariate benchmark models, especially when allowing for nonlinear predictive relations with the target variable in a flexible, nonparametric way as in kernel ridge regression.

In order to compare the performance of kernel-based and PC-based forecasts directly, we run a similar forecast combining regression

$$y_{t+h}^h = \alpha\, \hat{y}_{t+h|t}^{h,\,\text{KRR}} + (1-\alpha)\, \hat{y}_{t+h|t}^{h,\,\text{PC}} + u_{t+h}^h. \tag{4.10}$$

Table 4.4: Estimated coefficients $\hat{\alpha}$ from the forecast combining regression (4.10).

Forecast	Industrial Production				Personal Income			
method	$h=1$	$h=3$	$h=6$	$h=12$	$h=1$	$h=3$	$h=6$	$h=12$
Poly(1)	0.67*† (0.09)	0.63*† (0.16)	0.36*† (0.13)	0.53*† (0.15)	0.79* (0.17)	0.39† (0.24)	0.45*† (0.13)	0.24† (0.17)
Poly(2)	0.30*† (0.11)	0.40*† (0.15)	0.20† (0.20)	0.29*† (0.11)	0.69* (0.22)	0.34*† (0.13)	0.42*† (0.13)	0.41*† (0.07)
Gauss	0.57*† (0.17)	0.30† (0.19)	0.49*† (0.14)	0.45*† (0.11)	0.85* (0.17)	0.45*† (0.16)	0.63*† (0.16)	0.62*† (0.10)

Forecast	Manufacturing & Trade Sales				Employment			
method	$h=1$	$h=3$	$h=6$	$h=12$	$h=1$	$h=3$	$h=6$	$h=12$
Poly(1)	0.33*† (0.16)	0.29*† (0.15)	0.41*† (0.17)	0.72* (0.26)	0.17*† (0.08)	0.23† (0.14)	0.16† (0.16)	0.26† (0.19)
Poly(2)	0.24† (0.14)	-0.06† (0.17)	-0.29† (0.21)	0.18*† (0.09)	0.18† (0.13)	-0.05† (0.15)	0.05† (0.15)	0.04† (0.08)
Gauss	0.29† (0.16)	0.23† (0.18)	0.38*† (0.16)	0.32*† (0.13)	0.15† (0.12)	0.19† (0.13)	0.23*† (0.11)	0.23*† (0.11)

Notes: This table reports $\hat{\alpha}$, the weight placed on the kernel forecast in the forecast combining regression (4.10). HAC standard errors follow in parentheses. An asterisk ($*$) indicates rejection of the hypothesis $\alpha = 0$ and a dagger (†) indicates rejection of $\alpha = 1$, at 5% significance.

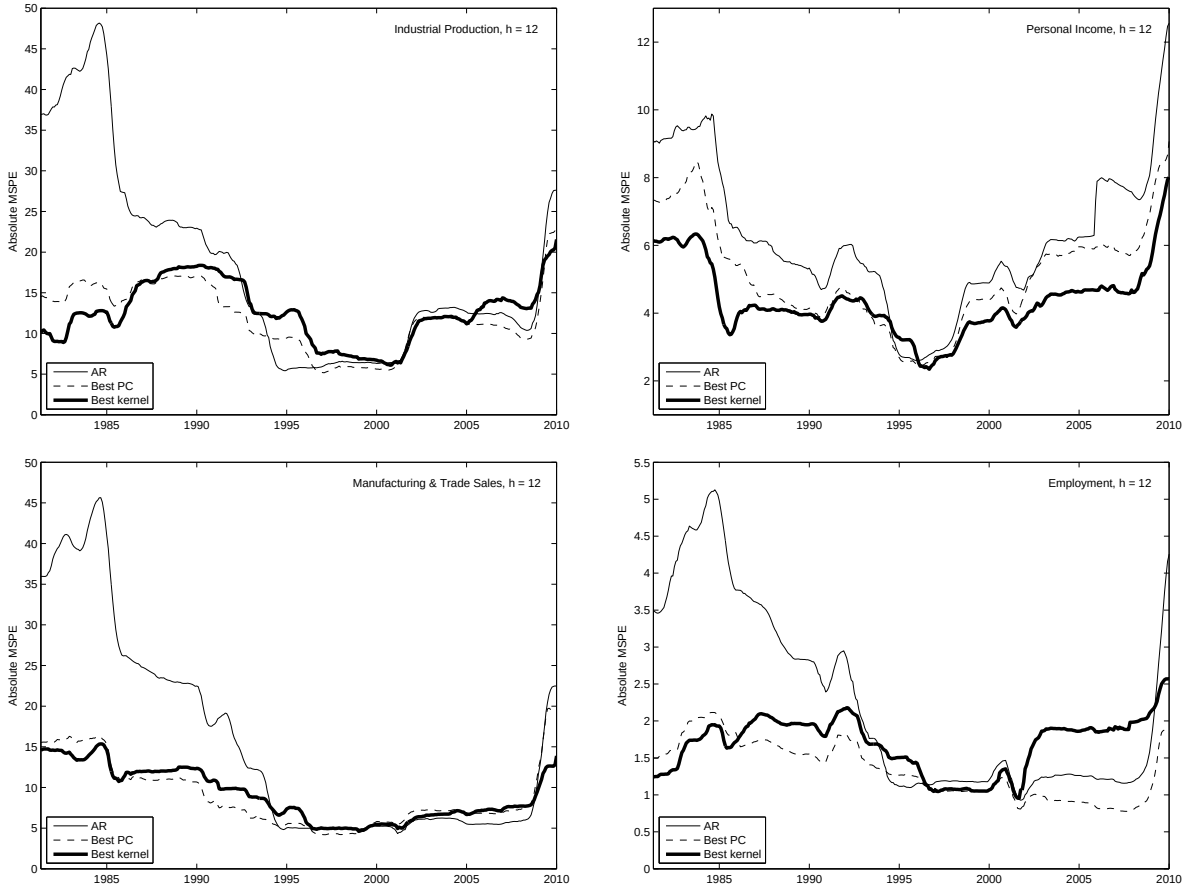
As linear PC performs better than PC^2 and SPC (see Table 4.2), we compare kernel methods to linear PC only. We report the estimates of α in Table 4.4. These results show that both hypotheses of interest ($\alpha = 0$ and $\alpha = 1$) are rejected in many cases (26 out of 48), suggesting that forecasts obtained from both types of models are complementary. Apparently, each forecast method uses relevant information that the other method misses.

Finally, we examine the stability of the forecasting performance of KRR and PC-based methods over time. For this purpose, Figure 4.2 shows time-series plots of rolling MSPEs for AR and for the best-performing PC and kernel methods, where the value plotted for date t is the MSPE computed over the ten-year subsample ending with the forecast for date t, that is, $\hat{y}^h_{t|t-h}$. We show only the series for $h = 12$, as the results for the other horizons are qualitatively similar. This figure confirms that, when KRR forecasts are less accurate than PC-based forecasts, this is because PC-based forecasts are very accurate, and not because KRR forecasts would be inaccurate. Another interesting feature evidenced by Figure 4.2 is that, while the recent crisis reduces the accuracy of all forecasts from 2008 onward, it affects kernel-based forecasts least.

Figure 4.3 shows the corresponding time series of relative MSPEs for ten-year rolling windows, together with the rolling variance of the series being predicted. Together with Figure 4.2, these graphs lead to the following two conclusions. First, predictability improves in an absolute sense during less volatile times, in the sense that the MSPEs in Figure 4.2 typically decline when the rolling variance of the series being predicted in Figure 4.3 goes down. Second, forecasting becomes more difficult in a relative sense during less volatile periods, in the sense that the relative MSPEs seem to be inversely related to the rolling variance of the series being predicted; see Figure 4.3. These results corroborate the findings of Stock and Watson (2007) for U.S. inflation. Concerning the second point, it is interesting to note that the fluctuations in relative MSPE generally are more pronounced for KRR than for PC-based methods. This suggests that kernel-based forecasts are most valuable during turmoil periods.

Figure 4.4 illustrates these conclusions by showing time series plots of the twelve-month growth rate of Personal Income. The choice of the three subperiods is motivated by dating the Great Moderation in 1984. The first subperiod contains only pre-Moderation data. As we estimate all models on 120-month rolling windows, the first forecast that is based only on post-Moderation data is the one for 1994, which marks the start of the last subperiod. During the second subperiod (see the middle panel of Figure 4.4), the kernel-based forecast is much more volatile than both the actual time series and the PC-based forecast. Apparently, kernel ridge

Figure 4.2: Ten-year rolling-window mean squared prediction errors for four macroeconomic series, for a forecast horizon of $h = 12$ months, for AR and for the best-performing PC and kernel methods.

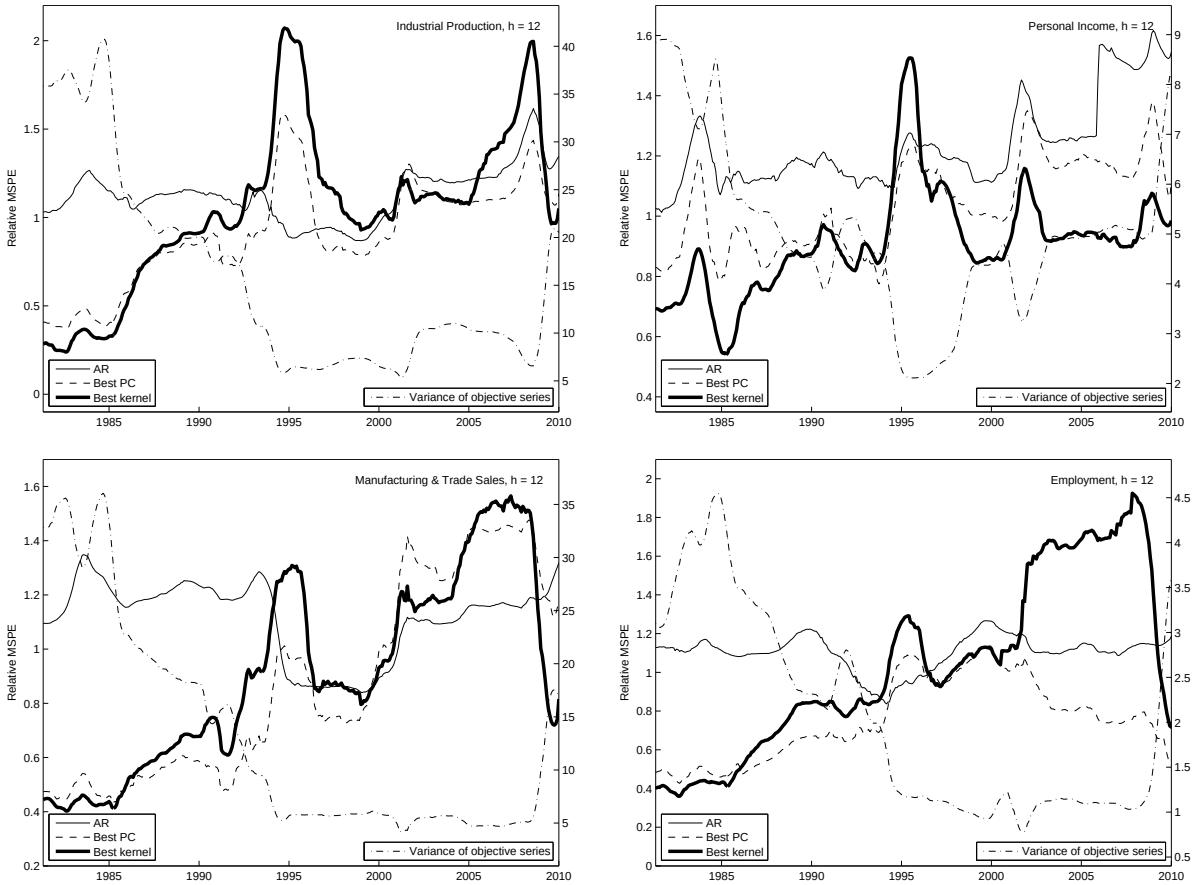


regression is relatively more heavily affected by the break in volatility in the Personal Income series at the Great Moderation (with a variance of 7.84 for 1970-1983, 4.56 for 1984-1993, and 7.75 for 1984-2010). On both other subsamples, however, allowing for nonlinearity through kernel methods enhances the forecast quality considerably, see the top and bottom panels of Figure 4.4. The relative MSPEs, compared to the AR benchmark, for the three subperiods 1970-1983, 1984-1993, and 1994-2010 are respectively 86%, 71%, and 76% for PC, as compared to 70%, 77%, and 67% for Gaussian kernel ridge regression. This result shows that the kernel method performs better than PC in the first and last subperiod. We also note the "overshooting" of the 2008-9 crisis by the PC forecasts in the bottom panel of Figure 4.4. This does not occur for kernel ridge regression, as such extreme forecasts are suppressed by the shrinkage parameter.

4.5 Conclusion

We have introduced kernel ridge regression as a framework for accommodating nonlinear predictive relations in a data-rich environment. We have extended the existing kernel methodology to enable its use in time-series contexts typical for macroeconomic and financial applications. These extensions involve the incorporation of unpenalized linear terms in the forecast equation and an efficient leave-one-out cross-validation procedure for model selection purposes. Our

Figure 4.3: Ten-year rolling-window mean squared prediction errors for four macroeconomic series, for a forecast horizon of $h = 12$ months, for AR and for the best-performing PC and kernel methods, relative to the variance of the series being predicted.

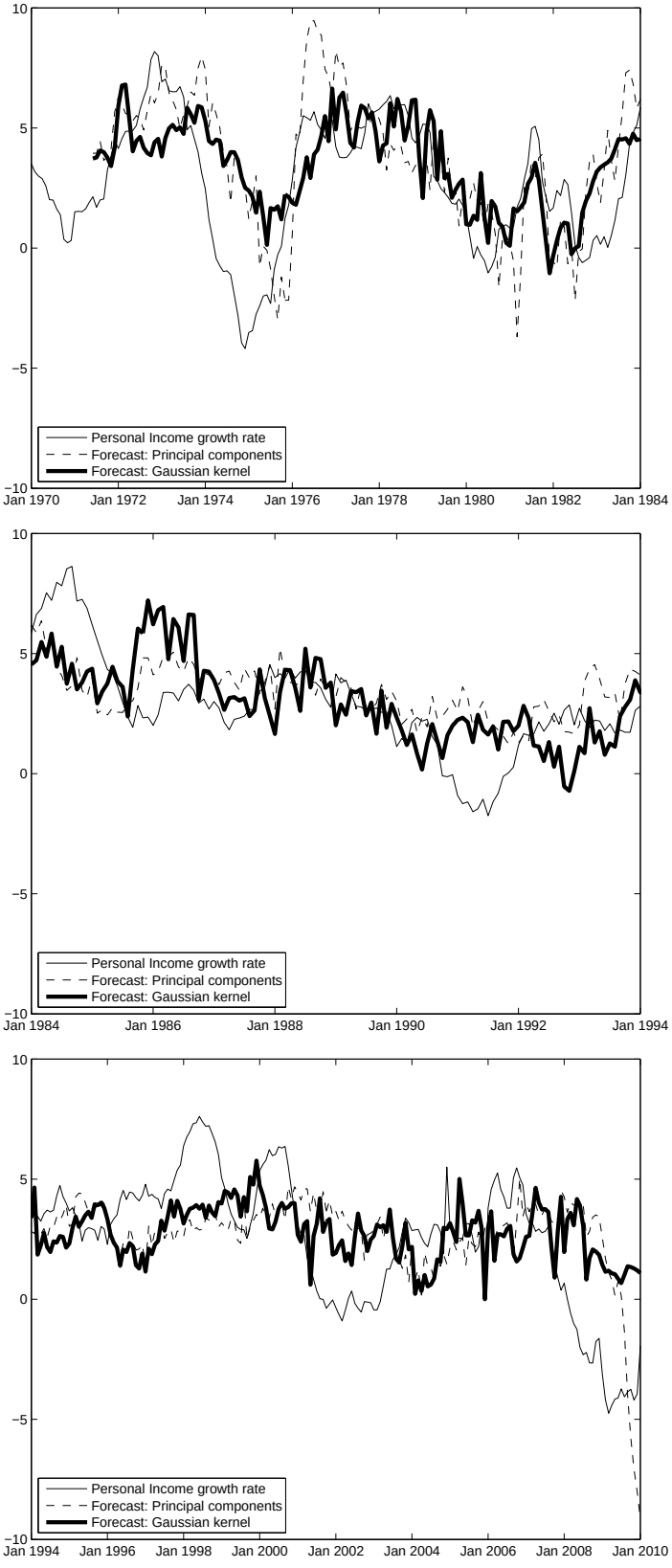


simulation study suggests that this method can deal with the type of data that comes up frequently in economic analysis, namely, data with a factor structure.

The empirical application to forecasting four key U.S. macroeconomic variables — production, income, sales, and employment — shows that kernel-based methods can provide more accurate forecasts than well-established autoregressive and principal-components-based methods. Further, kernel techniques consistently outperform previously proposed nonlinear extensions of the standard PC-based approach. Kernel ridge regression exhibits a relatively consistent good predictive performance, also during the crisis period in 2008-9. Among the kernel methods, linear and Gaussian kernels are found to produce the most reliable forecasts, and neither of these two kernels consistently outperforms the other. This finding implies that the ridge term contributes importantly to the predictive accuracy, while accounting for nonlinearity also helps in many cases. As using the Gaussian kernel does not require the forecaster to specify the form of nonlinearity in advance, this method is a powerful tool.

Finally, we have provided statistical evidence that kernel-based forecasts contain information missed by principal-components-based forecasts, and vice versa. This suggests a potential for forecast combinations. We conclude that the kernel methodology is a valuable addition to the macroeconomic forecaster's toolkit.

Figure 4.4: The twelve-month growth rate of Personal Income (thin line), with its PC-based (dashed line) and Gaussian-kernel forecast (heavy line). Top: 1970-1983. Middle: 1984-1993. Bottom: 1994-2010.



4.A Technical results

This appendix contains derivations of three results stated in Section 4.2. In Appendix 4.A.1 we derive the expression for the forecast equation (4.3) for kernel ridge regression with additional unpenalized linear terms. In Appendix 4.A.2 we obtain the mapping that corresponds to the Gaussian kernel function. Finally, in Appendix 4.A.3 we describe an efficient leave-one-out cross-validation method for selecting tuning parameters in KRR.

4.A.1 Kernel ridge regression with unpenalized linear terms

We have shown in Section 4.2.2 that minimization of the penalized least-squares criterion $||y - Z\gamma||^2 + \lambda ||\gamma||^2$ leads to the forecast $\hat{y}_* = k_*' (K + \lambda I)^{-1} y$ as given in (4.2). In this appendix, we modify this forecast equation to allow for unpenalized linear terms as in the generalized forecast equation $\hat{y}_* = w_*'\hat{\beta} + z_*'\hat{\gamma}$, where the $P \times 1$ vector w_* contains the variables to be treated linearly. In this case, we seek to minimize

$$||y - W\beta - Z\gamma||^2 + \lambda ||\gamma||^2 \tag{4.A.1}$$

over the $P \times 1$ vector β and the $M \times 1$ vector γ. For given $\hat{\beta}$, we can proceed as in Section 4.2.2 to find

$$\hat{\gamma} = Z' (K + \lambda I)^{-1} \left(y - W\hat{\beta}\right). \tag{4.A.2}$$

On the other hand, for given $\hat{\gamma}$, minimizing criterion (4.A.1) is equivalent to ordinary least squares regression, which gives

$$\hat{\beta} = (W'W)^{-1} W' (y - Z\hat{\gamma}). \tag{4.A.3}$$

If we pre-multiply both sides of (4.A.3) by $W'W$, substitute the expression for $\hat{\gamma}$ from (4.A.2) into (4.A.3), and recall that $K = ZZ'$, we get

$$\begin{aligned} W'W\hat{\beta} &= W' \left(y - K (K + \lambda I)^{-1} \left(y - W\hat{\beta}\right)\right) \\ &= W'y - W'K (K + \lambda I)^{-1} y + W'K (K + \lambda I)^{-1} W\hat{\beta}. \end{aligned}$$

Collecting the terms involving $\hat{\beta}$ on the left-hand side of this equation, and rearranging, we obtain

$$W' \left(I - K (K + \lambda I)^{-1}\right) W\hat{\beta} = W' \left(I - K (K + \lambda I)^{-1}\right) y.$$

Using the fact that $I - K (K + \lambda I)^{-1} = (K + \lambda I - K) (K + \lambda I)^{-1} = \lambda (K + \lambda I)^{-1}$, this leads to the expression

$$\hat{\beta} = \left(W' (K + \lambda I)^{-1} W\right)^{-1} W' (K + \lambda I)^{-1} y.$$

If we then substitute this result and (4.A.2) into the forecast equation $\hat{y}_* = z_*'\hat{\gamma} + w_*'\hat{\beta}$, and recall that $k_* = Zz_*$, we find

$$\begin{aligned} \hat{y}_* &= k_*' (K + \lambda I)^{-1} \left(I - W \left(W' (K + \lambda I)^{-1} W\right)^{-1} W' (K + \lambda I)^{-1}\right) y \\ &\quad + w_*' \left(W' (K + \lambda I)^{-1} W\right)^{-1} W' (K + \lambda I)^{-1} y. \end{aligned} \tag{4.A.4}$$

To obtain a more manageable equation, note that by the partitioned matrix inversion formula,

$$\begin{pmatrix} K+\lambda I & W \\ W' & 0 \end{pmatrix}^{-1} = \begin{pmatrix} (K+\lambda I)^{-1}\left(I - WSW'(K+\lambda I)^{-1}\right) & (K+\lambda I)^{-1}WS \\ SW'(K+\lambda I)^{-1} & -S \end{pmatrix}, \tag{4.A.5}$$

where $S = \left(W'(K+\lambda I)^{-1}W\right)^{-1}$. It follows from this result that (4.A.4) is equivalent to the forecast equation (4.3) in Section 4.2.2:

$$\hat{y}_* = \begin{pmatrix} k_* \\ w_* \end{pmatrix}' \begin{pmatrix} K+\lambda I & W \\ W' & 0 \end{pmatrix}^{-1} \begin{pmatrix} y \\ 0 \end{pmatrix}.$$

4.A.2 Expansion of the Gaussian kernel

In this appendix, we derive the mapping φ that corresponds to the Gaussian kernel function. As stated in (4.5) in Section 4.2.3, this kernel function is defined as $\kappa(a,b) = \exp\left(||a-b||^2/2\right)$. If we write $-(1/2)||a-b||^2 = -a'a/2 - b'b/2 + a'b$ and expand the Taylor series for $\exp(a'b)$, we obtain

$$\kappa(a,b) = e^{-a'a/2}\, e^{-b'b/2} \sum_{r=0}^{\infty} \frac{1}{r!}(a'b)^r. \tag{4.A.6}$$

We proceed by expanding $(a'b)^r$ as a multinomial series:

$$(a'b)^r = \left(\sum_{n=1}^{N} a_n b_n\right)^r = \sum\sum\cdots\sum_{\left\{\sum_{n=1}^{N} d_n = r,\ \text{all } d_n \geq 0\right\}} \left(\frac{r!}{\prod_{n=1}^{N} d_n!} \prod_{n=1}^{N} (a_n b_n)^{d_n}\right).$$

Substituting this result into (4.A.6), we find

$$\begin{aligned} \kappa(a,b) &= e^{-a'a/2}\, e^{-b'b/2} \sum_{r=0}^{\infty} \left(\frac{1}{r!} \sum\sum\cdots\sum_{\left\{\sum_{n=1}^{N} d_n = r,\ \text{all } d_n \geq 0\right\}} \left(\frac{r!}{\prod_{n=1}^{N} d_n!} \prod_{n=1}^{N} (a_n b_n)^{d_n}\right)\right) \\ &= e^{-a'a/2}\, e^{-b'b/2} \sum_{r=0}^{\infty} \left(\sum\sum\cdots\sum_{\left\{\sum_{n=1}^{N} d_n = r,\ \text{all } d_n \geq 0\right\}} \left(\prod_{n=1}^{N} \frac{(a_n b_n)^{d_n}}{d_n!}\right)\right) \\ &= e^{-a'a/2}\, e^{-b'b/2} \sum\sum\cdots\sum_{\{\text{all } d_n \geq 0,\ \text{for } n=1,2,\ldots,N\}} \left(\prod_{n=1}^{N} \frac{(a_n b_n)^{d_n}}{d_n!}\right). \end{aligned}$$

Finally, we split the product into two factors that depend only on a and only on b, respectively:

$$\kappa(a,b) = \sum_{d_1=0}^{\infty}\sum_{d_2=0}^{\infty}\cdots\sum_{d_N=0}^{\infty} \left(e^{-a'a/2} \prod_{n=1}^{N} \frac{a_n^{d_n}}{\sqrt{d_n!}}\right)\left(e^{-b'b/2} \prod_{n=1}^{N} \frac{b_n^{d_n}}{\sqrt{d_n!}}\right). \tag{4.A.7}$$

As this expression shows, $\kappa(a,b) = \varphi(a)'\varphi(b)$, where, as claimed in Section 4.2.3, $\varphi(a)$ contains as elements, for each combination of degrees $d_1, d_2, \ldots, d_N \geq 0$,

$$e^{-a'a/2} \prod_{n=1}^{N} \frac{a_n^{d_n}}{\sqrt{d_n!}}.$$

4.A.3 Computationally efficient leave-one-out cross-validation

In this appendix, we describe an efficient method for leave-one-out cross-validation, which we employ to select the tuning parameters in KRR. Our derivation extends the results in Cawley and Talbot (2008) to allow for the unpenalized linear terms in the forecast equation (4.3). The result of Appendix 4.A.1 can be summarized as follows: kernel ridge regression leads to the forecast

$$\hat{y}_* = \begin{pmatrix} k_* \\ w_* \end{pmatrix}' \begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} \quad \text{with} \quad \begin{pmatrix} K + \lambda I & W \\ W' & 0 \end{pmatrix} \begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} = \begin{pmatrix} y \\ 0 \end{pmatrix}. \tag{4.A.8}$$

The first step in leave-one-out cross-validation is to estimate the model on all observations except the first. As $K = ZZ'$, and each row of Z depends only on the corresponding row of X, the only elements in K that depend on the first observation are those in the first row and those in the first column. We therefore separate the first row and column from the other elements of K, and likewise, we split off the first row of W, the first element of $\hat{\alpha}$, and the first element of y. We denote these partitioned matrices and vectors by

$$K = \begin{pmatrix} k_{1,1} & k'_{-1,1} \\ k_{-1,1} & K_{-1,-1} \end{pmatrix}, \quad W = \begin{pmatrix} w'_1 \\ W_{-1} \end{pmatrix}, \quad \hat{\alpha} = \begin{pmatrix} \hat{\alpha}_1 \\ \hat{\alpha}_{-1} \end{pmatrix} \quad \text{and} \quad y = \begin{pmatrix} y_1 \\ y_{-1} \end{pmatrix}.$$

We then have from (4.A.8)

$$\begin{pmatrix} k_{1,1} + \lambda & k'_{-1,1} & w'_1 \\ k_{-1,1} & K_{-1,-1} + \lambda I & W_{-1} \\ w_1 & W'_{-1} & 0 \end{pmatrix} \begin{pmatrix} \hat{\alpha}_1 \\ \hat{\alpha}_{-1} \\ \hat{\beta} \end{pmatrix} = \begin{pmatrix} y_1 \\ y_{-1} \\ 0 \end{pmatrix},$$

or equivalently, separating the first equation from the others,

$$\hat{\alpha}_1 (k_{1,1} + \lambda) + \begin{pmatrix} k_{-1,1} \\ w_1 \end{pmatrix}' \begin{pmatrix} \hat{\alpha}_{-1} \\ \hat{\beta} \end{pmatrix} = y_1, \tag{4.A.9}$$

$$\hat{\alpha}_1 \begin{pmatrix} k_{-1,1} \\ w_1 \end{pmatrix} + \begin{pmatrix} K_{-1,1} + \lambda I & W_{-1} \\ W'_{-1} & 0 \end{pmatrix} \begin{pmatrix} \hat{\alpha}_{-1} \\ \hat{\beta} \end{pmatrix} = \begin{pmatrix} y_{-1} \\ 0 \end{pmatrix}. \tag{4.A.10}$$

The forecast of y_1 based on a model estimated on observations $2, 3, \ldots, T$ clearly equals

$$\tilde{y}_1 = \begin{pmatrix} k_{-1,1} \\ w_1 \end{pmatrix}' \begin{pmatrix} K_{-1,-1} + \lambda I & W_{-1} \\ W'_{-1} & 0 \end{pmatrix}^{-1} \begin{pmatrix} y_{-1} \\ 0 \end{pmatrix}$$

and using (4.A.9) and (4.A.10) we may write

$$\begin{aligned} \tilde{y}_1 &= \hat{\alpha}_1 \begin{pmatrix} k_{-1,1} \\ w_1 \end{pmatrix}' \begin{pmatrix} K_{-1,-1} + \lambda I & W_{-1} \\ W'_{-1} & 0 \end{pmatrix}^{-1} \begin{pmatrix} k_{-1,1} \\ w_1 \end{pmatrix} + \begin{pmatrix} k_{-1,1} \\ w_1 \end{pmatrix}' \begin{pmatrix} \hat{\alpha}_{-1} \\ \hat{\beta} \end{pmatrix} \\ &= \hat{\alpha}_1 \begin{pmatrix} k_{-1,1} \\ w_1 \end{pmatrix}' \begin{pmatrix} K_{-1,-1} + \lambda I & W_{-1} \\ W'_{-1} & 0 \end{pmatrix}^{-1} \begin{pmatrix} k_{-1,1} \\ w_1 \end{pmatrix} + y_1 - \hat{\alpha}_1 (k_{1,1} + \lambda) \\ &= y_1 - \hat{\alpha}_1 \left(k_{1,1} + \lambda - \begin{pmatrix} k_{-1,1} \\ w_1 \end{pmatrix}' \begin{pmatrix} K_{-1,-1} + \lambda I & W_{-1} \\ W'_{-1} & 0 \end{pmatrix}^{-1} \begin{pmatrix} k_{-1,1} \\ w_1 \end{pmatrix} \right). \end{aligned}$$

The expression $k_{1,1} + \lambda - \begin{pmatrix} k_{-1,1} \\ w_1 \end{pmatrix}' \begin{pmatrix} K_{-1,-1} + \lambda I & W_{-1} \\ W'_{-1} & 0 \end{pmatrix}^{-1} \begin{pmatrix} k_{-1,1} \\ w_1 \end{pmatrix}$ is equal to the reciprocal of element $(1,1)$ of $\begin{pmatrix} k_{1,1} + \lambda & k'_{-1,1} & w'_1 \\ k_{-1,1} & K_{-1,-1} + \lambda I & W_{-1} \\ w_1 & W'_{-1} & 0 \end{pmatrix}^{-1} = \begin{pmatrix} K + \lambda I & W \\ W' & 0 \end{pmatrix}^{-1}$, as can be seen by using the partitioned matrix inversion formula. Therefore, the first leave-one-out error equals

$$y_1 - \tilde{y}_1 = \hat{\alpha}_1 \Big/ \text{element } (1,1) \text{ of } \begin{pmatrix} K + \lambda I & W \\ W' & 0 \end{pmatrix}^{-1}.$$

In general, an analogous derivation shows that the t-th leave-one-out prediction error equals

$$y_t - \tilde{y}_t = \hat{\alpha}_t \Big/ \text{element } (t,t) \text{ of } \begin{pmatrix} K + \lambda I & W \\ W' & 0 \end{pmatrix}^{-1}. \tag{4.A.11}$$

That is, we can compute all leave-one-out errors by dividing each element of the vector $\hat{\alpha}$ by the corresponding diagonal element of the matrix $\begin{pmatrix} K + \lambda I & W \\ W' & 0 \end{pmatrix}^{-1}$. Observe that both $\hat{\alpha}$ and this matrix inverse are needed in computing the forecast $\hat{y}_*$. Thus, in the process of making the out-of-sample prediction, we can find all leave-one-out errors without performing any additional computations, aside from the division in (4.A.11).

As a final note, we mention that the matrix inverse in (4.A.11) can also be computed efficiently. As $K + \lambda I$ is symmetric and positive definite, its inverse can be computed from its Cholesky decomposition. The inverse of the full matrix can then be calculated using (4.A.5) in Appendix 4.A.1.

Chapter 5

Modelling Issues in Kernel Ridge Regression

This chapter is based on Exterkate (2011).

5.1 Introduction

In many areas of application, forecasters face a trade-off between model complexity and forecast accuracy. Due to the uncertainty associated with model choice and parameter estimation, a highly complex nonlinear predictive model is often found to produce less accurate forecasts than a simpler, e.g. linear, model. Thus, a researcher wishing to estimate a nonlinear relation for forecasting purposes generally restricts the search space drastically, for example to polynomials of low degree, or to regime-switching models (Teräsvirta, 2006) or neural networks (White, 2006). A recent comprehensive overview was given by Kock and Teräsvirta (2011). The improvement of such models upon the predictive accuracy of linear models is often found to be limited, see Stock and Watson (1999), Teräsvirta et al. (2005), and Medeiros et al. (2006), among others.

Another manifestation of this complexity-accuracy trade-off is that, while a very large number of potentially relevant predictors may be available, the *curse of dimensionality* implies that better forecasts can be obtained if a large proportion of the predictors is discarded. This situation arises, for example, in economic applications. Hundreds or even thousands of predictors are often available, and economic theory does not usually provide guidelines concerning which variables should or should not influence each other. A reduction in the number of predictors can of course be achieved by selecting a small subset of representative variables, but the most common way to proceed is to summarize the predictors by a small number of principal components. This approach has found successful forecasting applications in macroeconomics (e.g. Stock and Watson, 2002) and in finance (e.g. Ludvigson and Ng, 2007, 2009).

In this chapter we discuss *kernel ridge regression*, a forecasting technique that can overcome both aspects of this trade-off simultaneously, making it suitable for estimating nonlinear models with many predictors. While kernel methods are not widely known in the fields of economics and finance, they have found ample applications in machine learning; a recent review can be found in Hofmann et al. (2008). A typical application is classification, such as optical

recognition of scanned handwritten characters (Schölkopf et al., 1998). Recently, Exterkate et al. (2011a) use this technique in a macroeconomic forecasting application and they report an increase in forecast accuracy, compared to traditional linear methods.

The central idea in kernel ridge regression is to employ a flexible set of nonlinear prediction functions and to prevent overfitting by penalization, in a way that limits the computational complexity. This is achieved by mapping the set of predictors into a high-dimensional (or even infinite-dimensional) space of nonlinear functions of the predictors. A forecast equation is estimated in this high-dimensional space, using a penalty (or shrinkage, or ridge) term to avoid overfitting. Computational tractability is achieved by choosing the kernel in a convenient way, so that calculations in the high-dimensional space are actually prevented.

Kernel ridge regression provides the practitioner with a large amount of flexibility, but it also leaves him with a number of nontrivial decisions to make. One such decision concerns which kernel to use. Although any choice of kernel leads to restrictions on the functional form of the forecast equation, little attention is generally paid to such implications. Additionally, kernel ridge regression involves tuning parameters, and their practical interpretation is not always clear. This feature makes it difficult to select "reasonable" values for these parameters, resulting in time-consuming grid searches or in suboptimal performance.

To give a clear interpretation of the kernel functions and their associated tuning parameters, we review the kernel methodology from two different points of view, namely, function approximation and Bayesian statistics. This combination of perspectives enables us to relate one of the two tuning parameters that are found in most applications of kernel ridge regression to the signal-to-noise ratio in the data, and the other to smoothness measures of the prediction function. Based on these insights, we give explicit rules of thumb for selecting their values by using cross-validation over small grids. Cross-validation may also be used to select among different types of kernel. However, one needs to be somewhat careful with this procedure: we provide evidence against including the popular polynomial kernels in the cross-validation exercise.

In Section 5.2 we describe the kernel methodology, from the perspective of function approximation and from Bayesian statistics. We discuss several popular kernels and the functional forms of the associated forecast equations, and we interpret their tuning parameters. Section 5.3 presents a simulation study to show the effects of choosing the kernel or its tuning parameters incorrectly. Concerning the tuning parameters, selecting them using cross-validation from our grids affects the forecast quality only marginally, compared to using the true values. The choice of kernel can also be left to cross-validation; however, using a polynomial kernel when the data-generating process is non-polynomial, or vice versa, reduces forecast accuracy. We also present simulations in which all kernels estimate misspecified models, and we find that the Gaussian and Sinc kernels outperform polynomial kernels. We provide conclusions in Section 5.4.

5.2 Methodology

Kernel ridge regression can be understood as a function approximation tool, but it can also be given a Bayesian interpretation. We review the method from both viewpoints in Sections 5.2.1 and 5.2.2, respectively. We present some popular kernel functions in Section 5.2.3. In Section 5.2.4 we give an interpretation to the associated tuning parameters, and we derive "reasonable" values for these parameters.

5.2.1 Kernel ridge regression for function approximation

We first introduce some notation. We are given T observations $(y_1, x_1), (y_2, x_2), \ldots, (y_T, x_T)$, with $y_t \in \mathbb{R}$ and $x_t \in \mathbb{R}^N$, and our goal is to find a function f so that $f(x_t)$ is a "good" approximation to y_t for all $t = 1, 2, \ldots, T$. Then, we are given a new observation $x_* \in \mathbb{R}^N$ and asked to predict the corresponding y_*. We denote this prediction by $\hat{y}_* = f(x_*)$. By selecting f from a large and flexible class of functions while preventing overfitting, we hope to achieve that this prediction is accurate.

To describe the class of functions from which we select f, we first choose a function $\varphi : \mathbb{R}^N \to \mathbb{R}^M$. The regression function f will be restricted to a certain set of linear combinations of the form $\varphi(x)' \gamma$, with $\gamma \in \mathbb{R}^M$. The number of regressors M is either a finite integer with $M \geq N$, or $M = \mathbb{N}$, representing a countably infinite number of regressors. Examples of both types are presented in Section 5.2.3 below.

If a flexible functional form is desired, the number of regressors M needs to be large. Therefore we wish to avoid M-dimensional computations, and it turns out that we can do so by requiring only that the dot product $\kappa(x_s, x_t) = \varphi(x_s)' \varphi(x_t)$ can be found using only N-dimensional computations, for any $x_s, x_t \in \mathbb{R}^N$. The function $\kappa : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}$ is commonly called the kernel function. Conversely, functions κ for which a corresponding φ exists can be characterized by a set of conditions due to Mercer (1909). All kernel functions discussed in this study satisfy these conditions; a thorough justification can be found in Hofmann et al. (2008).

Finally, define a space of functions $\mathcal{H}_0$ which contains $f : \mathbb{R}^N \to \mathbb{R}$ if and only if there exists a finite set $x_1^f, x_2^f, \ldots, x_S^f \in \mathbb{R}^N$ and real numbers $\alpha_1^f, \alpha_2^f, \ldots, \alpha_S^f$ such that $f(x) = \sum_{s=1}^S \alpha_s^f \kappa\left(x, x_s^f\right)$. Every such $f(x)$ is a linear combination of the elements of $\varphi(x)$, as can be seen by recalling the definition of κ: we have $f(x) = \varphi(x)' \left(\sum_{s=1}^S \alpha_s^f \varphi\left(x_s^f\right)\right)$. We equip $\mathcal{H}_0$ with the following dot product:

$$\text{if } f(x) = \sum_{s=1}^{S} \alpha_s^f \kappa\left(x, x_s^f\right) \text{ and } g(x) = \sum_{s'=1}^{S'} \alpha_{s'} \kappa\left(x, x_{s'}^g\right),$$

$$\text{then } \langle f, g \rangle_{\mathcal{H}} = \sum_{s=1}^{S} \sum_{s'=1}^{S'} \alpha_s^f \alpha_{s'}^g \kappa\left(x_s^f, x_{s'}^g\right).$$

(For the verification that $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ is indeed a valid dot product, see Hofmann et al. (2008).) Finally, Aronszajn (1950) proved that completing $\mathcal{H}_0$ in the corresponding norm $||f||_{\mathcal{H}} = \sqrt{\langle f, f \rangle_{\mathcal{H}}}$ leads to a Hilbert space, which we call $\mathcal{H}$. This is the class of functions from which f will be selected.

In finite samples, an unrestricted search over the space $\mathcal{H}$ will lead to overfitting. Indeed, if $\mathcal{H}$ allows for sufficiently flexible functional forms, a prediction function f may be obtained for which the in-sample fit is perfect, but the out-of-sample predictive accuracy will generally be poor. Therefore, we consider the regularized problem

$$\min_{f \in \mathcal{H}} \sum_{t=1}^{T} \left(y_t - f(x_t)\right)^2 + \lambda ||f||_{\mathcal{H}}^2, \tag{5.1}$$

for some $\lambda > 0$. A result due to Kimeldorf and Wahba (1971), known as the *representer theorem*, states that the minimizer of this problem can be written as $f(x) = \sum_{t=1}^{T} \alpha_t \kappa(x, x_t)$, for

some sequence of real numbers $\alpha_1, \alpha_2, \ldots, \alpha_T$. That is, the optimal prediction function admits a kernel expansion in terms of the observations: the set of expansion points $\left\{x_1^f, x_2^f, \ldots, x_S^f\right\}$ may be taken equal to $\{x_1, x_2, \ldots, x_T\}$.

If we define $y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{pmatrix}$, $\alpha = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_T \end{pmatrix}$, and $K = \begin{pmatrix} \kappa(x_1, x_1) & \cdots & \kappa(x_1, x_T) \\ \kappa(x_2, x_1) & \cdots & \kappa(x_2, x_T) \\ \vdots & \ddots & \vdots \\ \kappa(x_T, x_1) & \cdots & \kappa(x_T, x_T) \end{pmatrix}$, we see that problem (5.1) is equivalent to

$$\min_{\alpha \in \mathbb{R}^T} (y - K\alpha)'(y - K\alpha) + \lambda \alpha' K \alpha. \tag{5.2}$$

Minimizing the quadratic form in (5.2) yields $\alpha = (K + \lambda I)^{-1} y$, where I is the $T \times T$ identity matrix. Finally, to forecast a new observation y_* if the corresponding x_* is given, we have

$$\hat{y}_* = f(x_*) = \sum_{t=1}^{T} \alpha_t \kappa(x_*, x_t) = k_*' \alpha = k_*' (K + \lambda I)^{-1} y, \tag{5.3}$$

where the vector $k_* \in \mathbb{R}^T$ has $\kappa(x_*, x_t)$ as its t-th element.

5.2.2 Kernel ridge regression for Bayesian prediction

In this section we retain the notation introduced above, but our point of view is different. We assume that, conditional on x_t, each y_t has a normal distribution, with mean $\varphi(x_t)'\gamma$ for some $\gamma \in \mathbb{R}^M$, and with some fixed variance θ^2. If we let Z be the $T \times M$ matrix[1] with t-th row equal to $\varphi(x_t)'$, the probability density function may be written as

$$p\left(y|Z, \gamma, \theta^2\right) \propto \left(\theta^2\right)^{-\frac{T}{2}} \exp\left(\frac{-1}{2\theta^2}(y - Z\gamma)'(y - Z\gamma)\right).$$

We specify our prior beliefs about γ and θ^2 as follows. We take the uninformative Jeffreys prior on θ^2 and, given θ^2, our prior on the distribution of γ is normal with mean zero and variance $(\theta^2/\lambda) I$:

$$p\left(\theta^2\right) \propto \left(\theta^2\right)^{-1}, \qquad p\left(\gamma|\theta^2\right) \propto \left(\theta^2\right)^{-\frac{M}{2}} \exp\left(\frac{-\lambda}{2\theta^2}\gamma'\gamma\right).$$

Using Bayes' rule, the posterior density of the parameters is given by

$$\begin{aligned} p\left(\gamma, \theta^2|Z, y\right) &\propto p\left(y|Z, \gamma, \theta^2\right) p\left(\gamma|\theta^2\right) p\left(\theta^2\right) \\ &\propto \left(\theta^2\right)^{-\frac{T+M+2}{2}} \exp\left(\frac{-1}{2\theta^2}\left[(y - Z\gamma)'(y - Z\gamma) + \lambda\gamma'\gamma\right]\right), \end{aligned}$$

see e.g. Raiffa and Schlaifer (1961). Now, for a new observation $x_* \in \mathbb{R}^N$, denote $z_* = \varphi(x_*)$ and assume that, just like $y_1, y_2, \ldots, y_T$, the unobserved y_* follows the normal distribution

$$p\left(y_*|z_*, \gamma, \theta^2, Z, y\right) \propto \left(\theta^2\right)^{-\frac{1}{2}} \exp\left(\frac{-1}{2\theta^2}(y_* - z_*'\gamma)^2\right).$$

[1] If M is infinite, applying the derivations in this section to a finite subset of the regressors and then letting $M \to \infty$ leads to the same final results.

Then, again by Bayes' rule, the predictive density of y_*, given all observed data, is

$$p\left(y_*|z_*, Z, y\right) = \int_{\mathbb{R}^M} \int_0^\infty p\left(y_*|z_*, \gamma, \theta^2, Z, y\right) p\left(\gamma, \theta^2|Z, y\right) d\theta^2 d\gamma$$

$$= \int_{\mathbb{R}^M} \int_0^\infty \left(\theta^2\right)^{-\frac{T+M+3}{2}} \exp\left(\frac{-1}{2\theta^2}\left[\left(y - Z\gamma\right)'\left(y - Z\gamma\right) + \left(y_* - z_*'\gamma\right)^2 + \lambda\gamma'\gamma\right]\right) d\theta^2 d\gamma.$$

This integral can be evaluated analytically (see e.g. Raiffa and Schlaifer, 1961) and the resulting predictive density has $\hat{y}_*$ from (5.3) as its mean, median, and mode. More precisely, introducing $k_{**} = z_*'z_*$ and

$$w = \frac{1}{T} y'\left(K + \lambda I\right)^{-1} y\left(k_{**} + \lambda - k_*'\left(K + \lambda I\right)^{-1} k_*\right),$$

the quantity $w^{-1/2}\left(y_* - \hat{y}_*\right)$ follows Student's t distribution with T degrees of freedom.

That is, two different approaches to forecasting y_* in terms of linear combinations of certain functions of x_* lead to the same point forecast $\hat{y}_*$. We shall exploit both points of view in the next section, which describes some common kernel functions, and in Section 5.2.4, where we discuss the associated tuning parameters.

5.2.3 Some popular kernel functions

A first obvious way of introducing nonlinearity in the prediction function $f\left(x\right) = \varphi\left(x\right)'\gamma$ is by making it a polynomial of some specified degree d. That is, we choose φ in such a way that $\varphi\left(x\right)$ contains all $\begin{pmatrix} N+d \\ d \end{pmatrix}$ monomials of the form $x_1^{d_1} x_2^{d_2} \cdots x_N^{d_N}$, with all d_n nonnegative integers with $\sum_{n=1}^N d_n \leq d$. As shown by Poggio (1975), the kernel function takes a simple form if we multiply each monomial by a constant: if a typical element of $\varphi\left(x\right)$ is

$$\left(\sigma^{-\sum_{n=1}^N d_n}\right) \sqrt{\frac{d!}{\left(d - \sum_{n=1}^N d_n\right)! \prod_{n=1}^N d_n!}} \prod_{n=1}^N x_n^{d_n}, \tag{5.4}$$

where $\sigma > 0$ is a tuning parameter, then the kernel function is simply

$$\kappa\left(x_s, x_t\right) = \left(1 + \frac{x_s' x_t}{\sigma^2}\right)^d. \tag{5.5}$$

A more sophisticated method for constructing kernels is to require that the resulting prediction function must be smooth in some sense. From the point of view of function approximation, this is a sensible requirement, as we do not want to overfit the data. In the context of Section 5.2.1, we can achieve this by selecting κ to generate a Hilbert space $\mathcal{H}$ for which $||f||_{\mathcal{H}}$ measures lack of smoothness of f; see the objective (5.1).

Following Smola et al. (1998), we restrict ourselves to functions f for which $\int_{\mathbb{R}^N} f\left(x\right)^2 dx$ is finite, and we measure the smoothness of such a function by examining its Fourier transform, defined by

$$\tilde{f} : \mathbb{R}^N \to \mathbb{R} \quad \text{with} \quad \tilde{f}\left(\omega\right) = \left(2\pi\right)^{-\frac{N}{2}} \int_{\mathbb{R}^N} \exp\left(-i\omega' x\right) f\left(x\right) dx.$$

The Fourier transform decomposes f according to frequency.[2] That is, if $\tilde{f}(\omega)$ takes large values for large values of $||\omega||$, this indicates that $f(x)$ fluctuates rapidly with x, i.e., that f is not smooth. It follows that lack of smoothness of f can be penalized by choosing κ in such a way that

$$||f||_{\mathcal{H}} = (2\pi)^{-N} \int_{\mathbb{R}^N} \frac{\left|\tilde{f}(\omega)\right|^2}{v(\omega)} d\omega, \tag{5.6}$$

where $(2\pi)^{-N}$ is a normalization constant, $|\cdot|$ denotes the absolute value of a complex number, and $v : \mathbb{R}^N \to \mathbb{R}$ is a suitably chosen penalization function. As explained, we want to penalize mainly the high-frequency components of f; thus, we choose v such that $v(\omega)$ is close to zero for large $||\omega||$.

Hofmann et al. (2008) show that it is possible to select a kernel function κ so that (5.6) holds, for any function v that satisfies the regularity conditions $v(\omega) \geq 0$, $\int_{\mathbb{R}^N} v(\omega)\, d\omega = 1$, and $v(\omega)$ is symmetric in ω. Specifically, they derive from a theorem of Bochner (1933) that the kernel function

$$\kappa(x_s, x_t) = (2\pi)^{\frac{N}{2}} \tilde{v}(x_s - x_t) \tag{5.7}$$

satisfies the Mercer (1909) conditions and leads to a norm $||\cdot||_{\mathcal{H}}$ that penalizes lack of smoothness as in (5.6).

We will now discuss two kernels that can be derived using (5.7). A popular choice is to use

$$v(\omega) = \left(\frac{2\pi}{\sigma^2}\right)^{-\frac{N}{2}} \exp\left(-\frac{\sigma^2}{2}\omega'\omega\right), \tag{5.8}$$

where $\sigma > 0$ is a tuning parameter. Components of f with frequency ω are penalized more heavily if $||\omega||$ is larger, and high-frequency components are more severely penalized for larger values of σ. It can be shown that substituting (5.8) into (5.7) yields

$$\kappa(x_s, x_t) = \exp\left(\frac{-1}{2\sigma^2}||x_s - x_t||^2\right), \tag{5.9}$$

where $||\cdot||$ is the usual Euclidean norm. Function (5.9), introduced by Broomhead and Lowe (1988), is known as the Gaussian kernel.

Notice that the Gaussian kernel allows all frequencies to be present in the prediction function f, albeit with very large penalties for high frequencies. One may alternatively choose to set an infinitely large penalty on certain frequencies ω by setting $v(\omega) = 0$, thereby explicitly disallowing noisy behavior of f.[3] One obvious way to accomplish this is by using the uniform penalty function

$$v(\omega) = \begin{cases} \left(\frac{\sigma}{2}\right)^N & \text{if } -\frac{1}{\sigma} < \omega_n < \frac{1}{\sigma} \text{ for all } n = 1, 2, \ldots, N; \\ 0 & \text{otherwise,} \end{cases} \tag{5.10}$$

[2]Note that $\tilde{f}(\omega)$ is symmetric in ω, so that no interpretation difficulties arise from "negative frequencies".

[3]More formally, we exclude the region where $v(\omega) = 0$ from the domain of integration in (5.6), and we restrict f to have $\tilde{f}(\omega) = 0$ in that region.

where again, $\sigma > 0$ is a tuning parameter. Substituting (5.10) into (5.7) yields the corresponding kernel function

$$\kappa(x_s, x_t) = \prod_{n=1}^{N} \text{sinc}\left(\frac{x_{sn} - x_{tn}}{\sigma}\right), \tag{5.11}$$

the Sinc kernel (see Yao, 1967), where $\text{sinc}(0) = 1$ and $\text{sinc}(u) = \sin(u)/u$ for all $u \neq 0$. Despite its intuitive interpretation given in (5.10), the Sinc kernel does not seem to have found wide application in the kernel literature.

As mentioned before, all kernels discussed in this study have the property that there exists a mapping φ such that $\kappa(x_s, x_t) = \varphi(x_s)' \varphi(x_t)$. However, the kernel functions derived here are much more easily understood by studying how v penalizes certain frequencies than by explicitly finding the regressors $\varphi(x)$. As an example, Exterkate et al. (2011a) derive the following expression for $\varphi(x)$ for the Gaussian kernel: it contains, for each combination of nonnegative degrees $d_1, d_2, \ldots, d_N$, the "dampened polynomial"

$$\left(\sigma^{-\sum_{n=1}^{N} d_n}\right) \exp\left(\frac{-x'x}{2\sigma^2}\right) \prod_{n=1}^{N} \frac{x_n^{d_n}}{\sqrt{d_n!}}.$$

5.2.4 Tuning parameters

Two tuning parameters have been introduced in our discussion of kernel ridge regression: a penalization parameter λ and a kernel-specific tuning parameter σ. In this section, we give an interpretation to both of these parameters. This interpretation will result in a small grid of "reasonable" values for both tuning parameters. Selection from this grid, as well as selection of the kernel function, can then be performed using leave-one-out cross-validation; see Cawley and Talbot (2008) for a computationally efficient implementation of this procedure. We choose this selection mechanism because of its close resemblance to the task at hand: the out-of-sample forecasting of the value of the dependent variable for one observation.

The parameter λ is most easily understood from the Bayesian point of view. We assumed that, conditional on x_t and the model parameters, y_t is normally distributed with mean $\varphi(x_t)'\gamma$ and variance θ^2. Equivalently, we may decompose y_t into signal and noise components: $y_t = \varphi(x_t)'\gamma + \varepsilon_t$, with $\varepsilon_t \sim \mathcal{N}(0, \theta^2)$. The entire analysis in Section 5.2.2 was conditional on x_t. If we now treat x_t as a random variable, of which the priors on γ and θ are independent, we can write

$$\begin{aligned}
\text{var}\left(\varphi(x_t)'\gamma\right) &= \text{E}\left(\varphi(x_t)'\gamma\gamma'\varphi(x_t)\right) = \text{E}\left(\text{E}\left(\varphi(x_t)'\gamma\gamma'\varphi(x_t)\,|x_t\right)\right) \\
&= \text{E}\left(\varphi(x_t)'\left(\frac{\theta^2}{\lambda} I\right)\varphi(x_t)\right) = \frac{\theta^2}{\lambda}\text{E}\left(\varphi(x_t)'\varphi(x_t)\right) = \frac{\theta^2}{\lambda}\text{E}\left(\kappa(x_t, x_t)\right).
\end{aligned}$$

This result enables us to relate λ to the signal-to-noise ratio,

$$\psi = \frac{\text{var}\left(\varphi(x_t)'\gamma\right)}{\text{var}(\varepsilon_t)} = \frac{\frac{\theta^2}{\lambda}\text{E}\left(\kappa(x_t, x_t)\right)}{\theta^2} = \frac{\text{E}\left(\kappa(x_t, x_t)\right)}{\lambda}. \tag{5.12}$$

For the Gaussian kernel (5.9) and the Sinc kernel (5.11), $\kappa(x_t, x_t) = 1$ does not depend on x_t and the signal-to-noise ratio is simply $\psi = 1/\lambda$. For the polynomial kernels (5.5), the

signal-to-noise ratio is still inversely proportional to λ, but the proportionality constant depends on the distribution of x_t. For example, if we assume that $x_t \sim \mathcal{N}(0, I)$, then $x_t' x_t$ follows a χ^2 distribution with N degrees of freedom, and hence

$$\psi = \frac{1}{\lambda}\mathrm{E}\left(\left(1 + \frac{x_t' x_t}{\sigma^2}\right)^d\right) = \frac{1}{\lambda}\sum_{j=0}^{d}\binom{d}{j}\sigma^{-2j}\mathrm{E}\left(\left(x_t' x_t\right)^j\right) = \frac{1}{\lambda}\sum_{j=0}^{d}\binom{d}{j}\sigma^{-2j}\prod_{i=0}^{j-1}(N+2i)\,. \tag{5.13}$$

We see that in all cases, the "correct" value of λ could be obtained if the signal-to-noise ratio ψ were known. We propose the following simple procedure for estimating ψ: obtain the R^2 from linear OLS regression of y on a constant and X (or, if N is not small relative to T, on a small number of principal components of X). If the estimated linear model were the true model, we would have $\psi_0 = R^2/(1-R^2)$, and its corresponding λ_0 can be found using (5.12). As one expects to obtain a better fit using nonlinear models, it is likely that a $\lambda < \lambda_0$ is required, and we propose to select λ from the grid $\left\{\frac{1}{8}\lambda_0, \frac{1}{4}\lambda_0, \frac{1}{2}\lambda_0, \lambda_0, 2\lambda_0\right\}$. The simulation study in Section 5.3 confirms that this grid is sufficiently fine, as well as the grids for σ that we define below.

On the other hand, the parameter σ is best understood in the context of function approximation. For the Gaussian and Sinc kernels, its interpretation is clear from the penalty functions v introduced in the previous section: a higher value of σ forces the prediction function to be smoother. For the sinc kernel, this works by explicitly disallowing components of frequency greater than $1/\sigma$. Recall that a component of $f(x)$ with a frequency of $1/\sigma$ oscillates $1/(2\pi\sigma)$ times as x changes by one unit. As we will always studentize the predictors, a one-unit change is equivalent to a one-standard-deviation change. We select a grid that implies that such a change in x may never result in more than two oscillations: $\left\{\frac{1}{4\pi}, \frac{1}{2\pi}, \frac{1}{\pi}, \frac{2}{\pi}, \frac{4}{\pi}\right\}$.

For the Gaussian kernel, although all frequencies are allowed, the penalty function (5.8) decreases to zero faster for larger values of σ. In fact, along each dimension, 95% of its mass lies in the interval $(-2/\sigma, 2/\sigma)$, leaving very little mass (that is, very high penalties) for frequencies greater than $2/\sigma$. Therefore, the same reasoning as above leads to a grid in which all values are twice those in the grid for the Sinc kernel: $\left\{\frac{1}{2\pi}, \frac{1}{\pi}, \frac{2}{\pi}, \frac{4}{\pi}, \frac{8}{\pi}\right\}$.

Finally, for the polynomial kernels, the contributions of terms of different orders to the variance of y_t are given in (5.13). Irrespective of the distribution of x_t, a higher value of σ allows higher-order terms to contribute less. Thus, as for the other kernels, a higher σ imposes more smoothness on the function f. To derive a rule of thumb for the σ grid, we propose that in most applications the first-order effects should dominate in terms of variance contributions, followed by the second-order, third-order, etc. If we assume that $x_t \sim \mathcal{N}(0, 1)$, we can derive from the right-hand-side of (5.13) that this ordering is preserved if $\sigma > \sigma_0 = \sqrt{(d-1)(N+2)/2}$. Thus, for $d > 1$ we select σ from the grid $\left\{\frac{1}{2}\sigma_0, \sigma_0, 2\sigma_0, 4\sigma_0, 8\sigma_0\right\}$. For $d = 1$, this formula yields $\sigma_0 = 0$, which cannot be used. We set $\sigma_0 = \sqrt{N/2}$ instead and construct the grid in the same manner.

5.3 Monte Carlo simulation

In order to assess the empirical validity of the rules of thumb for selecting tuning parameters described in Section 5.2.4, and to investigate the impact of kernel choice on forecast quality, we perform two simulation studies. In the first Monte Carlo study, the data-generating processes

correspond to the kernels discussed in Section 5.2.3. For estimation, we consider four different cases:

- treating the kernel and the tuning parameters as known;
- treating the kernel as known, but selecting the tuning parameters using cross-validation;
- deliberately picking an incorrect kernel, and selecting the tuning parameters using cross-validation; and
- selecting the kernel and the tuning parameters jointly using cross-validation.

In the second Monte Carlo experiment, the data-generating process is such that all kernels estimate a misspecified model. This experiment is intended to resemble practical situations, in which nothing is known about the data-generating process.

5.3.1 Setup

In each replication of the kernel simulation study, we obtain $T + 1$ random draws x_t from the N-variate normal distribution with mean zero and variance the identity matrix. The prediction function $f(x)$ is then defined using the kernel expansion given below equation (5.1), using random draws $\alpha_t \sim \mathcal{N}(0, 1)$ for the expansion coefficients. An additional set of $T + 1$ random draws ε_t from the univariate normal distribution is generated, with mean zero and a variance selected to control the signal-to-noise ratio, and $y_t = f(x_t) + \varepsilon_t$ is computed for $t = 1, 2, \ldots, T+1$. Finally, the y_t are rescaled to have mean zero and unit variance. Kernel ridge regression is then used to forecast y_{T+1}, given x_{T+1} and the pairs (y_t, x_t) for $t = 1, 2, \ldots, T$.

We simulate univariate ($N = 1$), intermediate ($N = 10$), and data-rich ($N = 100$) models, fixing the number of observations at $T = 100$. The kernels that we consider are the polynomial kernels (5.5) of degrees 1, 2, and 3, the Gaussian kernel (5.9), and the Sinc kernel (5.11). The signal-to-noise ratio ψ is varied over $\{0.5, 1, 2\}$, and the smoothness parameter σ is varied over the middle three values in the grids in Section 5.2.4.

Each kernel is used for forecasting in each data-generating process, to allow us to assess the impact on forecast accuracy of selecting an incorrect kernel. The tuning parameter σ is selected from the grids that we defined in Section 5.2.4. As the correct value of λ is known in this simulation study, we do not estimate it as described in Section 5.2.4. Instead, we select it from the grid $\left\{\frac{1}{4}\lambda_0, \frac{1}{2}\lambda_0, \lambda_0, 2\lambda_0, 4\lambda_0\right\}$, where λ_0 is the true value. This procedure allows us to determine whether such a grid, which is of the same form as the grid we proposed for situations in which λ_0 is unknown, is sufficiently fine.

In the second simulation study, we consider the univariate model $y_t = (1 + \exp(-10x_t))^{-1} + \varepsilon_t$. We shall refer to this experiment as the logistic simulation study. The factor 10 in the exponent is present to make the data-generating process sufficiently nonlinear, see also Figure 5.1. Note that in this case, the true model differs substantially from the prediction functions associated with each of the kernels. As $|x|$ grows large, a prediction function estimated using a polynomial kernel has $|f(x)| \to \infty$, while the Gaussian and Sinc kernels both have $f(x) \to 0$. In contrast, the logistic function approaches different (but finite) values: $f(x) \to 0$ as $x \to -\infty$ and $f(x) \to 1$ as $x \to \infty$.

As in the kernel simulation study, we vary the signal-to-noise ratio ψ over $\{0.5, 1, 2\}$, and we set $T = 100$. Forecasts are obtained using the same five kernels as above.

Table 5.1: Average relative mean squared prediction errors in the kernel simulation study.

data-generating process		kernel, λ, σ	select λ and σ using CV					CV for
kernel	N	correct	Poly(1)	Poly(2)	Poly(3)	Gauss	Sinc	kernel, λ, σ
Poly(1)	1	0.978	**0.981**	0.987	1.002	1.011	1.012	1.007
	10	1.128	**1.109**	1.119	1.119	1.220	1.316	1.128
	100	1.042	**1.019**	1.018	1.015	2.014	2.014	1.020
Poly(2)	1	0.985	1.089	**0.989**	1.003	1.033	1.040	1.016
	10	1.165	1.145	**1.147**	1.149	1.227	1.315	1.160
	100	1.021	1.025	**1.014**	1.010	1.841	1.841	1.024
Poly(3)	1	0.991	1.073	1.002	**1.004**	1.039	1.046	1.019
	10	1.147	1.134	1.140	**1.137**	1.216	1.304	1.152
	100	1.009	1.013	1.006	**1.002**	1.856	1.856	1.014
Gaussian	1	1.014	1.403	1.321	1.271	**1.029**	1.036	1.038
	10	1.065	1.144	1.185	1.188	**1.116**	1.127	1.143
	100	0.952	0.990	1.002	0.994	**0.952**	0.952	0.991
Sinc	1	1.012	1.570	1.484	1.441	1.040	**1.036**	1.045
	10	1.053	1.117	1.144	1.143	1.101	**1.098**	1.122
	100	0.952	0.990	1.002	0.994	0.952	**0.952**	0.991

Notes: This table reports mean squared prediction errors (MSPEs) over 2500 replications of the kernel simulation study, relative to the expected value of the MSPE if the DGP is known, which is $1/(\psi+1)$. The MSPEs in this table were obtained by averaging over all values of the DGP parameters ψ and σ. In the group of columns headed "select λ and σ using CV", MSPEs obtained using the correct kernel are printed in boldface.

5.3.2 Results

Mean squared prediction errors (MSPEs) over 2500 replications of the kernel simulation study are shown in Tables 5.A.1-5.A.3 in Appendix 5.A, and a summary of these results is reported in Table 5.1. For ease of comparison, we have divided all MSPEs by $1/(\psi+1)$, the expected MSPE that would result if the data-generating process were known and used. The summarized results in Table 5.1 were obtained by averaging the relative MSPEs over all DGPs with the same kernel and number of predictors; the differences in results across different values of the parameters ψ and σ are minor.

The column headed "kernel, λ, σ correct" lists the MSPEs that are obtained if kernel ridge regression is used with the same kernel and tuning parameter σ as in the DGP, and with the value of λ corresponding to the true signal-to-noise ratio. As we would expect by our normalization, most numbers in this column are close to unity.

We now shift our attention to the MSPEs resulting from using the correct kernel, but selecting λ and σ using cross-validation, which are indicated in boldface in Table 5.1. Interestingly, these numbers are not much different from those obtained when fixing λ and σ at their correct values; we find that not knowing the correct values of these parameters leads to an increase in MSPE of only around 0.5%. Recall that the values of λ and σ are selected from a grid that allows each of them to be off by a factor of four. Thus, while very extreme values of the tuning parameters might lead to poor forecasts, our relatively crude rule of thumb for selecting their

values seems sufficient. Inspecting the selected values, we find that λ is generally selected correctly, whereas for σ a larger value than that in the data-generating process is often selected in all kernels. This suggests that kernel ridge regression is somewhat biased toward smoother prediction functions, although the effect of this bias on forecast accuracy is minor.

Next, we investigate what happens if we use an incorrect kernel. The results from this procedure can be found in the group of columns headed "select λ and σ using CV" (where CV stands for cross-validation), excluding the numbers printed in boldface. Four features clearly emerge from these results. First, we observe that if the data-generating process is polynomial, using a polynomial kernel of too high degree hardly hurts the forecasting performance. Apparently, the ridge term is an effective safeguard against overfitting in this case. Using a polynomial kernel of too low degree does deteriorate the quality of the forecasts, as expected. Second, the "smooth" Gaussian and Sinc kernels perform almost interchangeably, despite the very different appearance of their kernel functions (5.9) and (5.11). Third, there is an important difference between polynomial and non-polynomial kernels. Using a kernel from one group when the data is generated from a process in the other group almost invariably leads to large forecast errors. Fourth, we observe from the full tables in Appendix 5.A that the differences between kernels are mitigated if the true value of σ goes up. Notice that for all types of kernels under consideration, a higher value of σ translates into a smoother prediction function. The smoother a function is, the less the estimation method matters.

In the last column of Table 5.1 we show the results from selecting not only the tuning parameters, but also the kernel function using cross-validation. We find that in about half of the cases, the cross-validation procedure selects the correct kernel. Moreover, incorrectly selected kernels usually fall in the correct group of polynomials or non-polynomials. As a result, the MSPEs are on average less than 2% larger than when use of the correct kernel is imposed. The selection frequency of the correct kernel is lower for larger values of σ; again, the smoothest functions are easily estimable using any method.

Most of these observations apply to the results with one, ten, or one hundred predictors alike. The main exception is that the difference in results between using polynomial and non-polynomial kernels increases with the number of predictors, N. Related to this finding, we observe that using cross-validation to select the kernel also performs somewhat worse for larger N, since occasionally selecting an incorrect kernel makes a larger difference in that case. For this reason, it seems desirable to consider the correct group of kernels only in cross-validation, that is, to select only among polynomial or non-polynomial kernels. Of course, in practice one does not observe the data-generating process. However, given the more flexible and smoother functional forms provided by the Gaussian and Sinc kernels, we argue that a practitioner is in general better off using only this set of kernels, unless he would have strong prior knowledge that the true predictive relation is polynomial.

We now turn to the results of the logistic simulation study, in which kernel ridge regression always estimates an incorrectly specified model. The relative MSPEs, again over 2500 replications, are reported in Table 5.2. It is clear from theses results that the Gaussian kernel performs best in approximating the logistic function, with the Sinc kernel ranking second best. For the smallest signal-to-noise ratio that we consider ($\psi = 0.5$), the differences between the kernels are minor. As ψ increases, however, the polynomial kernels perform much worse than the non-polynomial ones. That is, if the DGP is reflected by the data more clearly, it becomes more apparent that a polynomial prediction function is not a suitable approximation.

Table 5.2: Relative mean squared prediction errors in the logistic simulation study.

signal-to-noise	select λ and σ using CV					CV for
ratio (ψ)	Poly(1)	Poly(2)	Poly(3)	Gauss	Sinc	kernel, λ, σ
0.5	1.040	1.046	1.051	1.031	1.035	1.037
1.0	1.097	1.106	1.094	1.057	1.072	1.073
2.0	1.221	1.228	1.173	1.083	1.101	1.091

Notes: This table reports mean squared prediction errors (MSPEs) over 2500 replications of the logistic simulation study, relative to the expected value of the MSPE if the DGP is known, which is $1/(\psi + 1)$.

Selecting the kernel using cross-validation leads to a forecast accuracy that ranks in between the polynomial and non-polynomial kernels. Cross-validation selects the Gaussian kernel in 59% and the Sinc kernel in 32% of the replications; however, the remaining 9% in which polynomial kernels are selected still brings the forecast accuracy down substantially. This result illustrates our recommendation that in general, it is not advisable to use the popular polynomial kernels.

As an illustrative example, we show a scatter plot of one simulated data set in Figure 5.1. The true prediction function f is also shown, as well as its estimates using the third-degree polynomial, Gaussian, and Sinc kernels. This figure shows that in contrast with the non-polynomial estimates, the polynomial prediction function is not sufficiently flexible to capture the behavior of the true f. This is particularly evident near $x_t = 0$, where most data points are located.

5.4 Conclusion

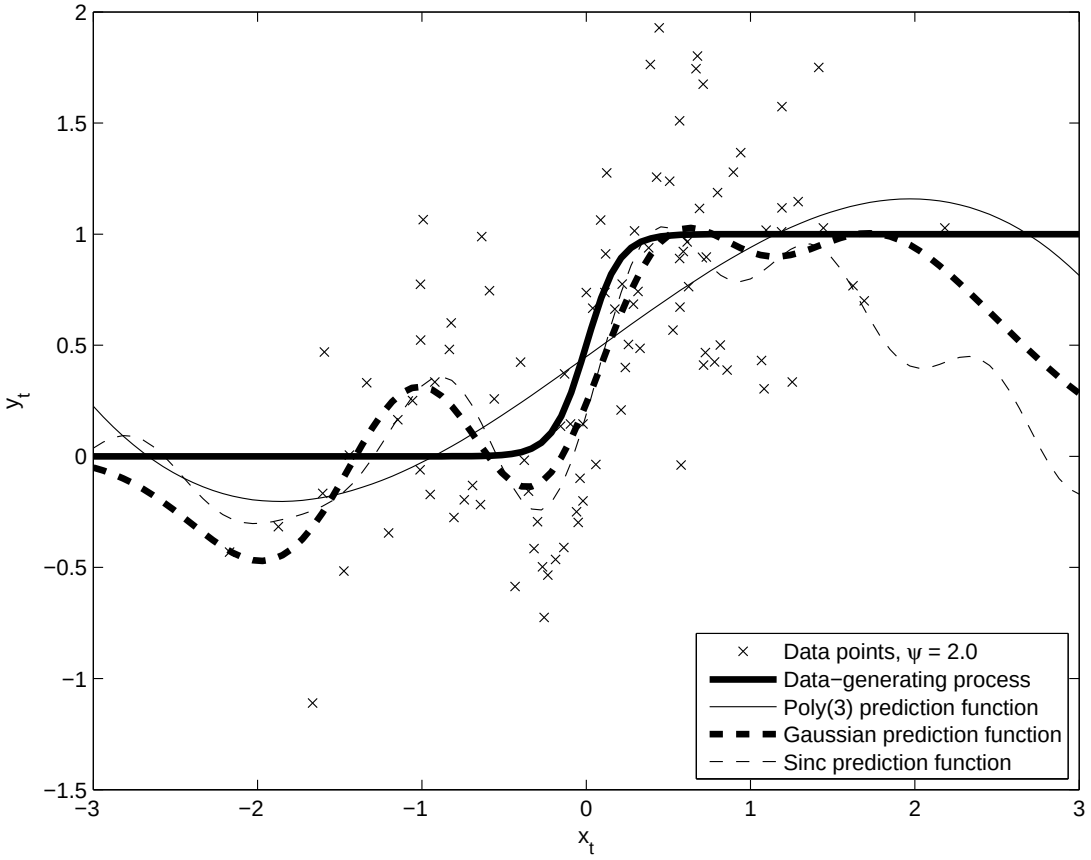
We review the technique of kernel ridge regression from two different points of view, namely from a function approximation perspective and from a Bayesian statistical point of view. This combination of perspectives enables us to give a clear interpretation to two tuning parameters that are generally present in kernel ridge regression. We relate one of these parameters to the signal-to-noise ratio, and the other to the overall smoothness of the regression function. Moreover, we provide rules of thumb for selecting their values.

In addition to the well-known polynomial and Gaussian kernels, we discuss the Sinc kernel. Kernel ridge regression using this kernel function acts as a low-pass filter, so that any high-frequency patterns observed in the data are considered noise and are discarded. Despite this attractive feature, the Sinc kernel has not received widespread attention in the kernel literature.

Our simulation studies confirm the empirical usefulness of our parameter selection rules. Compared to using the true values of the tuning parameters, selecting their values using our rules of thumb leads to an increase of mean squared prediction errors of only 0.5%.

Cross-validation can also be used relatively safely to distinguish among different kernel functions, with a 2% increase in mean squared prediction errors when compared to using the correct kernel. This method performs less favorably for a larger number of predictor variables. We argue that this problem is mainly due to the large difference between non-polynomial and polynomial kernels. For this reason, and because of their smoother and more flexible prediction functions, we recommend to use only non-polynomial kernels if no prior knowledge of the true prediction function is available.

Figure 5.1: The logistic data-generating process with 100 data points, generated with signal-to-noise ratio $\psi = 2.0$. Three estimated prediction functions, using the third-degree polynomial, Gaussian, and Sinc kernels, are also shown.



5.A Detailed simulation results

On the next three pages, we report the mean squared prediction errors for all data-generating processes in the kernel simulation study. A summary of the results was presented in Table 5.1.

Table 5.A.1: Relative mean squared prediction errors in the kernel simulation study, $T = 100, N = 1$.

data-generating process			kernel, λ, σ	select λ and σ using CV					CV for
kernel	σ	ψ	correct	Poly(1)	Poly(2)	Poly(3)	Gauss	Sinc	kernel, λ, σ
Poly(1)	0.71	0.5	0.978	**0.980**	0.991	1.004	1.010	1.006	1.009
		1.0	0.978	**0.980**	0.991	1.008	1.014	1.012	1.008
		2.0	0.978	**0.980**	0.989	1.010	1.028	1.025	1.013
	1.41	0.5	0.978	**0.983**	0.988	1.001	1.004	1.006	1.007
		1.0	0.978	**0.981**	0.986	1.000	1.010	1.016	1.009
		2.0	0.978	**0.980**	0.986	1.000	1.018	1.022	1.011
	2.83	0.5	0.977	**0.981**	0.982	0.997	0.998	1.002	1.000
		1.0	0.978	**0.981**	0.983	0.997	1.005	1.006	0.999
		2.0	0.978	**0.981**	0.984	0.997	1.009	1.014	1.003
Poly(2)	1.22	0.5	0.991	1.108	**0.997**	1.007	1.053	1.054	1.031
		1.0	0.992	1.234	**0.998**	1.011	1.085	1.102	1.023
		2.0	0.992	1.486	**0.995**	1.014	1.123	1.154	1.046
	2.45	0.5	0.984	0.995	**0.989**	1.001	1.005	1.008	1.008
		1.0	0.987	1.007	**0.990**	1.003	1.009	1.018	1.015
		2.0	0.989	1.029	**0.988**	1.003	1.017	1.018	1.024
	4.90	0.5	0.976	0.979	**0.982**	0.994	0.995	1.000	1.001
		1.0	0.978	0.980	**0.982**	0.996	1.003	1.002	0.998
		2.0	0.979	0.981	**0.983**	0.998	1.008	1.008	1.002
Poly(3)	1.73	0.5	1.002	1.090	1.014	**1.010**	1.064	1.071	1.043
		1.0	1.009	1.199	1.033	**1.016**	1.100	1.110	1.031
		2.0	1.014	1.414	1.064	**1.018**	1.158	1.178	1.067
	3.46	0.5	0.983	0.994	0.987	**1.001**	1.003	1.008	1.006
		1.0	0.987	1.002	0.988	**1.003**	1.008	1.019	1.010
		2.0	0.991	1.020	0.989	**1.003**	1.016	1.019	1.020
	6.93	0.5	0.976	0.977	0.981	**0.994**	0.993	0.998	1.000
		1.0	0.977	0.978	0.981	**0.994**	1.002	1.002	0.996
		2.0	0.978	0.979	0.981	**0.995**	1.008	1.007	1.001
Gaussian	0.32	0.5	1.020	1.279	1.252	1.235	**1.034**	1.045	1.041
		1.0	1.037	1.571	1.507	1.465	**1.047**	1.058	1.057
		2.0	1.056	2.155	2.015	1.914	**1.064**	1.084	1.073
	0.64	0.5	1.001	1.171	1.134	1.110	**1.020**	1.022	1.024
		1.0	1.010	1.349	1.264	1.203	**1.025**	1.031	1.035
		2.0	1.020	1.705	1.522	1.383	**1.037**	1.042	1.047
	1.27	0.5	0.988	1.052	1.028	1.022	**1.006**	1.006	1.015
		1.0	0.994	1.112	1.057	1.044	**1.012**	1.015	1.023
		2.0	1.000	1.231	1.112	1.064	**1.013**	1.018	1.027
Sinc	0.16	0.5	1.020	1.340	1.325	1.330	1.044	**1.044**	1.047
		1.0	1.034	1.710	1.673	1.656	1.064	**1.059**	1.064
		2.0	1.049	2.449	2.366	2.323	1.086	**1.070**	1.085
	0.32	0.5	1.001	1.269	1.235	1.224	1.033	**1.026**	1.034
		1.0	1.009	1.551	1.471	1.433	1.042	**1.032**	1.045
		2.0	1.016	2.112	1.952	1.830	1.052	**1.047**	1.057
	0.64	0.5	0.988	1.096	1.049	1.038	1.011	**1.009**	1.017
		1.0	0.993	1.198	1.096	1.053	1.015	**1.016**	1.025
		2.0	0.997	1.400	1.187	1.082	1.012	**1.022**	1.028

Notes: This table reports mean squared prediction errors (MSPEs) over 2500 replications of the kernel simulation study with $N = 1$ predictor, relative to the expected value of the MSPE if the DGP is known, which is $1/(\psi + 1)$. In the group of columns headed "select λ and σ using CV", MSPEs obtained using the correct kernel are printed in boldface.

Table 5.A.2: Relative mean squared prediction errors in the kernel simulation study, $T = 100, N = 10$.

data-generating process			kernel, λ, σ	select λ and σ using CV					CV for
kernel	σ	ψ	correct	Poly(1)	Poly(2)	Poly(3)	Gauss	Sinc	kernel, λ, σ
Poly(1)	2.24	0.5	1.138	**1.131**	1.142	1.144	1.176	1.215	1.156
		1.0	1.157	**1.149**	1.163	1.161	1.242	1.319	1.169
		2.0	1.169	**1.161**	1.175	1.181	1.353	1.500	1.173
	4.47	0.5	1.111	**1.092**	1.103	1.102	1.146	1.195	1.116
		1.0	1.134	**1.103**	1.110	1.111	1.201	1.286	1.124
		2.0	1.154	**1.114**	1.124	1.124	1.285	1.447	1.130
	8.94	0.5	1.076	**1.070**	1.079	1.080	1.135	1.183	1.089
		1.0	1.094	**1.076**	1.084	1.086	1.182	1.271	1.096
		2.0	1.117	**1.082**	1.090	1.087	1.259	1.428	1.097
Poly(2)	2.45	0.5	1.158	1.165	**1.174**	1.175	1.183	1.211	1.195
		1.0	1.222	1.219	**1.214**	1.214	1.248	1.308	1.228
		2.0	1.328	1.313	**1.269**	1.275	1.353	1.475	1.282
	4.90	0.5	1.120	1.105	**1.115**	1.114	1.153	1.198	1.126
		1.0	1.146	1.118	**1.128**	1.133	1.213	1.294	1.140
		2.0	1.180	1.137	**1.146**	1.149	1.303	1.458	1.157
	9.80	0.5	1.087	1.077	**1.087**	1.087	1.138	1.185	1.097
		1.0	1.108	1.083	**1.092**	1.096	1.183	1.274	1.110
		2.0	1.131	1.091	**1.100**	1.100	1.266	1.433	1.109
Poly(3)	3.46	0.5	1.141	1.153	1.163	**1.159**	1.166	1.194	1.184
		1.0	1.190	1.201	1.207	**1.201**	1.225	1.283	1.223
		2.0	1.268	1.282	1.263	**1.253**	1.320	1.442	1.269
	6.93	0.5	1.113	1.096	1.111	**1.107**	1.153	1.195	1.121
		1.0	1.139	1.109	1.120	**1.120**	1.205	1.285	1.133
		2.0	1.170	1.127	1.134	**1.136**	1.292	1.451	1.141
	13.86	0.5	1.080	1.072	1.082	**1.082**	1.135	1.183	1.093
		1.0	1.100	1.079	1.088	**1.089**	1.183	1.272	1.105
		2.0	1.123	1.086	1.094	**1.090**	1.262	1.429	1.102
Gaussian	0.32	0.5	1.045	1.077	1.092	1.092	**1.072**	1.066	1.093
		1.0	1.045	1.091	1.113	1.110	**1.085**	1.080	1.111
		2.0	1.045	1.120	1.150	1.149	**1.105**	1.111	1.139
	0.64	0.5	1.046	1.078	1.094	1.093	**1.074**	1.066	1.093
		1.0	1.047	1.094	1.114	1.114	**1.086**	1.082	1.112
		2.0	1.047	1.123	1.155	1.153	**1.109**	1.113	1.140
	1.27	0.5	1.091	1.148	1.178	1.183	**1.128**	1.133	1.152
		1.0	1.105	1.218	1.288	1.289	**1.164**	1.191	1.195
		2.0	1.118	1.351	1.485	1.509	**1.218**	1.301	1.252
Sinc	0.16	0.5	1.045	1.077	1.092	1.091	1.072	**1.066**	1.093
		1.0	1.045	1.091	1.113	1.110	1.085	**1.080**	1.111
		2.0	1.045	1.120	1.150	1.149	1.105	**1.111**	1.139
	0.32	0.5	1.045	1.077	1.093	1.092	1.073	**1.066**	1.093
		1.0	1.046	1.091	1.114	1.110	1.086	**1.080**	1.111
		2.0	1.046	1.121	1.151	1.150	1.106	**1.111**	1.140
	0.64	0.5	1.062	1.106	1.132	1.128	1.097	**1.093**	1.109
		1.0	1.068	1.146	1.184	1.184	1.124	**1.118**	1.132
		2.0	1.074	1.220	1.270	1.273	1.160	**1.156**	1.166

Notes: This table reports relative MSPEs over 2500 replications of the kernel simulation study with $N = 10$ predictors.

Table 5.A.3: Relative mean squared prediction errors in the kernel simulation study, $T = 100$, $N = 100$.

data-generating process			kernel, λ, σ	select λ and σ using CV					CV for
kernel	σ	ψ	correct	Poly(1)	Poly(2)	Poly(3)	Gauss	Sinc	kernel, λ, σ
Poly(1)	7.07	0.5	1.033	**1.026**	1.024	1.017	1.361	1.361	1.026
		1.0	1.099	**1.063**	1.056	1.053	1.781	1.781	1.062
		2.0	1.214	**1.119**	1.107	1.105	2.624	2.624	1.114
	14.14	0.5	0.990	**0.993**	0.995	0.991	1.412	1.412	0.995
		1.0	1.023	**1.005**	1.007	1.005	1.883	1.883	1.007
		2.0	1.086	**1.022**	1.024	1.023	2.830	2.831	1.023
	28.28	0.5	0.966	**0.977**	0.977	0.977	1.426	1.426	0.979
		1.0	0.975	**0.981**	0.982	0.981	1.914	1.914	0.983
		2.0	0.994	**0.987**	0.988	0.988	2.895	2.895	0.989
Poly(2)	7.14	0.5	0.996	1.014	**1.005**	0.999	1.179	1.179	1.013
		1.0	1.019	1.050	**1.028**	1.020	1.417	1.417	1.048
		2.0	1.049	1.120	**1.060**	1.050	1.895	1.895	1.111
	14.28	0.5	1.000	0.999	**1.001**	0.997	1.378	1.378	1.002
		1.0	1.037	1.019	**1.016**	1.013	1.815	1.815	1.019
		2.0	1.097	1.048	**1.044**	1.039	2.693	2.693	1.046
	28.57	0.5	0.974	0.982	**0.984**	0.983	1.420	1.420	0.985
		1.0	0.991	0.990	**0.991**	0.991	1.901	1.902	0.993
		2.0	1.027	1.000	**1.001**	1.000	2.869	2.869	1.003
Poly(3)	10.10	0.5	0.987	1.001	0.994	**0.994**	1.184	1.184	1.003
		1.0	1.005	1.033	1.017	**1.008**	1.427	1.427	1.031
		2.0	1.027	1.093	1.047	**1.033**	1.915	1.915	1.086
	20.20	0.5	0.993	0.995	0.997	**0.993**	1.389	1.389	0.998
		1.0	1.025	1.009	1.009	**1.006**	1.837	1.837	1.009
		2.0	1.078	1.032	1.029	**1.029**	2.737	2.738	1.031
	40.40	0.5	0.970	0.979	0.980	**0.979**	1.423	1.423	0.981
		1.0	0.983	0.986	0.989	**0.986**	1.907	1.907	0.988
		2.0	1.011	0.994	0.995	**0.994**	2.881	2.881	0.996
Gaussian	0.32	0.5	0.952	0.974	0.980	0.976	**0.952**	0.952	0.972
		1.0	0.952	0.984	0.999	0.991	**0.952**	0.952	0.986
		2.0	0.952	1.012	1.026	1.016	**0.952**	0.952	1.013
	0.64	0.5	0.952	0.974	0.980	0.976	**0.952**	0.952	0.972
		1.0	0.952	0.984	0.999	0.991	**0.952**	0.952	0.986
		2.0	0.952	1.012	1.026	1.016	**0.952**	0.952	1.013
	1.27	0.5	0.952	0.974	0.980	0.976	**0.952**	0.952	0.972
		1.0	0.952	0.984	0.999	0.991	**0.952**	0.952	0.986
		2.0	0.952	1.012	1.026	1.016	**0.952**	0.952	1.013
Sinc	0.16	0.5	0.952	0.974	0.980	0.976	0.952	**0.952**	0.972
		1.0	0.952	0.984	0.999	0.991	0.952	**0.952**	0.986
		2.0	0.952	1.012	1.026	1.016	0.952	**0.952**	1.013
	0.32	0.5	0.952	0.974	0.980	0.976	0.952	**0.952**	0.972
		1.0	0.952	0.984	0.999	0.991	0.952	**0.952**	0.986
		2.0	0.952	1.012	1.026	1.016	0.952	**0.952**	1.013
	0.64	0.5	0.952	0.974	0.980	0.976	0.952	**0.952**	0.972
		1.0	0.952	0.984	0.999	0.991	0.952	**0.952**	0.986
		2.0	0.952	1.012	1.026	1.016	0.952	**0.952**	1.013

Notes: This table reports relative MSPEs over 2500 replications of the kernel simulation study with $N = 100$ predictors.

Nederlandse Samenvatting (Summary in Dutch)

Dit proefschrift omvat vier afzonderlijk te lezen hoofdstukken over vernieuwende technieken voor het maken van economische voorspellingen. Elk hoofdstuk behandelt methoden die efficiënt omgaan met de informatie in grote hoeveelheden gegevens. Elk van deze methoden wordt vergeleken met de technieken die momenteel gebruikelijk zijn en in het algemeen vinden we verbeteringen van de voorspelkwaliteit. De geïntroduceerde methoden worden toegepast voor het voorspellen van de rente op Amerikaanse staatsobligaties in hoofdstuk 2, van huizenprijzen in hoofdstuk 3, en van de reële macro-economische grootheden productie, inkomen, verkopen en werkgelegenheid in de hoofdstukken 3 en 4. Daarnaast bevatten de hoofdstukken 3, 4 en 5 simulatiestudies om het gebruik van deze nieuwe methode in verschillende contexten te onderzoeken.

In de hoofdstukken 2 en 3 behandelen we lineaire modellen, en in de hoofdstukken 4 en 5 niet-lineaire modellen. Wat betreft het gebruik van lineaire methoden voor grote hoeveelheden gegevens, is het veruit het meest gebruikelijk om de klassieke principalecomponentenanalyse toe te passen. Wij wijzen op een aantal tekortkomingen van deze methode en we bevelen aan om geavanceerdere technieken te gebruiken die deze tekortkomingen niet hebben. In hoofdstuk 2 wordt aangetoond dat de recentelijk hernieuwde populariteit van partiële kleinste kwadraten in de econometrie terecht is. Voorts wordt in hoofdstuk 3 een nieuwe factorconstructiemethode voorgesteld. Deze nieuwe methode resulteert in robuuste en interpreteerbare factoren, die ook goed presteren wanneer er voorspellingen op gebaseerd worden.

In hoofdstuk 2 vergelijken we verscheidene methoden om de informatie in een groot macro-economisch gegevensbestand te gebruiken om de rentetermijnstructuur te voorspellen binnen het Nelson-Siegel model. Vijf vraagstukken gerelateerd aan het construeren van factoren uit zo'n grote verzameling gegevens worden behandeld, namelijk het selecteren van een gedeelte van de beschikbare informatie, het gebruik van de voorspeldoelstelling bij het construeren van factoren, het gelijktijdig voorspellen van meerdere variabelen, het groeperen van gegevens alvorens factoren te construeren, en het automatisch selecteren van het aantal factoren. Onze empirische resultaten tonen aan dat elk van deze punten bijdraagt aan een verhoogde voorspelkwaliteit, vooral voor obligaties met een relatief korte of juist lange looptijd. Het gebruik van macro-economische informatie helpt het meest in volatiele perioden, inclusief de crisis in 2008-2009, wanneer eenvoudigere modellen onvoldoende blijken te werken. De best werkende techniek om deze informatie samen te vatten blijkt partiële kleinste kwadraten, gevolgd door principalecomponentenanalyse. De gemiddelde kwadratische voorspelfout wordt 20-30% verminderd, in vergelijking met het Nelson-Siegel model zonder macro-economische factoren.

Hoofdstuk 3 begint met enkele punten van kritiek op bestaande methoden voor factorconstructie, die zeer algemeen gebruikt worden om grote gegevensbestanden samen te vatten in enkele representatieve factoren: ze zijn niet robuust tegen extreme waarnemingen, en ze omschrijven elke factor als een lineaire combinatie van alle variabelen, hetgeen de interpretatie van de factoren bemoeilijkt. Wij stellen een nieuwe procedure voor die deze nadelen niet heeft. We vinden dat deze procedure beter interpreteerbare factoren oplevert dan meer gebruikelijke technieken. Tevens blijken deze factoren goed te presteren wanneer ze worden gebruikt om mee te voorspellen, zowel in een Monte-Carlo-experiment als in een macro-economische en een micro-economische empirische toepassing.

In vergelijking met het gebruik van lineaire methoden staat het gebruik van niet-lineaire methoden in grote gegevensverzamelingen nog in de kinderschoenen. In de hoofdstukken 4 en 5 gebruiken we de zogenoemde "kernel"-methodologie, die afkomstig is uit de wereld van het zogeheten "machinaal leren". Een gebruikelijke toepassing is het geautomatiseerd herkennen van met de hand geschreven tekens, zoals in postcodes. Deze techniek wordt in hoofdstuk 4 uitgebreid om tijdreeksvoorspellingen mogelijk te maken, en het blijkt dat deze "kernel-ridge-regressie" in veel gevallen beter presteert dan traditionele lineaire voorspeltechnieken. Er staan echter nog belangrijke vragen open met betrekking tot modelselectie in kernel-ridge-regressie. Er lijkt tot nu toe weinig bekend te zijn over hoe een kernel gekozen moet worden, laat staan over hoe de bijbehorende parameters gekozen moeten worden. Het theoretische onderzoek in hoofdstuk 5 leidt tot een gemakkelijk toe te passen vuistregel om deze vragen te beantwoorden, en we tonen aan dat deze vuistregel modellen selecteert die goede voorspellingen opleveren.

Meer in het bijzonder behandelt hoofdstuk 4 het gebruik van kernel-ridge-regressie voor het doen van voorspellingen op basis van niet-lineaire modellen met veel predictoren. Deze methode omhelst het niet-lineair transformeren van de voorspelvariabelen naar een hoog-dimensionale ruimte. In deze ruimte wordt een voorspelmodel geschat, gebruikmakend van een "shrinkage"-techniek om ondanks de hoge dimensie toch een accurate schatting te kunnen maken. We breiden deze techniek uit om haar toe te kunnen passen op het voorspellen van economische tijdreeksen, door het mogelijk te maken dat het effect van vertraagde waarnemingen van de te voorspellen variabele of andere individuele variabelen apart behandeld kan worden. Middels Monte-Carlo-simulaties en een empirische toepassing op verscheidene maatstaven van de reële economische activiteit wordt bevestigd dat kernel-ridge-regressie preciezere voorspellingen kan leveren dan traditionele lineaire, op principale componenten gebaseerde methoden.

In hoofdstuk 5 onderzoeken we modelleringsvraagstukken met betrekking tot kernel-ridge-regressie. Meer specifiek bestuderen we de invloed van het kiezen van de kernel en van zijn parameters op de voorspelkwaliteit. Een aantal populaire kernels wordt behandeld, te weten polynomiale kernels, de Gaussische kernel en de Sinc-kernel. We geven een interpretatie aan deze laatste twee kernels door de "gladheid" van de bijbehorende voorspelfunctie te bestuderen. Vervolgens relateren we de parameters van al deze kernels aan maatstaven van gladheid en aan de signaal/ruis-verhouding. Op basis van deze resultaten worden aanbevelingen gedaan voor het selecteren van de parameters middels kruisvalidatie over kleine rasters. Een Monte-Carlo-studie toont aan dat deze vuistregels in de praktijk goed werken. Ten slotte wordt duidelijk gemaakt dat de flexibele en gladde voorspelfuncties die bij de Gaussische kernel en de Sinc-kernel behoren deze kernels breed toepasbaar maken. We bevelen daarom aan dat in het algemeen, wanneer niets bekend is over het data-genererende proces, deze kernels zouden moeten worden gebruikt in plaats van de thans meer gangbare polynomiale kernels.

Bibliography

M. Aiolfi and C.A. Favero. Model uncertainty, thick modelling and the predictability of stock returns. *Journal of Forecasting*, 24:233–254, 2005.

A. Ang and M. Piazzesi. A no-arbitrage vector autoregression of term structure dynamics with macroeconomic and latent variables. *Journal of Monetary Economics*, 50:745–787, 2003.

N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68:337–404, 1950.

J. Bai and S. Ng. Forecasting economic time series using targeted predictors. *Journal of Econometrics*, 146:304–317, 2008.

M. Bańbura, D. Giannone, and L. Reichlin. Large Bayesian vector autoregressions. *Journal of Applied Econometrics*, 25:71–92, 2010.

R.R. Bliss. Testing term structure estimation methods. *Advances in Futures and Options Research*, 9:197–231, 1997.

S. Bochner. Monotone Funktionen, Stieltjessche Integrale und harmonische Analyse. *Mathematische Annalen*, 180:378–410, 1933.

J. Boivin and S. Ng. Are more data always better for factor analysis? *Journal of Econometrics*, 132:169–194, 2006.

B.E. Boser, I.M. Guyon, and V.N. Vapnik. A training algorithm for optimal margin classifiers. In D. Haussler, editor, *Proceedings of the Annual Conference on Computational Learning Theory*, pages 144–152. ACM Press, Pittsburgh, Pennsylvania, 1992.

D.S. Broomhead and D. Lowe. Multivariable functional interpolation and adaptive networks. *Complex Systems*, 2:321–355, 1988.

A. Carriero, G. Kapetanios, and M. Marcellino. Forecasting large datasets with Bayesian reduced rank multivariate models. *Journal of Applied Econometrics*, 26:735–761, 2011.

G.C. Cawley and N.L.C. Talbot. Efficient approximate leave-one-out cross-validation for kernel logistic regression. *Machine Learning*, 71:243–264, 2008.

C. Çakmaklı and D. van Dijk. Getting the most out of macroeconomic information for predicting stock returns and volatility. *Tinbergen Institute Discussion Paper 2010-115/4*, 2010.

J.H.E. Christensen, F.X. Diebold, and G.D. Rudebusch. The affine arbitrage-free class of Nelson-Siegel term structure models. *Journal of Econometrics*, 164:4–20, 2011.

J.C. Cox, J.E. Ingersoll, and S.A. Ross. A theory of the term structure of interest rates. *Econometrica*, 53:385–407, 1985.

C. Croux and P. Exterkate. Sparse and robust factor modelling. *Tinbergen Institute Discussion Paper 2011-122/4*, 2011.

C. Croux and G. Haesbroeck. Principal component analysis based on robust estimators of the covariance or correlation matrix: Influence functions and efficiencies. *Biometrika*, 87:603–618, 2000.

C. Croux, P. Filzmoser, G. Pison, and P.J. Rousseeuw. Fitting multiplicative models by robust alternating regressions. *Statistics and Computing*, 13:23–36, 2003.

Q. Dai and K.J. Singleton. Specification analysis of affine term structure models. *Journal of Finance*, 55:1943–1978, 2000.

F. De la Torre and M.J. Black. Robust principal component analysis for computer vision. In *International Conference on Computer Vision*, pages 362–369, Vancouver, Canada, 2001.

C. De Mol, D. Giannone, and L. Reichlin. Forecasting using a large number of predictors: Is Bayesian shrinkage a valid alternative to principal components? *Journal of Econometrics*, 146:318–328, 2008.

M. De Pooter, F. Ravazzolo, and D. van Dijk. Predicting the term structure of interest rates: Incorporating parameter uncertainty, model uncertainty and macroeonomic information. *Tinbergen Institute Discussion Paper 2007-028/4*, 2007.

C. Dehon, M. Gassner, and V. Verardi. A Hausman-type test to detect the presence of influential outliers in regression analysis. *Economics Letters*, 105:64–67, 2009.

H. Dewachter and M. Lyrio. Macro factors and the term structure of interest rates. *Journal of Money, Credit, and Banking*, 38:119–140, 2006.

F.X. Diebold and C. Li. Forecasting the term structure of government bond yields. *Journal of Econometrics*, 130:337–364, 2006.

F.X. Diebold and R.S. Mariano. Comparing predictive accuracy. *Journal of Business and Economic Statistics*, 13:134–144, 1995.

F.X. Diebold, G.D. Rudebusch, and S.B. Aruoba. The macroeconomy and the yield curve: A dynamic latent factor approach. *Journal of Econometrics*, 131:309–338, 2006.

G.R. Duffee. Term premia and interest rate forecasts in affine models. *Journal of Finance*, 57: 405–443, 2002.

G.R. Duffee. Forecasting with the term structure: The role of no-arbitrage restrictions. *Johns Hopkins University working paper*, 2011.

D. Duffie and R. Kan. A yield-factor model of interest rates. *Mathematical Finance*, 6:379–406, 1996.

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Annals of Statistics*, 32:407–451, 2004.

P. Exterkate. Modelling issues in kernel ridge regression. *Tinbergen Institute Discussion Paper 2011-138/4*, 2011.

P. Exterkate, P.J.F. Groenen, C. Heij, and D. van Dijk. Nonlinear forecasting with many predictors using kernel ridge regression. *Tinbergen Institute Discussion Paper 2011-007/4*, 2011a.

P. Exterkate, D. van Dijk, C. Heij, and P.J.F. Groenen. Forecasting the yield curve in a data-rich environment using the Factor-Augmented Nelson-Siegel model. *Journal of Forecasting*, forthcoming, doi:10.1002/for.1258, 2011b.

G. Fagiolo, M. Napoletano, and A. Roventini. Are output growth-rate distributions fat-tailed? Some evidence from OECD countries. *Journal of Applied Econometrics*, 23:639–669, 2008.

E.F. Fama and R.R. Bliss. The information in long-maturity forward rates. *American Economic Review*, 77:680–692, 1987.

J. Faust and J.H. Wright. Comparing Greenbook and reduced form forecasts using a large realtime dataset. *Journal of Business and Economic Statistics*, 27:468–479, 2009.

C.A. Favero, L. Niu, and L. Sala. Term structure forecasting: No-arbitrage restrictions versus large information set. *Journal of Forecasting*, forthcoming, doi:10.1002/for.1181, 2011.

M. Forni, M. Hallin, M. Lippi, and L. Reichlin. The generalized dynamic factor model: Identification and estimation. *Review of Economics and Statistics*, 82:540–554, 2000.

M. Forni, M. Hallin, M. Lippi, and L. Reichlin. The generalized dynamic factor model: One-sided estimation and forecasting. *Journal of the American Statistical Association*, 100:830–840, 2005.

J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33:1–22, 2010.

P.H. Garthwaite. An interpretation of partial least squares. *Journal of the American Statistical Association*, 89:122–127, 1994.

R. Giacomini and H. White. Tests of conditional predictive ability. *Econometrica*, 74:1545–1578, 2006.

J.J.J. Groen and G. Kapetanios. Revisiting useful approaches to data-rich macroeconomic forecasting. *Federal Reserve Bank of New York Staff Report 327*, 2008.

M. Hallin and R. Liška. Dynamic factors in the presence of blocks. *Journal of Econometrics*, 163:29–41, 2011.

D. Harrison and D.L. Rubinfeld. Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, 5:81–102, 1978.

C. Heij, P.J.F. Groenen, and D. van Dijk. Time series forecasting by principal covariate regression. *Econometric Institute Report 2006-37*, 2006.

C. Heij, P.J.F. Groenen, and D. van Dijk. Forecast comparison of principal component regression and principal covariate regression. *Computational Statistics and Data Analysis*, 51: 3612–3625, 2007.

T. Hofmann, B. Schölkopf, and A.J. Smola. Kernel methods in machine learning. *Annals of Statistics*, 36:1171–1220, 2008.

P. Hördahl, O. Tristani, and D. Vestin. A joint econometric model of macroeconomic and term-structure dynamics. *Journal of Econometrics*, 131:405–444, 2006.

H. Huang and T.-H. Lee. To combine forecasts or to combine information? *Econometric Reviews*, 29:534–570, 2010.

I.T. Jolliffe, N.T. Trendafilov, and M. Uddin. A modified principal component technique based on the Lasso. *Journal of Computational and Graphical Statistics*, 12:531–547, 2003.

B. Jungbacker, S.J. Koopman, and M. van der Wel. Dynamic factor models with smooth loadings for analyzing the term structure of interest rates. *Tinbergen Institute Discussion Paper 2009-041/4*, 2010.

G.S. Kimeldorf and G. Wahba. Some results on Tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications*, 33:82–95, 1971.

A.B. Kock and T. Teräsvirta. Forecasting with non-linear models. In M.P. Clements and D.F. Hendry, editors, *Oxford Handbook of Economic Forecasting*. Oxford University Press, Oxford, 2011.

S.J. Koopman, M.I.P. Mallee, and M. van der Wel. Analyzing the term structure of interest rates using the dynamic Nelson-Siegel model with time-varying parameters. *Journal of Business and Economic Statistics*, 28:329–343, 2010.

S.J. Koopman, D. van Dijk, M. van der Wel, and J.H. Wright. Forecasting interest rates with shifting endpoints. *Working paper, Econometric Institute, Erasmus University Rotterdam*, 2011.

S.C. Ludvigson and S. Ng. The empirical risk-return relation: A factor analysis approach. *Journal of Financial Economics*, 83:171–222, 2007.

S.C. Ludvigson and S. Ng. Macro factors in bond risk premia. *Review of Financial Studies*, 22: 5027–5067, 2009.

R.A. Maronna and V.J. Yohai. Robust low-rank approximation of data matrices with element-wise contamination. *Technometrics*, 50:295–304, 2008.

R.A. Maronna, D.R. Martin, and V.J. Yohai. *Robust Statistics: Theory and Methods*. Wiley, New York, 2006.

M.C. Medeiros, T. Teräsvirta, and G. Rech. Building neural network models for time series: A statistical approach. *Journal of Forecasting*, 25:49–75, 2006.

J. Mercer. Functions of positive and negative type, and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society of London: Series A*, 209:415–446, 1909.

E. Mönch. Forecasting the yield curve in a data-rich environment: A no-arbitrage factor-augmented VAR approach. *Journal of Econometrics*, 146:26–43, 2008.

K.-R. Müller, A. Smola, G. Rätsch, B. Schölkopf, J. Kohlmorgen, and V. Vapnik. Predicting time series with support vector machines. In W. Gerstner, A. Germond, M. Hasler, and J.-D. Nicoud, editors, *Artificial Neural Networks ICANN'97*, pages 999–1004. Springer, Berlin, 1997.

C.R. Nelson and A.F. Siegel. Parsimonious modeling of yield curves. *Journal of Business*, 60: 473–489, 1987.

R.K. Pace and O.W. Gilley. Using the spatial configuration of the data to improve estimation. *Journal of Real Estate Finance and Economics*, 14:333–340, 1997.

A.R. Pagan and A. Ullah. *Nonparametric Econometrics*. Cambridge University Press, Cambridge, United Kingdom, 1999.

G. Pison, P.J. Rousseeuw, P. Filzmoser, and C. Croux. Robust factor analysis. *Journal of Multivariate Analysis*, 84:145–172, 2003.

T. Poggio. On optimal nonlinear associative recall. *Biological Cybernetics*, 19:201–209, 1975.

J. Racine. Consistent cross-validatory model-selection for dependent data: hv-block cross-validation. *Journal of Econometrics*, 99:39–61, 2000.

H. Raiffa and R. Schlaifer. *Applied Statistical Decision Theory*. Harvard University Press, Cambridge, Massachusetts, 1961.

D.E. Rapach, J.K. Strauss, and G. Zhou. Out-of-sample equity premium prediction: Combination forecasts and links to the real economy. *Review of Financial Studies*, 23:821–862, 2010.

S.N. Roy. *Some aspects of multivariate analysis*. Wiley, New York, 1957.

G.D. Rudebusch and T. Wu. A macro-finance model of the term structure, monetary policy, and the economy. *Economic Journal*, 118:906–926, 2008.

B. Schölkopf, A. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10:1299–1319, 1998.

T. Similä and J. Tikka. Common subset selection of inputs in multiresponse regression. In *Proceedings of the IEEE International Joint Conference on Neural Networks*, pages 1908–1915, Vancouver, Canada, 2006.

A.J. Smola and B. Schölkopf. A tutorial on support vector regression. *Statistics and Computing*, 14:199–222, 2004.

A.J. Smola, B. Schölkopf, and K.-R. Müller. The connection between regularizaton operators and support vector kernels. *Neural Networks*, 11:637–649, 1998.

J.H. Stock and M.W. Watson. A comparison of linear and nonlinear univariate models for forecasting macroeconomic time series. In R.F. Engle and H. White, editors, *Cointegration, Causality and Forecasting. A Festschrift in Honour of Clive W.J. Granger*, pages 1–44. Oxford University Press, Oxford, 1999.

J.H. Stock and M.W. Watson. Macroeconomic forecasting using diffusion indexes. *Journal of Business and Economic Statistics*, 20:147–162, 2002.

J.H. Stock and M.W. Watson. Implications of dynamic factor models for VAR analysis. *NBER Working Paper No. 11467*, 2005.

J.H. Stock and M.W. Watson. Forecasting with many predictors. In G. Elliot, C.W.J. Granger, and A. Timmermann, editors, *Handbook of Economic Forecasting*, pages 515–554. Elsevier, Amsterdam, 2006.

J.H. Stock and M.W. Watson. Why has U.S. inflation become harder to forecast? *Journal of Money, Credit and Banking*, 39:3–33, 2007.

J.H. Stock and M.W. Watson. Generalized shrinkage methods for forecasting using many predictors. *Manuscript, Harvard University*, 2009.

J.B. Taylor. Discretion versus policy rules in practice. *Carnegie-Rochester Conference Series on Public Policy*, 39:195–214, 1993.

T. Teräsvirta. Forecasting economic variables with nonlinear models. In G. Elliot, C.W.J. Granger, and A. Timmermann, editors, *Handbook of Economic Forecasting*, pages 413–458. Elsevier, Amsterdam, 2006.

T. Teräsvirta, D. van Dijk, and M.C. Medeiros. Linear models, smooth transition autoregressions, and neural networks for forecasting macroeconomic time series: A re-examination. *International Journal of Forecasting*, 21:755–774, 2005.

R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B*, 58:267–288, 1996.

V.N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, New York, 1995.

O. Vašíček. An equilibrium characterization of the term structure. *Journal of Financial Economics*, 5:177–188, 1977.

H. Wang, G. Li, and G. Jiang. Robust regression shrinkage and consistent variable selection through the LAD-Lasso. *Journal of Business and Economic Statistics*, 25:347–355, 2007.

H. White. Approximate nonlinear forecasting methods. In G. Elliot, C.W.J. Granger, and A. Timmermann, editors, *Handbook of Economic Forecasting*, pages 459–514. Elsevier, Amsterdam, 2006.

D.M. Witten, R. Tibshirani, and T. Hastie. A penalized matrix decomposition, with applications to sparse principal component analysis and canonical correlation analysis. *Biostatistics*, 10: 515–534, 2009.

H. Wold. Nonlinear estimation by iterative least squares procedures. In F. David, editor, *Research Papers in Statistics: Festschrift for J. Neyman*, pages 411–444. Wiley, New York, 1966.

J.H. Wright. Forecasting US inflation by Bayesian model averaging. *Journal of Forecasting*, 28:131–144, 2009.

K. Yao. Applications of reproducing kernel Hilbert spaces: Bandlimited signal models. *Information and Control*, 11:429–444, 1967.

H. Zou, T. Hastie, and R. Tibshirani. On the "degrees of freedom" of the Lasso. *Annals of Statistics*, 35:2173–2192, 2007.

Curriculum Vitae

Education

Sep. 2008 - Sep. 2011: Ph.D. in Econometrics
Tinbergen Institute, Erasmus University Rotterdam
Research topic: Novel techniques for data-rich economic forecasting
Defence expected: December 2011

Sep. 2006 - Aug. 2008: M.Phil. in Economics
Tinbergen Institute, Erasmus University Rotterdam
Major field: Econometrics; minor field: Macroeconomics
Diploma received: August 2008, with honors

Sep. 2003 - Aug. 2006: B.Sc. in Econometrics and Management Science
Econometric Institute, Erasmus University Rotterdam
Major field: Econometrics
Diploma received: October 2006, with honors

Publications

- "Forecasting the yield curve in a data-rich environment using the Factor-Augmented Nelson-Siegel model" (with D. van Dijk, C. Heij, and P.J.F. Groenen): *Journal of Forecasting*, 2011, forthcoming.

Discussion papers

- "Modelling issues in kernel ridge regression": *Tinbergen Institute Discussion Paper 2011-138/4*, 2011.
- "Sparse and robust factor modelling" (with C. Croux): *Tinbergen Institute Discussion Paper 2011-122/4*, 2011.
- "Nonlinear forecasting with many predictors using kernel ridge regression" (with P.J.F. Groenen, C. Heij, and D. van Dijk): *Tinbergen Institute Discussion Paper 2011-007/4*, 2011.

Scholarships

- Full Tinbergen Institute Scholarship for the second year of M.Phil. studies (academic year 2007-2008), as a reward for excellent performance (#1 of cohort) in 2006-2007.

Professional experience

Oct. 2011 - Present:	Post-doctoral Researcher at CREATES, Aarhus University Research topic: Forecasting methods for financial variables in a data-rich environment
Sep. 2008 - Sep. 2011:	Ph.D. Candidate at Tinbergen Institute, Erasmus University Rotterdam Research topic: Novel techniques for data-rich economic forecasting
Feb. 2011 - May 2011:	Visiting Ph.D. Candidate at Katholieke Universiteit Leuven Research topic: Robust and sparse factor modelling

Teaching experience

2008-'09, 2009-'10:	Programming in Matlab B.Sc. programme, with Prof. Patrick Groenen
2008-'09, 2009-'10:	Multivariate and Nonparametric Statistics B.Sc. programme, with Prof. Patrick Groenen and Dr. Alex Koning
2007-'08, 2009-'10:	Case Studies in Applied Econometrics M.Sc. programme, with Dr. Christiaan Heij
Sep. 2009 - Oct. 2009:	Completed the Basic Didactics Course at Risbo Research and Training Institute. Certificate received in November 2009.

References (in alphabetical order)

- Christophe Croux, Professor of Statistics, Katholieke Universiteit Leuven, Belgium: christophe.croux@econ.kuleuven.be, +32 16 326958.
- Patrick J.F. Groenen, Professor of Statistics, Erasmus University Rotterdam, Netherlands: groenen@ese.eur.nl, +31 10 4081281.
- Dick J.C. van Dijk, Professor of Financial Econometrics, Erasmus University Rotterdam, Netherlands: djvandijk@ese.eur.nl, +31 10 4081263.

Tinbergen Institute Research Series

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:

466 J.A. BOLHAAR, *Health Insurance: Selection, Incentives and Search*.
467 T. FARENHORST-YUAN, *Efficient Simulation Algorithms for Optimization of Discrete Event Based on Measure Valued Differentiation*.
468 M.I. OCHEA, *Essays on Nonlinear Evolutionary Game Dynamics*.
469 J.L.W. KIPPERSLUIS, *Understanding Socioeconomic Differences in Health: An Economic Approach*.
470 A. AL-IBRAHIM, *Dynamic Delay Management at Railways: A Semi-Markovian Decision Approach*.
471 R.P. FABER, *Prices and Price Setting*.
472 J. HUANG, *Education and Social Capital: Empirical Evidences from Microeconomic Analyses*.
473 J.W. VAN DER STRAATEN, *Essays on Urban Amenities and Location Choice*.
474 K.M. LEE, *Filtering Non Linear State Space Models: Methods and Economic Applications*.
475 M.J. REINDERS, *Managing Consumer Resistance to Innovations*.
476 A. PARAKHONYAK, *Essays on Consumer Search, Dynamic Competition and Regulation*.
477 S. GUPTA, *The Study of Impact of Early Life Conditions on Later Life Events: A Look Across the Individual's Life Course*.
478 J. LIU, *Breaking the Ice between Government and Business: From IT Enabled Control Procedure Redesign to Trusted Relationship Building*.
479 D. RUSINOVA, *Economic Development and Growth in Transition Countries*.
480 H. WU, *Essays on Top Management and Corporate Behavior*.
481 X. LIU, *Three Essays on Real Estate Finance*.
482 E.L.W. JONGEN, *Modelling the Impact of Labour Market Policies in the Netherlands*.
483 M.J. SMIT, *Agglomeration and Innovations: Evidence from Dutch Microdata*.
484 S. VAN BEKKUM, *What is Wrong With Pricing Errors? Essays on Value Price Divergence*.

485 X. HU, *Essays on Auctions*.
486 A.A. DUBOVIK, *Economic Dances for Two (and Three)*.
487 A.M. LIZYAYEV, *Stochastic Dominance in Portfolio Analysis and Asset Pricing*.
488 B. SCHWAAB, *Credit Risk and State Space Methods*.
489 N. BASTÜRK, *Essays on parameter heterogeneity and model uncertainty*.
490 E. GUTIÉRREZ PUIGARNAU, *Labour markets, commuting and company cars*.
491 M.W. VORAGE, *The Politics of Entry*.
492 A.N. HALSEMA, *Essays on Resource Management: Ownership, Market Structures and Exhaustibility*.
493 R.E. VLAHU, *Three Essays on Banking*.
494 N.E. VIKANDER, *Essays on Teams and the Social Side of Consumption*.
495 E. DEMIREL, *Economic Models for Inland Navigation in the Context of Climate Change*.
496 V.A.C. VAN DEN BERG, *Congestion pricing with Heterogeneous travellers*.
497 E.R. DE WIT, *Liquidity and Price Discovery in Real Estate Assets*.
498 C. LEE, *Psychological Aspects of the Disposition Effect: An Experimental Investigation*.
499 M.H.A. RIDHWAN, *Regional Dimensions of Monetary Policy in Indonesia*.
500 J. GARCÍA, *The moral herd: Groups and the Evolution of Altruism and Cooperation*.
501 F.H. LAMP, *Essays in Corporate Finance and Accounting*.
502 J. SOL, *Incentives and Social Relations in the Workplace*.
503 A.I.W. HINDRAYANTO, *Periodic Seasonal Time Series Models with applications to U.S. macroeconomic data*.
504 J.J. DE HOOP, *Keeping Kids in School: Cash Transfers and Selective Education in Malawi*.
505 O. SOKOLINSKIY, *Essays on Financial Risk: Forecasts and Investor Perceptions*.
506 T. KISELEVA, *Structural Analysis of Complex Ecological Economic Optimal Management Problems*.
507 U. KILINC, *Essays on Firm Dynamics, Competition and Productivity*.
508 M.J.L. DE HEIDE, *R&D, Innovation and the Policy Mix*.
509 F. DE VOR, *The Impact and Performance of Industrial Sites: Evidence from the Netherlands*.
510 J.A. NON, *Do ut Des: Incentives, Reciprocity, and Organizational Performance*.
511 S.J.J. KONIJN, *Empirical Studies on Credit Risk*.
512 H. VRIJBURG, *Enhanced Cooperation in Corporate Taxation*.
513 P. ZEPPINI, *Behavioural Models of Technological Change*.
514 P.H. STEFFENS, *It's Communication, Stupid! Essays on Communication, Reputation and (Committee) Decision-Making*.
515 K.C. YU, *Essays on Executive Compensation: Managerial Incentives and Disincentives*.