Forecasting and Oracle Efficient Econometrics

By Anders Bredahl Kock

A dissertation submitted to
Business and Social Sciences, Aarhus University,
in partial fulfilment of the requirements of
the PhD degree in
Economics and Management

Preface

This thesis was written in the period September 2007-August 2011 during my graduate studies at the School of Economics and Management at Aarhus University and the Department of Economics at the University of California, Berkeley. I would like to thank both of these institutions for providing me with stimulating research environments. I am grateful to the Danish National Research Foundation for funding the Center for Research in the Econometric Analysis of Time Series (CREATES) and hence my PhD studies. This has given me the oportunity to travel to and present my work at several conferences across the world.

I would also like to thank two my supervisors, Niels Haldrup and Timo Teräsvirta. I would like to thank Niels for encouraging me to write a PhD thesis already when he supervised my bachelor thesis and for always providing me with support and guidance. I am grateful to Timo for letting me work with him on three different papers – one of which can be found in this thesis and for always providing me with unbelievably detailed feedback on my work. I look very much forward to continuing this collaboration.

I am grateful to Michael Jansson for inviting me to visit UC Berkeley. I enjoyed the inspiring research environments and courses at the Departments of Economics and Statistics very much. Special thanks also to Svend Erik Graversen for teaching me probability theory and even spending his free time doing so. His rigor and attention to detail is something I will strive to live up to. The value of his enthusiastic help and teaching can not be overestimated for Chapter 3 and will also be of much use in my future career.

I would also like to thank all my friends at the International House in Berkeley for making my stay there an unforgettable experience and for giving me the chance to meet wonderful people from all over the world.

At the School of Economics and Management I would like to thank my fellow PhD students for providing me with a very pleasant evironment with heaps of fun. I also appreciate your patience with me in my less productive moments. Thanks to Malene for trying to raise me and to Johannes and Stefan for helping me with LaTeX-related problems. Thomas, Kajetan and Anders deserve my gratitude for sharing an office with me and bearing over with my messy desk. Kenneth and Christian deserve thanks for having supported me all the way from the beginning of our joint studies at university in 2003. Finally, I would like to thank Anne, Lasse, Laurent, Mikkel, Niels H/S and all the other PhD students for many funny hours.

I would like to thank my parents, Anette and Hans, my brother Henrik, and my aunt Gitte for always being curious and supportive about my endeavours and work.

Anders Bredahl Kock, Aarhus, August 2011

Updated preface

The predefense was held on Monday, October 17th, 2011. I would like to thank the committee consisting of Henning Bunzel, Dick van Dijk and Jurgen Doornik for their careful reading and constructive comments. In particular, I appreciate the numerous suggestions pointing me towards future avenues of research.

Anders Bredahl Kock, Aarhus, November 2011

Contents

English Summary

If it exists, the overall theme for this thesis could be called machine learning in econometrics, i.e. trying to enable computers to learn patterns from data. These patterns could be figuring out which covariates from a huge database are relevant in explaining a particular phenomenon. Having learned these covarites, one can use these to form out of sample predictions – an extemely important topic in econometrics. This thesis conists of three self-contained chapters dealing with various aspects of machine learning, forecasting, and variable selection.

In Chapter 1 we examine the forecast performance of nonlinear models compared to that of linear autoregressions. Linear models have the advantage that they can be understood and analyzed in great detail. However, it might be inappropriate to assume that the generating mechanism of a series is linear. Hence, nonlinear models have become increasingly popular, see e.g. Granger and Teräsvirta (1993) and Teräsvirta, Granger, and Tjøstheim (2010). However, the nonlinear models are still restricted by the fact that modeling takes place within a prespecified family of models. Since the modeler often has little prior knowledge regarding the functional form of the data generating process, choosing the correct family is still not an easy task. If one wants to avoid making this choice, one may apply universal approximators which are able to approximate broad classes of functions arbitrarily well in various metrics.

The universal approximators are data driven in the sense that little a priori knowledge is needed about the functional relationship between the left- and the right-hand side variables. Artificial Neural Networks (ANN) are a particular type of universal approximators which have been applied in numerous forecasting studies such as Stock and Watson (1999a) and Teräsvirta, van Dijk, and Medeiros (2005). Other universal approximators have also been studied, but in our experience no comparison of the relative forecasting performance has been made.

The paper has three purposes. First, a comparison of the forecast perfor-

mance of three universal approximators is made. Second, we wish to investigate the stability of forecasts from the universal approximators since recursive forecasts by nonlinear models can be erratic. Third, we investigate the performance of a forecast combination algorithm developed in the computer science literature. Its main virtue is that its worst case performance can be explicitly bounded independently of the joint distribution of forecasts combined from.

Besides the ANNs the universal approximators considered are the Kolmogorov-Gabor polynomials and Elliptic Basis Function Networks (EBF). The latter nest the more well known Radial Basis Function Networks as a special case. All the universal approximators applied nest the linear autoregression, and we are hence able to investigate how much (if at all) the nonlinear structure adds to the forecasting performance. While a comparison of recursive and direct nonlinear forecasting procedures is of interest in its own right, we focus on the former here. Hence standard direct procedures such as kernel regression are not considered.

In investigating the stability of the forecasts we are particularly interested in dealing with the importance of handling insane forecasts. Here insane forecasts are to be understood as forecasts that are clearly unrealistic in the light of the hitherto observed values of the time series to be forecast. A precise definition shall be given later. Handling insane forecasts turned out to be particularly important for the Kolmogorov-Gabor polynomials since their polynomial structure could yield explosive forecasts.

Forecast combination has a long history in econometrics. The first to study this were Bates and Granger (1969). The literature has proliferated since then, and a recent survey is given in Timmermann (2006). Two caveats apply, however, to many of these combination algorithms. First, nothing can be said a priori about the performance of the algorithm (combination rule) compared to the individual forecasts. And even if bounds are provided, they often depend on the joint distribution of the vector consisting of the forecasts made by the individual models. The *Weighted Average Algorithm* (WAA) of Kivinen and Warmuth (1999) developed in the computer science literature does not share any of these problems. First, explicit loss bounds for the worst case performance of the algorithm are available. Furthermore, these bounds do not depend on the distribution of the vector of forecasts from the individual models.

We argue that forecasting with universal approximators and combining these into a single forecast by an algorithm for which explicit bounds can be derived forms a solid theoretical foundation for combining forecasts. The empirical per-

formance of the universal approximators as well as the WAA will be investigated by considering various monthly postwar macroeconomic data sets for the G7 and the Scandinavian countries.

Chapter 2 is concerned with the task of building Artificial Neural Network Models and using them for forecasting. It is joint work with one of my supervisors, Timo Teräsvirta.

Artificial Neural Networks (ANN) have been quite popular in many areas of science for describing various phenomena and forecasting them. They have also been used in forecasting macroeconomic time series and financial series, see Kuan and Liu (1995) for a successful example on exchange rate forecasting, and Zhang, Patuwo, and Hu (1998) and Rech (2002) for more mixed results. The main argument in their favour is that ANNs are universal approximators, which means that they are capable of approximating arbitrarily accurately functions satisfying only mild regularity conditions. The ANN models thus have a strong nonparametric flavour. One may therefore expect them to be a versatile tool in economic forecasting and adapt quickly to rapidly changing forecasting situations. Recently, Ahmed, Atiya, El Gayer, and El-Shishiny (2010) conducted an extensive forecasting study comprising more than 1000 economic time series from the M3 competition Makridakis and Hibon (2000), and a large number of what they called machine learning tools. They concluded that the ANN model that we are going to consider, the single hidden-layer feedforward ANN model or multi-layer perceptron with one hidden layer, was one of the best or even the best performer in their study. A single hidden-layer ANN model is already a universal approximator; see Cybenko (1989) and Hornik, Stinchcombe, and White (1989).

A major problem in the application of ANN models is the specification and estimation of these models. A large number of modelling strategies have been developed for the purpose. It is possible to begin with a small model and increase its size ("specific-to-general", "bottom up", or "growing the network"). Conversely, one can specify a network with a large number of variables and hidden units or "neurons" and then reduce its size ("general-to-specific", "top down" or "pruning the network"). Since the ANN model is nonlinear in parameters, its parameters have to be estimated numerically, which may be a demanding task if the number of parameters in the model is large. Recently, White (2006) devised a clever strategy for modelling ANNs that converts the specification and ensuing nonlinear estimation problem into a linear model selection problem. This greatly simplifies the estimation stage and alleviates the computational effort. It is there-

fore of interest to investigate how well this strategy performs in macroeconomic forecasting. A natural benchmark in that case is a linear autoregressive model.

Quite often, application of White's strategy leads to a situation in which the number of variables in the set of candidate variables exceeds the number of observations. The strategy handles these cases without problems, because it essentially works from specific to general and then back again. We shall also consider a one-way variant from specific to general in this study. One may want to set a maximum limit for variables to be included in the model to control its size.

There exist other modelling strategies that can also be applied to selecting the variables. In fact, White (2006) encouraged comparisons between his method and other alternatives, and here we shall follow his suggestion. In this work, we consider two additional specification techniques. One is Autometrics by Doornik (2009), see also Krolzig and Hendry (2001) and Hendry and Krolzig (2005), and the other one is the Marginal Bridge Estimator (MBE), see Huang, Horowitz, and Ma (2008b). The former is designed for econometric modelling, whereas the latter one has its origins in statistics. Autometrics works from general to specific, and the same may be said about MBE. We shall compare the performance of these three methods when applying White's idea of converting the specification and estimation problem into a linear model selection problem and selecting hidden units for our ANN models. That is one of the main objectives of this paper.

The focus in this study is on multiperiod forecasting. There are two ways of generating multiperiod forecasts. One consists of building a single model and generating the forecasts for more than one period ahead recursively. The other one, called direct forecasting, implies that a separate model is built for each forecasting horizons, and no recursions are involved. For discussion, see for example Teräsvirta (2006), Teräsvirta et al. (2010, Chapter 14), or Kock and Teräsvirta (2011). In nonlinear forecasting, the latter method appears to be more common, see for example Stock and Watson (1999b) and Marcellino (2002), whereas Teräsvirta, van Dijk, and Medeiros (2005) constitutes an example of the former alternative. A systematic comparison of the performance of the two methods exists, see Marcellino, Stock, and Watson (2006), but it is restricted to linear autoregressive models. Our aim is to extend these comparisons to nonlinear ANN models.

Nonlinear models can sometimes generate obviously insane forecasts. One way of alleviating this problem is to use insanity filters as in Swanson and White

(1995, 1997a,b) who discuss this issue. We will compare two filters to the unfiltered forecasts and see how they impact on the forecasting performance of the neural networks.

In this work the ANN models are augmented by including lags of the variable to be forecast linearly in them. As a result, the augmented models nest a linear autoregressive model. It is well known that if the data-generating process is linear, the augmented ANN model is not even locally identified; see for example Lee, White, and Granger (1993), Teräsvirta, Lin, and Granger (1993) or Teräsvirta et al. (2010, Chapter 5) for discussion. A general discussion of identification problems in ANN models can be found in Hwang and Ding (1997). It may then be advisable to first test linearity of each series under consideration before applying any ANN modelling strategy to it. But then, it may also be argued that linearity tests are unnecessary, because the set of candidate variables can be (and in our case is) defined to include both linear lags and hidden units. The modelling technique can then choose among all of them and find the combination that is superior to the others. We shall compare these two arguments. This is done by carrying out pretesting and only fitting an ANN model to the series if linearity is rejected. Forecasts are generated from models specified this way and compared with forecasts from the ANN models obtained using White's method and the three automatic modelling techniques.

The main criterion of comparing forecasts is the Root Mean Square Forecast Error (RMSFE), which implies a quadratic loss function. Other alternatives are possible, but the RMSFE is commonly used and thus even applied here. We rank the methods, which makes some comparisons possible. Furthermore, we also carry out Wilcoxon signed rank tests but principally for descriptive purposes, so the tests are not used as an ex post model selection criterion; see Costantini and Kunst (2011) for a discussion.

It might be desirable to compare White's method with modelling strategies which are not based on linearising the problem but in which statistical methods such as hypothesis testing and nonlinear maximum likelihood estimation are applied. Examples of these include Swanson and White (1995, 1997a,b), Anders and Korn (1999) and Medeiros, Teräsvirta, and Rech (2006). These approaches do, however, require plenty of human resources, unless the number of time series under consideration and forecasts generated from them are small. This is because nonlinear iterative estimation cannot be automated and the algorithms left to their own devices. Each estimation needs a nonnegligible amount of tender loving care, and when the number of time series to be considered is large,

ANN model building and forecasting tend to require a substantial amount of resources.

In this paper we investigate the forecasting performance of the above techniques. We first conduct a small simulation study to see how well these techniques perform when the data are generated by a known nonlinear model. The economic data sets consist of the monthly unemployment and consumer price index series from the 1960's until 2009.

Chapter 3 generalizes some recent results on oracle efficient variable selection in cross sectional models to a panel data setting.

When building a model one of the first steps is to decide which variables to include. Sometimes theory can guide the researcher towards a set of potential explanatory variables but which variables in this set are relevant and which are to be left out? Huang, Horowitz, and Ma (2008a) showed that the Bridge estimator is able to discriminate between relevant and irrelevant explanatory variables in a cross section setting with fixed covariates whose number is allowed to increase with the sample size. In fact, oracle efficient estimation has received quite some attention in the statistics literature in the recent years, see (among others) Zou (2006), Candes and Tao (2007), Fan and Lv (2008), and Meinshausen and Yu (2009). However, we are not aware of any similar results for panel data models. For the case of fewer explanatory variables than observations we show that the oracle efficiency of the Bridge estimator carries over to linear panel data models with random regressors in the random and fixed effects settings. More precisely, it suffices that either the number of cross sectional units (N) or the number of observations within each cross sectional unit (T_N) goes to infinity in order to establish consistency and correct elimination of irrelevant variables. To obtain the oracle efficient asymptotic distribution (the distribution obtained by only including the relevant covariates) of the estimators of the nonzero coefficients, further restrictions are needed. In the classical setting of fixed T_N and large N these restrictions are satisfied. Further sufficient conditions for oracle efficiency are given. By fixing T_N and the number of covariates we obtain as a corollary that the asymptotic distribution of the estimators of the non-zero coefficients is exactly the classical fixed effects or random effects limit law.

If the set of potential explanatory variables is larger than the number of observations we show that the Marginal Bridge estimator of Huang et al. (2008a) can be used to distinguish between relevant and irrelevant variables in random and fixed effects panel data models. A partial orthogonality condition restricting the dependence between the relevant and the irrelevant variables of the same

type as in Huang et al. (2008a) is imposed. Furthermore, the error terms must be Gaussian – a price paid for letting the covariates be random. The random covariates also rendered the maximum inequalities based on exponential Orlicz norms used in Huang et al. (2008a) inapplicable. However, more simple maximum inequalities in L^q spaces can still be applied but the result is that the number of irrelevant variables must be $o(N^{q/2})$ for some $q \geq 1$ (this is for fixed T_N for comparability to the known cross sectional results) as opposed to $\exp(o(N))$ (a subexponential rate). Since q is arbitrary this still allows m_N to increase at any polynomial rate. The number of relevant variables may still be $o(N^{1/2})$ (again T_N is considered fixed for comparison).

Furthermore, the Marginal Bridge estimator is very fast to implement which also makes it useful as an initial screening device to weed out the most irrelevant variables before initiating the actual modeling stage.

Since cross section data can be viewed as panel data with only one observation per individual, all our results are also valid for cross section data and hence generalize the results for these.

Bibliography

Ahmed, N. K., A. F. Atiya, N. El Gayer, and H. El-Shishiny (2010). An empirical comparison of machine learning tools for time series forecasting. *Econometric Reviews 29*, 594–621.

Anders, U. and O. Korn (1999). Model selection in neural networks. *Neural Networks 12*, 309–323.

Bates, J. and C. Granger (1969). Combination of forecasts . *Operations Research Quarterly 20*(4), 451–468.

Candes, E. and T. Tao (2007). The Dantzig selector: statistical estimation when p is much larger than n. *Annals of Statistics 35*(6), 2313–2351.

Costantini, M. and R. M. Kunst (2011). On the usefulness of the Diebold-Mariano test in the selection of prediction models: some Monte Carlo evidence. Working paper, University of Vienna.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems 2*, 303–314.

Doornik, J. A. (2009). Autometrics. In J. L. Castle and N. Shephard (Eds.), *The Methodology and Practice of Econometrics*, pp. 88–122. Oxford University Press, Oxford.

Fan, J. and J. Lv (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70*(5), 849–911.

Granger, C. W. J. and T. Teräsvirta (1993). *Modelling Nonlinear Economic Relationships*. Oxford University Press, USA.

Hendry, D. F. and H. M. Krolzig (2005). The properties of automatic Gets modelling. *Economic Journal 115*, 32–61.

Hornik, K., M. Stinchcombe, and H. White (1989). Multilayer feedforward networks are universal approximators. *Neural Networks 2*, 359–366.

Huang, J., J. Horowitz, and S. Ma (2008a). Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Annals of Statistics 36*(2), 587–613.

Huang, J., J. L. Horowitz, and S. Ma (2008b). Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Annals of Statistics 36*, 587–613.

Hwang, J. T. G. and A. A. Ding (1997). Prediction intervals for artificial neural networks. *Journal of the American Statistical Association 92*, 748–757.

Kivinen, J. and M. K. Warmuth (1999). Averaging Expert Predictions. *In Paul Fischer and Hans Ulrich Simon, editors, Proceedings of the 4th European Conference on Computational Learning Theory EuroCOLT '99*, 153–167.

Kock, A. B. and T. Teräsvirta (2011). Forecasting with nonlinear time series models. In M. P. Clements and D. F. Hendry (Eds.), *Oxford Handbook of Economic Forecasting*, pp. 61–87. Oxford University Press, Oxford.

Krolzig, H. M. and D. F. Hendry (2001). Computer automation of general-to-specific model selection procedures. *Journal of Economic Dynamics and Control 25*, 831–866.

Kuan, C.-M. and T. Liu (1995). Forecasting exchange rates using feedforward and recurrent neural networks. *Journal of Applied Econometrics 10*, 347–364.

Lee, T.-H., H. White, and C. W. J. Granger (1993). Testing for neglected nonlinearity in time series models: A comparison of neural network methods and alternative tests. *Journal of Econometrics 56*, 269–290.

Makridakis, S. and M. Hibon (2000). The M3-Competition: results, conclusions and implications. *International Journal of Forecasting 16*, 451–476.

Marcellino, M. (2002). Instability and non-linearity in the EMU. Discussion Paper No. 3312, Centre for Economic Policy Research.

Marcellino, M., J. H. Stock, and M. W. Watson (2006). A comparison of direct and iterated multistep AR methods for forecasting macroeconomic time series. *Journal of Econometrics 135*, 499–526.

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

Meinshausen, N. and B. Yu (2009). Lasso-type recovery of sparse representations for high-dimensional data. *Annals of Statistics 37*(1), 246–270.

Rech, G. (2002). Forecasting with artificial neural network models. SSE/EFI Working Paper Series in Economics and Finance 491, Stockholm School of Economics.

Stock, J. H. and M. W. Watson (1999a). *A comparison of linear and nonlinear univariate models for forecasting macroeconomic time series, in R.F. Engle and H. White (eds)*. Oxford: Oxford University Press.

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

Swanson, N. R. and H. White (1995). A model-selection approach to assessing the information in the term structure using linear models and artificial neural networks. *Journal of Business & Economic Statistics 13*, 265–275.

Swanson, N. R. and H. White (1997a). A model selection approach to real-time macroeconomic forecasting using linear models and artificial neural networks. *Review of Economics and Statistics 79*, 540–550.

Swanson, N. R. and H. White (1997b). Forecasting economic time series using flexible versus fixed specification and linear versus nonlinear econometric models. *International Journal of Forecasting 13*, 439–461.

Teräsvirta, T. (2006). Forecasting economic variables with nonlinear models. In G. Elliott, C. W. J. Granger, and A. Timmermann (Eds.), *Handbook of Economic Forecasting*, Volume 1, pp. 413–457. Elsevier, North-Holland.

Teräsvirta, T., C. W. J. Granger, and D. Tjøstheim (2010). *Modelling Nonlinear Economic Time Series*. Oxford University Press, Oxford.

Teräsvirta, T., C. F. Lin, and C. W. J. Granger (1993). Power of the neural network linearity test. *Journal of Time Series Analysis 14*, 209–220.

Teräsvirta, T., D. van Dijk, and M. C. Medeiros (2005). Linear models, smooth transition autoregressions, and neural networks for forecasting macroeconomic time series: A re-examination. *International Journal of Forecasting 21*(4), 755–774.

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

Timmermann, A. (2006). Forecast Combination, in G. Elliott, C.W.J. Granger, A Timmermann (eds). *Handbook of Economic Forecasting 1*, 135–196.

White, H. (2006). Approximate nonlinear forecasting methods. In G. Elliott, C. W. J. Granger, and A. Timmermann (Eds.), *Handbook of Economic Forecasting*, Volume 1, pp. 459–512. Elsevier, Amsterdam.

Zhang, G., B. E. Patuwo, and M. Y. Hu (1998). Forecasting with artificial neural networks: The state of the art. *International Journal of Forecasting 14*, 35–62.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association 101*(476), 1418–1429.

Dansk resumé

Dansk resumé (Danish summary)

Om den findes, så er den røde tråd i denne afhandling anvendelsen af teknikker fra maskinlæring til at finde mønstre og information i økonomisk data. De seneste år har fremvist en stor vækst i store datasæt, og et oplagt spørgsmål er i den forbindelse, hvordan man finder frem til den ønskede information i data uden at blive overdøvet af støj. At finde frem til de variable, der faktisk driver den proces man er interesseret i, er af stor betydning, når man laver forecasts af fremtidige økonomiske variable. Denne afhandling beskæftiger sig med, hvordan man finder frem til de relevante forklarende variable i meget store datasæt. Vi beskriver også flere metoder til at forecaste økonomiske variable på.

Kapitel 1 har titlen "Forecasting with Universal Approximators and a Learning Algorithm" og omhandler forecasting i situationer, hvor man ikke er villig til at gøre sig særlige antagelser om den datagenererende proces. Til dette formål benytter vi os af mængder af funktioner, der ligger tæt i diverse større funktionsrum. Ofte kaldes disse tætte mængder af funktioner for universelle approksimatorer, da de netop har den egenskab at de kan approksimere brede klasser af funktioner til en vilkårlig grad (de ligger tæt i dem). Kapitlet har tre formål, hvoraf det første er at sammenligne præcisionen af neurale netværk, elliptiske basisfunktionsnetværk samt Kolmogorov-Gabor-polynomier som alle er universelle approksimatorer. For det andet undersøges stabiliteten af forecasts med universelle approksimatorer, idet disse kan have en tendens til at være eksplosive. Endeligt sammenlignes forskellige måder at kombinere forecasts på. Specielt er vi interesseret i "The Weighted Average Algorithm", da denne på langt sigt beviseligt ikke vil klare sig dårligere end den bedste af de modeller fra den mængde af modeller man kombinerer fra. Så selvom vi ikke ved hvilken model, der er den bedste, kan vi klare os lige så godt på langt sigt som om vi kendte den.

Vi finder, at de elliptiske basisfunktionsnetværk er de mest præcise af de universelle approksimatorer, samt at det er nødvendigt at filtrere for vilde forecasts bort, hvis man ønsker en bare nogenlunde hæderlig præcision. Endeligt er det en god ide at kombinere forecasts, men the Weighted Average Algorithm klarer sig ikke meget bedre end mere simple kombinationsskemaer så som ligelig vægtning.

Kaptitel 2, "Forecasting Macroconomic Variables using Neural Network Models and Three Automated Model Selection Techniques", er skrevet sammen med den ene af mine vejledere, Timo Teräsvirta. Dette kapitel går mere i dybden med én af de universelle approksimatorer fra Kapitel 1, nemlig de neurale netværk. Disse er ikke lette at estimere, og vi benytter et snedigt lineariseringstrick til at gøre dette. Efter denne linearisering har fundet sted, skal vi dog vælge mellem en lang række af potentielle variable. Der er altså tale om et højdimensioalt variabelselektionsproblem og vi sammenligner, hvordan QuickNet, Marginal Bridge estimatoren og Autometrics klarer sig i forecasthenseende med hensyn til at vælge variable. Oftest forecaster man enten direkte eller rekursivt og vi kender ikke til nogen sammenligninger af disse metoder for ikke-lineære modeller. Vi udfylder dette hul i litteraturen ved at sammenligne direkte og rekursive forecasts for neurale netværk. Ligesom i kapitel 1 undersøger vi forskellige måder at håndtere vilde forecasts på. Endeligt undersøger vi, om det kan betale sig at teste for linearitet før man bygger en ikke-lineær model.

Vi finder, at Marginal Bridge estimatoren er bedst til at estimere de neurale netværk, samt at direkte forecast ofte er mere præcise end rekursive forecasts. Det lader ikke til at det at teste linearitetshypotesen før vi bygger en ikke-lineær model resulterer i mere præcise forecasts. Dette kan dog skyldes, at alle vores ikke-lineære modeller indeholder en lineær model som ægte delmængde.

Tredje kapitel omhandler variabelselektion i højdimensionale modeller, dvs modeller hvor antallet af variable tillades at vokse i stikprøvestørrelsen og potentielt overstiger denne. Dette emne har fået meget opmærksomhed i de seneste år i statistiklittereaturen. Vi bidrager til denne ved at udvide resultater for Bridge og Marginal Bridge estimatorerne fra cross section modeller til panel data modeller. Specielt viser vi, at Bridge estimatoren er orakelefficient. Det vil sige at den korrekt kan skelne mellem relevante og irrelevante variable og kun beholder de relevante variable i modellen. Ydermere er den asymptotiske fordeling for koefficienterne hørende til de relevante variable den samme som om man fra starten af kun havde inkluderet de relevante variable. Vi kan altså opnå samme asymptotiske efficiens som om et orakel havde fortalt os, hvad den sande model

er.

Ovenstående resultater er for tilfældet, hvor der er færre variable end observationer. I tilfældet med flere variable end observationer viser vi at Marginal Bridge estimatoren stadig kan skelne mellem relevante og irrelevante variable. Dog estimeres koefficienterne til de relevante variable ikke længere konsistent, og vi forreslår derfor en totrinprocedure, hvor man i første trin vælger de relevante forklarende variable ved hjælp af Marginal Bridge estimatoren og i andet trin estimerer koefficienterne til de relevante variable med en vilårlig konsistent estimator såsom OLS eller Bridge estimatoren. Endeligt viser vi ved simulationer, at de foreslåede estimatorer klarer sig godt i endelige stikprøver.

Chapter 1

Forecasting with Universal Approximators and a Learning Algorithm

Anders Bredahl Kock
Aarhus University and CREATES

Journal of Time Series Econometrics (2011), Vol. 3: Iss. 3, Article 3

Abstract

This paper applies three universal approximators for forecasting. They are the Artificial Neural Networks, the Kolmogorov-Gabor polynomials, as well as the Elliptic Basis Function Networks. We are particularly interested in the relative performance and stability of these. Even though forecast combination has a long history in econometrics focus has not been on proving loss bounds for the combination rules applied. We apply the Weighted Average Algorithm (WAA) of Kivinen and Warmuth (1999) for which such loss bounds exist. Specifically, one can bound the worst case performance of the WAA compared to the performance of the best single model in the set of models combined from. The

Financial support from the Center for Research in the Econometric Analysis of Time Series, CREATES funded by the Danish National Research Foundation is gratefully acknowledged by the author. The author also wishes to thank Niels Haldrup, Peter Reinhard Hansen, Timo Teräsvirta, and an anonymous referee for helpful comments. Finally, the author would like to thank his PhD committee consisting of Henning Bunzel, Dick van Dijk and Jurgen Doornik for their careful reading and constructive comments.

use of universal approximators along with a combination scheme for which explicit loss bounds exist should give a solid theoretical foundation to the way the forecasts are performed. The practical performance will be investigated by considering various monthly postwar macroeconomic data sets for the G7 as well as the Scandinavian countries.

1.1 Introduction

In this paper we examine the forecast performance of nonlinear models compared to that of linear autoregressions. Linear models have the advantage that they can be understood and analyzed in great detail. However, it might be inappropriate to assume that the generating mechanism of a series is linear. Hence, nonlinear models have become increasingly popular, see e.g. Granger and Teräsvirta (1993) and Teräsvirta et al. (2010). However, the nonlinear models are still restricted by the fact that modeling takes place within a prespecified family of models. Since the modeler often has little prior knowledge regarding the functional form of the data generating process, choosing the correct family is still not an easy task. If one wants to avoid making this choice, one may apply universal approximators which are able to approximate broad classes of functions arbitrarily well in a way to be made clear in Section 1.2.

The universal approximators are data driven in the sense that little a priori knowledge is needed about the functional relationship between the left- and the right-hand side variables. Artificial Neural Networks (ANN) are a particular type of universal approximators which have been applied in numerous forecasting studies such as Stock and Watson (1999) and Teräsvirta et al. (2005). Other universal approximators have also been studied, but in our experience no comparison of the relative forecasting performance has been made.

The paper has three purposes. First, a comparison of the forecast performance of three universal approximators is made. Second, we wish to investigate the stability of forecasts from the universal approximators since recursive forecasts by nonlinear models can be erratic. Third, we investigate the performance of a forecast combination algorithm developed in the computer science literature. Its main virtue is that its worst case performance can be explicitly bounded independently of the joint distribution of forecasts combined from.

Besides the ANNs the universal approximators considered are the Kolmogorov-Gabor polynomials and Elliptic Basis Function Networks (EBF). The latter nest the more well known Radial Basis Function Networks as a special

case. All the universal approximators applied nest the linear autoregression, and we are hence able to investigate how much (if at all) the nonlinear structure adds to the forecasting performance. While a comparison of recursive and direct nonlinear forecasting procedures is of interest in its own right, we focus on the former here. Hence standard direct procedures such as kernel regression are not considered.

In investigating the stability of the forecasts we are particularly interested in dealing with the importance of handling insane forecasts. Here insane forecasts are to be understood as forecasts that are clearly unrealistic in the light of the hitherto observed values of the time series to be forecast. A precise definition shall be given later. Handling insane forecasts turned out to be particularly important for the Kolmogorov-Gabor polynomials since their polynomial structure could yield explosive forecasts.

Forecast combination has a long history in econometrics. The first to study this were Bates and Granger (1969). The literature has proliferated since then, and a recent survey is given in Timmermann (2006). Two caveats apply, however, to many of these combination algorithms. First, nothing can be said a priori about the performance of the algorithm (combination rule) compared to the individual forecasts. And even if bounds are provided, they often depend on the joint distribution of the vector consisting of the forecasts made by the individual models. The *Weighted Average Algorithm* (WAA) of Kivinen and Warmuth (1999) developed in the computer science literature does not share any of these problems. First, explicit loss bounds for the worst case performance of the algorithm are available. Furthermore, these bounds do not depend on the distribution of the vector of forecasts from the individual models.

We argue that forecasting with universal approximators and combining these into a single forecast by an algorithm for which explicit bounds can be derived forms a solid theoretical foundation for combining forecasts. The empirical performance of the universal approximators as well as the WAA will be investigated by considering various monthly postwar macroeconomic data sets for the G7 and the Scandinavian countries.

The outline of the paper is as follows. Section 1.2 introduces the universal approximators applied in the paper and contains a review of important theoretical results. Next, Section 1.3 introduces the benchmark models, and Section 1.4 discusses forecasting with expert advice with particular emphasis on the Weighted Average Algorithm and its theoretical underpinnings. Section 1.5 presents the results of the application, and Section 1.6 concludes.

1.2 Universal Approximators

We begin by defining precisely what we mean by universal approximators and then discuss the three types employed in this paper.

In order to define universal approximators some preliminary notation is necessary. Let X be a topological space and A a subset of X. Let $\bar{A}$ denote the closure of A. Then A is dense in X if $\bar{A} = X$. Since all topologies used in this paper will be induced by metrics, we may define the closure of A as

$$\bar{A} = \left\{ x \in X \mid \exists (x_n)_{n \geq 1} \subseteq A \text{ such that } d(x_n, x) \to 0 \right\},$$

where d is the metric on X. Hence, A is dense in X if for each $x \in X$ one can choose an element $a \in A$ that is arbitrarily close (in the metric on X) to x. We are now ready to define what we mean by universal approximators.

Definition 1. *Let $\mathscr{H}$ be a subset of functions of a topological space $\mathscr{F}$. Then $\mathscr{H}$ is a universal approximator of $\mathscr{F}$ if $\bar{\mathscr{H}} = \mathscr{F}$.*

An example could be $C_C(\mathbb{R}^n)$, the compactly supported continuous functions on $\mathbb{R}^n$, being dense in $L^p(\lambda_n)$ for $1 \leq p < \infty$, where λ_n is the Lebesgue measure on $(\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n))$ with $\mathscr{B}(\mathbb{R}^n)$ denoting the Borel σ-field on $\mathbb{R}^n$. So for any function $f \in L^p(\lambda_n)$ one can choose a function $h \in C_C(\mathbb{R}^n)$ that is arbitrarily close to f, where closeness is expressed in terms of the metric induced by the L^p-norm. This result can be found many places, see e.g. Rudin (1987) or Kallenberg (1997). Next, we will introduce the universal approximators used in this paper.

Artificial Neural Networks

Artificial Neural Networks (ANN) form a very popular family of universal approximators. They are defined in the following way:

$$\mathscr{H}_{ANN} = \left\{ h : \mathbb{R}^n \to \mathbb{R} \mid h(x) = \sum_{i=1}^{q} \beta_i G(x' \gamma_i + \delta_i),\ \beta_i, \delta_i \in \mathbb{R}, \gamma_i \in \mathbb{R}^n,\ q \in \mathbb{N} \right\}$$

To be precise, $\mathscr{H}_{ANN}$ is the set of *single hidden layer* neural network models. Hornik, Stinchcombe, and White (1989) show that if one chooses the hidden

units $G : \mathbb{R} \to \mathbb{R}$ to be nondecreasing sigmoidal functions[1] (i.e. a squashing function), then $\mathscr{H}_{ANN}$ is uniformly dense on compacta in $C(\mathbb{R}^n)$. More formally, for any $f \in C(\mathbb{R}^n)$ and any compact set $K \subseteq \mathbb{R}^n$, it holds that

$$\forall \varepsilon > 0 \; \exists h \in \mathscr{H}_{ANN} : \sup_{x \in K} |f(x) - h(x)| < \varepsilon.$$

Furthermore, they prove that $\mathscr{H}_{ANN}$ is "dense in measure" in $M(\mathscr{B}(\mathbb{R}^n))$ which denotes the set of Borel measurable functions on $\mathbb{R}^n$. For any finite measure μ on $(\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n))$ and $f, g \in M(\mathscr{B}(\mathbb{R}^n))$ they define the metric

$$\rho_\mu(f,g) = \inf\left\{\varepsilon > 0 \mid \mu(x \in \mathbb{R}^n | \; |f(x) - g(x)| > \varepsilon) < \varepsilon\right\}$$

and show that $\mathscr{H}_{ANN}$ is ρ_μ-dense in $M(\mathscr{B}(\mathbb{R}^n))$. So for any $f \in M(\mathscr{B}(\mathbb{R}^n))$ and any $\varepsilon > 0$ there exists an $h \in \mathscr{H}_{ANN}$ such that $\rho_\mu(f,h) < \varepsilon$. Since convergence in the metric ρ_μ is a metrization of convergence in the measure μ, this result can also be understood as establishing the existence of an $h \in \mathscr{H}_{ANN}$ for which the measure of the set $\{x \in \mathbb{R}^n | f(x) - h(x)| > \varepsilon\}$ can be made arbitrarily small for any $\varepsilon > 0$. Hence, the term dense in measure is appropriate.

Regarding uniform denseness of $\mathscr{H}_{ANN}$ in $M(\mathscr{B}(\mathbb{R}^n))$ Hornik et al. (1989), show that for any $f \in M(\mathscr{B}(\mathbb{R}^n))$ and for any $\varepsilon > 0$ there exists an $h \in \mathscr{H}_{ANN}$ and a compact set $K \subseteq \mathbb{R}^n$ such that $\mu(K^C) < \varepsilon$ and $|f(x) - h(x)| < \varepsilon$ for all $x \in K$.

As a final interesting result we mention that $\mathscr{H}_{ANN}$ is dense in L^p for any $p \in [1, \infty)$ if there exists a compact set $K \subseteq \mathbb{R}^n$ such that $\mu(K^C) = 0$. This is true for any finite measure μ and any squashing function G. The choice of G has not been discussed but an obvious choice is any cumulative distribution function since these are squashing functions (they are sigmoidal and non decreasing).

Elliptic Basis Function Networks

The Elliptic Basis Function Networks (EBF) introduced in Park and Sandberg (1994) have been less frequently applied in econometrics than the more common Artificial Neural Networks. The better known Radial Basis Function Networks

[1] The defining property of a sigmoidal function in Cybenko (1989) is $\sigma : \mathbb{R} \to \mathbb{R}$,

$$\sigma(t) \to \begin{cases} 1 & \text{for } t \to \infty \\ 0 & \text{for } t \to -\infty \end{cases}.$$

Some authors also require σ to be increasing and/or continuous.

(RBF) may be regarded as a special case of the EBF. The set of EBF is defined as

$$\mathscr{H}_{EBF} = \Bigg\{ h : \mathbb{R}^n \to \mathbb{R} |\ h(x) = \sum_{i=1}^{q} w_i G(\frac{x_1 - c_{i1}}{\sigma_{i1}}, ..., \frac{x_n - c_{in}}{\sigma_{in}}),$$

$$c_i, \sigma_i \in \mathbb{R}^n,\ 1 \leq i \leq q,\ q \in \mathbb{N} \Bigg\}.$$

The parameters c_{ij} and σ_{ij} are often referred to as the centroids and width (or smoothing) factors respectively. Though it is not necessary for the theorems stated below to be valid, it is often assumed that $G : \mathbb{R}^n \to \mathbb{R}$ is radially symmetric. Put differently, $G(x) = G(y)$ if $||x|| = ||y||$ where $||.||$ denotes the Euclidean norm on $\mathbb{R}^n$. If G is radially symmetric and $\sigma_{ij} = \sigma_i$ for $j = 1, ..., n,\ i = 1, ..., q,\ \mathscr{H}_{EBF}$ reduces to the set of RBF Networks[2].

A frequent choice of G is the Gaussian, for which

$$G(x) = \exp\left(\frac{-\sum_{j=1}^{n} x_j^2}{2} \right).$$

For this choice the output of the ith hidden unit is given by

$$G(\frac{x_1 - c_{i1}}{\sigma_{i1}}, ..., \frac{x_n - c_{in}}{\sigma_{in}}) = \exp\left(-\sum_{j=1}^{n} \frac{(x_j - c_{ij})^2}{2\sigma_{ij}^2} \right). \tag{1.1}$$

Formula (1.1) simply defines a rescaling of the probability density function of a multivariate Gaussian vector with a diagonal covariance matrix. From (1.1) it is seen why the c_{ij}s are called the centroids. The vector c_i determines where in $\mathbb{R}^n$ the ith hidden unit is centered. In practice one wants to center the hidden units in areas of high data intensity. Since the σ_{ij}'s are allowed to vary across j for each i, the level sets of G will be elliptic (think of $n = 2$) which explains the term *Elliptic Basis Function Network*. In the RBFs the value of G only depends on the distance to the center in the sense that if $||x - c_i|| = ||y - c_i||$ the ith hidden unit takes the same value at x and y. In other words, the level sets are circles (think of $n = 2$ again) - a special case of the ellipse.

Regarding the universal approximation ability of the EBF, it follows from Park and Sandberg (1991, 1993, 1994) that if $G \in L^1(\mathbb{R}^n)$ is bounded and continuous almost everywhere wrt. the Lebesgue measure and satisfies $\int_{\mathbb{R}^n} G(x)dx \neq 0$, then $\mathscr{H}_{EBF}$ is dense in $L^p(\mathbb{R}^n)$ for $1 \leq p < \infty$.

[2] If $\sigma_{ij} = \sigma$ for $j = 1, ..., n,\ i = 1, ..., q$ one still calls $\mathscr{H}_{EBF}$ the set of RBF.

With respect to uniform approximation the following result holds. If G is continuous and satisfies the conditions above then $\mathscr{H}_{EBF}$ is uniformly dense on compacta in $C(\mathbb{R}^n)$. In other words any continuous function may be approximated arbitrarily well in the supremum norm on any compact set. For more results and details, see Park and Sandberg (1991, 1993, 1994).

Finally, we notice that a probability density function is an obvious choice for G in light of the above conditions on G. This is in contrast to the sigmoidal hidden units in the case of ANNs. A cumulative distribution function was an obvious choice in that case. This illustrates an interesting difference between the EBF (RBF) networks and the ANN. The former may be seen as local approaches since probability distribution functions tend to zero as the distance from the centroids goes to infinity. So the hidden units are only active close (locally) to their centroids. The ANN may be seen as a global approach since the hidden units take values close to one for sufficiently large $x'\gamma+\delta$. Put differently, a cumulative distribution function does not tend to 0 as the norm of the input vector goes to infinity. Global effects can, however, cancel out and become local.

Kolmogorov-Gabor polynomials

The set of Kolmogorov-Gabor polynomials is defined as follows:

$$\mathscr{H}_{KG} = \Bigg\{ h : \mathbb{R}^n \to \mathbb{R} \mid h(x) = \phi + \sum_{i_1=1}^{n} \phi_{i_1} x_{i_1} + \sum_{i_1=1}^{n} \sum_{i_2=i_1}^{n} \phi_{i_1 i_2} x_{i_1} x_{i_2} + \dots$$
$$+ \sum_{i_1=1}^{n} \sum_{i_2=i_1}^{n} \dots \sum_{i_q=i_{q-1}}^{n} \phi_{i_1 i_2 \dots i_q} x_{i_1} x_{i_2} \dots x_{i_q},\ \phi, \phi_{i_1 \dots i_j} \in \mathbb{R}, 1 \le j \le q,\ q \in \mathbb{N} \Bigg\}.$$

The Kolmogorov-Gabor polynomials are qth degree polynomials with all possible cross-products included. They are the truncated sum analogue of the Volterra expansions, see Teräsvirta et al. (2010) and references therein for more details.

By the Stone-Weierstrass Theorem it follows that $\mathscr{H}_{KG}$ is uniformly dense on compacta in $C(\mathbb{R}^n)$. $\mathscr{H}_{KG}$ is clearly an algebra of (real) functions on $K \subseteq \mathbb{R}^n$ for any compact set K. It vanishes at no point since $\mathscr{H}_{KG} \ni h(x) = 1 + \sum_{i=1}^{n} x_i^2 > 0$ for all $x \in K$. $\mathscr{H}_{KG}$ also separates points in K. To see this, let $x, y \in \mathbb{R}^n$ and assume $x \neq y$. So for at least one $1 \le i \le n$ it holds that $x_i \neq y_i$. Since $h(x) = x_i$ belongs to $\mathscr{H}_{KG}$, the uniform closure of $\mathscr{H}_{KG}$ consists of all continuous functions on K.

1.3 Benchmark Models

Smooth Transition Models

The Smooth Transition regression model is not a universal approximator. The reason for including it in this work is that it is a benchmark nonlinear model. A standard Smooth Transition regression model is given by

$$y_t = \phi' x_t + \theta' x_t G(s_t) + \varepsilon_t \tag{1.2}$$

where G is the transition function and ε_t is an $i.i.d.$ error term. As is usual in the literature, we choose G to be the logistic function, i.e.,

$$G(s_t) = \frac{1}{1+\exp(-\gamma(s_t - c))}$$

where s_t is the transition variable. Examples include $s_t = y_{t-d}$ for some $d \geq 1$ or $s_t = t$. The parameter γ controls the speed of transition and c determines the position of the transition function.

Autoregressions

Finally, the pth order linear autoregression

$$y_t = c + \sum_{i=1}^{p} \phi_i y_{t-i} + u_t \tag{1.3}$$

for some $p \in \mathbb{N}$ is employed as a benchmark. This allows a check of whether or not the universal approximators are able to outperform the autoregression when it comes to forecasting.

1.4 Forecasting with Experts

In this section we shall focus on the third aim of the paper: forecasting with experts. The emphasis will be on the Weighted Average Algorithm (WAA) of Kivinen and Warmuth (1999). This is one of several algorithms of this type: for other choices, see Freund and Schapire (1997), Littlestone and Warmuth (1994), and Cesa-Bianchi and Lugosi (2006). To establish the notation, consider a setting with n "experts" (models) and let $\mathscr{E}_i$ denote expert i,

$i = 1,2,...,n$, and l the number of trials, i.e., the number of forecasts made with each model. Now consider a sequence $S_l = \left\{(\mathbf{x}_{t+\tau,t}, y_{t+\tau})\right\}_{t=1}^{l}$ where $(\mathbf{x}_{t+\tau,t}, y_{t+\tau}) \in [0,1]^{n+1}$ for $t \in \{1,2,...,l\}$; for every t we have access to a vector of forecasts $\mathbf{x}_{t+\tau,t} = (x_{t+\tau,t,1},...,x_{t+\tau,t,n}) \in [0,1]^n$ whose elements are the τ period ahead forecasts made by each expert at time t. $y_{t+\tau} \in [0,1]$ denotes the actual outcome of the variable to be forecast. Upon observing $y_{t+\tau}$ expert i incurs a loss $L(y_{t+\tau}, x_{t+\tau,t,i})$. A frequently applied loss function to be used in this paper as well is the quadratic loss, i.e., $L(y_{t+\tau}, x_{t+\tau,t,i}) = (y_{t+\tau} - x_{t+\tau,t,i})^2$. The total loss of expert i given the sequence S_l is defined as $L_{\mathscr{E}_i}(S_l) = \sum_{t=1}^{l} L(y_{t+\tau}, x_{t+\tau,t,i})$. Similarly, the total loss of an algorithm A that gives a sequence of weighted forecasts $\left\{\hat{y}_{t+\tau,t}\right\}_{t=1}^{l}$ is $L_A(S_l) = \sum_{t=1}^{l} L(y_{t+\tau}, \hat{y}_{t+\tau,t})$.

In economic applications one cannot assume that $(\mathbf{x}_{t+\tau,t}, y_{t+\tau}) \in [0,1]^{n+1}$. This assumption can, however, be relaxed to $(\mathbf{x}_{t+\tau,t}, y_{t+\tau}) \in [a,b]^{n+1}$. Thus, by choosing $[a,b]$ sufficiently wide, one may circumvent this problem. The exact choice of $[a,b]$ depends on the problem at hand and the procedure used in this paper to determine it will be described in Section 1.5.

The Weighted Average Algorithm. The Weighted Average Algorithm (WAA) of Kivinen and Warmuth (1999) provides a way of combining the forecasts $\mathbf{x}_{t+\tau,t}$ of the experts at trial t into a single forecast. As mentioned in the introduction, an attractive feature of the WAA is that, as opposed to many other forecast combination schemes applied in econometrics (see e.g. Timmermann (2006)), the WAA does not make any assumptions regarding the joint distribution of the forecasts made by the experts. The loss bounds presented below are purely arithmetic results that hold for any distribution of $\mathbf{x}_{t+\tau,t}$. Letting $\mathbf{v}_t$ denote an $n \times 1$ vector of weights, i.e.,

$$\mathbf{v}_t \in \left\{ \mathbf{s} \in \mathbb{R}^n \middle| \sum_{i=1}^{n} s_i = 1,\ s_i \geq 0,\ i = 1,2,...,n \right\}$$

the forecasts of the WAA are given by $\hat{y}_{t+\tau,t} = \mathbf{v}_t' \mathbf{x}_{t+\tau,t}$, $t = 1,...,l$. This way of forecasting explains the terminology *Weighted Average Algorithm* since the forecasts made by the algorithm are weighted averages of the forecasts made by the experts. The weights in the WAA are constructed in the following way:

1. Initialize the algorithm by choosing $\mathbf{v}_1$. If no prior knowledge regarding the performance of the experts is available, an obvious choice is to give equal weights to all of them.

2. For $t = 1, ..., l$

   a) Observe vector of expert forecasts $\mathbf{x}_{t+\tau,t}$.

   b) Calculate the forecast of the algorithm, $\hat{y}_{t+\tau,t} = \mathbf{v}_t' \mathbf{x}_{t+\tau,t}$.

   c) Observe the actual value of y_t.

   d) Update the weights according to

$$v_{t+1,i} = \frac{v_{t,i} \exp(-L(y_t, x_{t,t-\tau,i})/c)}{\sum_{i=1}^{n} v_{t,i} \exp(-L(y_t, x_{t,t-\tau,i})/c)}$$

   where the denominator ensures that the weights sum to 1 and c is a positive constant to be defined below.

Notice that if two experts, $\mathscr{E}_1$ and $\mathscr{E}_2$ have $v_{t,1}/v_{t,2} \neq 1$ due to differences in their previous performance but perform equally well in all future periods, we will not have $v_{t,1}/v_{t,2} \to 1$ as $t \to \infty$. Their ratio will stay unchanged unless they actually incur different losses.

What makes the WAA attractive from a theoretical point of view is the following result in Kivinen and Warmuth (1999):

Theorem 1. *Let $L(y,x)$ be a convex twice differentiable loss function of x for every y. Assume $L_2'(y,y) = 0$. Letting WAA denote the Weighted Average Algorithm with uniform initial weights, i.e., $v_{1,i} = 1/n$, $i = 1, ..., n$, and $S_l = \left\{(\mathbf{x}_{t+\tau,t}, y_{t+\tau})\right\}_{t=1}^{l}$ an arbitrary input sequence, it holds that*

$$L_{WAA}(S_l) \leq \left(\min_{1 \leq i \leq n} L_{\mathscr{E}_i}(S_l) \right) + c \ln(n) \tag{1.4}$$

where c is a constant that depends on the loss function.

In particular, Kivinen and Warmuth (1999) show that it is enough that

$$c \geq \sup_{0 \leq x,y \leq 1} \frac{\left(L_2'(y,x)\right)^2}{L_{22}''(y,x)} \tag{1.5}$$

in order for the inequality (1.4) to be valid. For a quadratic loss function this implies $c \geq 2$. As mentioned above, one cannot in general know in advance

that $(\mathbf{x}_{t+\tau,t}, y_{t+\tau}) \in [0,1]^{n+1}$. But if there exists an interval $[a,b]$ such that $(\mathbf{x}_{t+\tau,t}, y_{t+\tau}) \in [a,b]^{n+1}$ then inequality (1.4) is still valid if one chooses

$$c \geq \sup_{a \leq x,y \leq b} \frac{\left(\frac{d}{dx}(y-x)^2\right)^2}{\frac{d^2}{dx^2}(y-x)^2} = \sup_{a \leq x,y \leq b} \frac{\left(-2(y-x)\right)^2}{2} = 2(b-a)^2.$$

Regarding the conditions on c for other loss functions we refer to Kivinen and Warmuth (1999).

The inequality (1.4) is the theoretical foundation of the Weighted Average Algorithm since it gives an explicit bound to the loss of the algorithm as compared to the best expert in the set of experts. In particular, the WAA will perform no worse than the best expert plus some constant independent of the number of trials. This implies

$$\limsup_{l \to \infty} \frac{L_{WAA}(S_l)}{l} \leq \limsup_{l \to \infty} \frac{\left(\min_{1 \leq i \leq n} L_{\mathscr{E}_i}(S_l)\right)}{l}.$$

Thus, the average loss of the WAA will be no greater than the average loss incurred by the best expert as the number of trials (forecasts) approaches infinity. For more results regarding the WAA we refer to Kivinen and Warmuth (1999) but Theorem 1 gives the main point.

Theorem 1 gives the theoretical motivation for applying the WAA in econometric forecasting. A drawback of the WAA is that it does not apply to the absolute loss function $L(y_{t+\tau}, x_{t+\tau,t,i}) = \left|y_{t+\tau} - x_{t+\tau,t,i}\right|$ due to the non-differentiability of this function. Other algorithms such as the *Hedge-*β algorithm by Freund and Schapire (1997) are available in this case.

1.5 Application

In order to investigate the performance of the models introduced in Sections 1.2 and 1.3 as well as the Weighted Average Algorithm, we consider monthly postwar macroeconomic data sets for the G7 countries as well as the Scandinavian countries including Finland. Five different macroeconomic series were considered for each country: annual Inflation (INF), Industrial Production (IP), long term Interest Rates (I), narrow Money Supply (M), and Unemployment (U). For some countries certain series were missing, and in total 47 series were analyzed. The series have been obtained from the OECD Main Economic Indicators

database and the IMF database. The starting date of the majority of the series is 1960 and most series are available until 2008. The series were seasonally adjusted except for INF and I. For IP and M the models are specified for yearly growth rates.

Estimation and Forecasting Methodology

In this section we describe the details of the estimation procedure for the various models as well as the details of the forecasting procedure. For all series forecasts were made 1, 3, 6, 12, and 24 months ahead. 72 forecasts were made at each horizon for each series. For all series, specification, estimation, and forecasting were carried out using an expanding window (a recursive scheme) with the last window closing 24 months prior to the last observation. All models were respecified and reestimated each time the window was expanded by one observation[3]. All models were univariate and nested the linear autoregression. The details of the individual models will follow.

Autoregressions. For each estimation window, autoregressions with up to five lags were estimated. The one with the lowest value of our Choice Criterion $CC = \log(MSE) + \delta k/T$, where k denotes the number of parameters, $\delta = 1$ and T is the number of observations in the window, was chosen for forecasting. Compared to the Akaike Information Criterion (AIC), in which $\delta = 2$, CC is a rather liberal criterion[4].

Since autoregressions are affine, $E(y_{t+\tau}|\mathscr{F}_t)$ equals the skeleton forecast τ periods ahead made at t (i.e. the recursive forecast ignoring the noise) where $\mathscr{F}_t$ is the σ-algebra generated by $\{y_s\}_{s=1}^t$.

Even though preliminary experiments indicated that an insanity filter as introduced in Swanson and White (1995) was not necessary for the linear autoregression, we adopted the following rule (similar in spirit, but not identical to the one in Swanson and White (1995)) in order to safeguard ourselves against too extreme forecasts. If a forecast did not belong to the interval given by the

[3]By respecification we mean that the number of lags and hidden units/basis functions/cross terms (depending on the type of model) were chosen anew. Reestimation here refers to the fact that parameters were reestimted every time the length of the estimation window was increased by one observation.

[4]The CC criterion was chosen since more standard information criteria such as AIC and BIC (BIC in particular) often constructed ANN models with few or no hidden units, i.e. purely linear models, and so these criteria might not be the right ones for this type of model. Other criteria such as cross validation which we have considered in later work (Kock and Teräsvirta (2011)) could also be applied.

last observation of the estimation window plus/minus three times the standard deviation of the 120 most recent observations in the window, it was replaced by the last observation of the window. In the words of Swanson and White (1995), craziness was replaced by ignorance. The purpose of this insanity filter was to weed out unreasonable forecasts and thereby more closely mimic the behavior of a real forecaster. The reason for only calculating the standard deviation based on the last 120 observations of the window is that for many data sets the standard deviation of all observations in the window is often very high due to large historic fluctuations. As a result, basing the standard deviation on all observations in the window would lead to occasionally accepting wild forecasts.

No Change Forecasts. In order to investigate whether any of the estimated models was able to beat naive No Change (NC) forecasts, these were also included. Inability to beat the NC forecasts can be seen as an indication of a martingale (e.g. a random walk) like behavior of the series considered. This is due to the fact that the No Change forecasts are optimal in an expected square error sense when the underlying process is a martingale.

Smooth Transition Autoregressions. For each window a search over lag orders up to five was performed. The transition variables searched over were 1, 2, 6, and 12 lags of the left-hand side variable. The model with the lowest *CC* value was chosen for forecasting. This model could be, and was indeed quite often, a linear one. In order to avoid biased forecasts the forecasts were generated using the same bootstrap approach as in Teräsvirta et al. (2005). It works in the following way:

Let $y_t = f(y_{t-1},...,y_{t-p};\boldsymbol{\theta}) + \varepsilon_t$ for some parameter vector $\boldsymbol{\theta}$. In our case $f(\cdot)$ is the (nonlinear) function chosen by the *CC* criterion and hence it is known[5]. Letting h denote the maximal forecast horizon, N_B the number of bootstrap replications and $\{\hat{\varepsilon}_t\}$ the sequence of error terms from the estimation we resampled $h-1$ errors N_B times. Put differently, we created $(\hat{\varepsilon}^i_{t+1,t},...,\hat{\varepsilon}^i_{t+h-1,t})$ for $i = 1,...,N_B$ and generated the τ-step ahead forecast in the following way

$$\hat{y}_{t+\tau,t} = \frac{1}{N_B}\sum_{i=1}^{N_B} f(\hat{y}^i_{t+\tau-1,t} + \hat{\varepsilon}^i_{t+\tau-1,t},...,\hat{y}^i_{t+\tau-p,t} + \hat{\varepsilon}^i_{t+\tau-p,t}) \qquad (1.6)$$

with $\hat{y}^i_{t+\tau-j,t} + \hat{\varepsilon}^i_{t+\tau-j,t}$ replaced by $y_{t+\tau-j}$ for $j \geq \tau,\ j = 1,...,p$. In this paper $N_B = 1000$ was used. Furthermore, an insanity filter was applied at the level of the individual bootstrap replications. Specifically, if *any* forecast of a bootstrap

[5]So in case of the Smooth Transition Autoregression in (1.2), $f(y_{t-1},...,y_{t-p};\boldsymbol{\theta}) = \phi' x_t + \theta' x_t G(s_t)$ with $x_t = (y_{t-1},...,y_{t-p})'$.

sample path did not belong to the interval consisting of the last observation of the given window plus minus 3 times the standard deviation of the 120 most recent observations of the window, then the whole bootstrap path was discarded.

Kolmogorov-Gabor polynomials. For each window we searched over models with at most five lags and the highest degree of the polynomial was five. Since the number of parameters increases rapidly as a function of the number of lags and the degree of the polynomial, we implemented a parameter cap of 50 such that specifications containing more than 50 parameters were ignored. Among the remaining models the one with the lowest CC value was chosen. Due to the non-affinity of the Kolmogorov-Gabor polynomials, the forecasts were obtained using the bootstrap technique outlined above. An insanity filter of the same kind as the one applied to the STR was used.

Artificial Neural Networks. In each window *single hidden layer feedforward* networks with at most five lags and five hidden units were estimated. Model specification and estimation were carried out using the *QuickNet* algorithm of White (2006)[6]. The model with the lowest CC value was chosen for forecasting, and as for all non-affine models, the forecasts were generated using a bootstrap approach combined with the insanity filter outlined for the STR.

Elliptic Basis Function Networks. Models with at most five lags and no more than five hidden units were estimated in each window. G was chosen as in (1.1). The models considered were of the following form:

$$y_t = \alpha_q + \beta_q' x_t + \sum_{i=1}^{q} w_{iq} \exp\left(- \sum_{j=1}^{n} \frac{(x_{t,j} - c_{ij})^2}{2(n\sigma_{ij})^2} \right) \tag{1.7}$$

where $x_t = (y_{t-1}, ..., y_{t-n})$, $n = 1, ..., 5$. The multiplication of σ_{ij} by n was done for practical reasons. In some initial experiments the hidden units had a very small radius of activity – in particular if many explanatory variables were included. By rescaling the width parameters proportionally to the number of explanatory variables this numerical problem was alleviated.

EBF networks can be estimated in many ways. We settled for a procedure which learns the centroids and width parameters unsupervised (only the right hand side variables were used to determine the centroids and width parameters

[6]As discussed in White (2006) the parameter estimates from *QuickNet* can in principle be used as starting values in a non linear least squares estimation. However, White (2006) also shows in a series of experiments that this does not seem to improve the forecast accuracy of the ANN. In many cases the estimation simply breaks down in this highly nonlinear problem.

without access to the associated left hand side variables). After having determined these, the problem is linear and the w_is can be found by linear regression. This resembles the structure of *QuickNet* since the problem is split into two parts. First, the centroids and smoothing parameters are found by a grid search. Having done this, the problem becomes linear, and one can determine the w_is by linear regression.

We now explain how our grid is constructed. Let $1 \leq n \leq 5$ be given.

1. For $1 \leq j \leq n$ sort y_{t-j} in ascending order and divide it into five clusters of equally many observations, i.e the splits were made at the quintiles[7].

2. Within each cluster of y_{t-j} calculate the mean and standard deviations. These are our candidates for c_{ij} and σ_{ij} in (1.7). This yields five pairs of centroids and width parameters for each $y_{t-j},\ j = 1,...,n$.

3. Take all possible combinations of centroid-width parameter pairs. This yields a grid of cardinality 5^n. A typical element of the grid takes the form $\left\{(c_{i_1 1},\ \sigma_{i_1 1}),...,(c_{i_n n},\ \sigma_{i_n n})\right\},\ 1 \leq i_1,...,i_n \leq 5$.

We are now in a position to describe the details of our proposed estimation algorithm which one could call the *QuickEBF* due to its similarity to the *QuickNet*. Choose $n \in \{1,...,5\}$ and let $x_t = (y_{t-1},...,y_{t-n})$ be given, i.e., consider a fixed vector of explanatory variables.

1. Determine $\hat{\alpha}_0$ and $\hat{\beta}_0$ by regressing y_t on a constant and x_t. Also calculate the value of the choice criterion CC.

2. Each of the 5^n grid points corresponds to a hidden unit (elliptic basis function). For $q = 1,...,5$, add hidden units one by one in the following way: Determine $\hat{\alpha}_q,\ \hat{\beta}_q$, and $\hat{w}_{i,q},\ i = 1,...,q$ by OLS for each potential hidden unit not previously chosen, i.e., for each point in the grid of c's and σ's. Notice that the weights of previously added hidden units as well as $\hat{\alpha}_q$ and $\hat{\beta}_q$ are allowed to change as further hidden units are added, whereas the centroids and smoothing parameters remain fixed once they have been determined. This resembles the *QuickNet* algorithm of White (2006). Calculate the value of the choice criterion and add the hidden unit which yields the lowest value of CC.

[7] One could also split each explanatory variable into more than five clusters and thereby obtain an even finer grid. Here five clusters were chosen for each series since the maximum number of hidden units was five.

3. Repeat (1) and (2) for all choices of explanatory variables which in our case corresponds to $x_t = (y_{t-1}, ..., y_{t-n})$, $n = 1, ..., 5$. Choose $\hat{n} \in \{1, ..., 5\}$ and $\hat{q} \in \{0, ..., 5\}$ such that they minimize the choice criterion and forecast with the parameter estimates corresponding to these values.

The forecasts of the Elliptic Basis Function Network were produced by the same bootstrap procedure as that for the aforementioned nonlinear models. Unreasonable forecasts were weeded out by the same insanity filter as outlined for the STR.

One notices that the *QuickEBF* is easy to implement since after having fixed the centroids and width parameters it consists of linear regressions. This is convenient since at the outset the EBF is not easy to estimate because the centroids and width parameters are not identified when more than one hidden unit is included. The *QuickEBF* deals with this in an easy way.

Forecast Combinations. As mentioned in the Introduction, a major aim of the work is considering the effects of combining forecasts on forecast accuracy. Two equal weighting schemes were employed: (i) Equal weighting of all models and (ii) equal weighting of the three universal approximators.

As mentioned in Section 1.4, one must assume the existence of an interval $[a,b]$ such that $(\mathbf{x}_{t+\tau,t}, y_{t+\tau}) \in [a,b]^{n+1}$ in order to give explicit loss bounds for the WAA. Our solution to this problem was the following. Let Y denote the last observation in the first estimation window and s the standard deviation of the 120 most recent observations of the first estimation window. Then we chose $[a,b] = [Y - 3s, Y + 3s]$ which in the vast majority of the cases was more than wide enough to contain all forecasts and realizations[8]. The corresponding value of c was denoted c_B. The reason for only calculating s on the basis of the last 120 observations was the same as that given in the treatment of the insanity filter.

The results for the WAA with $c = c_B$ were called $WAA(c_B)$. In order to investigate the performance of the WAA for smaller values of c, i.e., a faster adjustment of the weights towards the models that have performed well in the more recent past, we calculated the forecasts of the WAA with $c = c_L = \frac{c_B}{100}$. These forecasts were called $WAA(c_L)$. Both WAA based forecasts were applied separately to each horizon. This allowed the WAA to attach different weights to each model for different horizons. This is sensible since models performing well at short horizons need not perform well in long term forecasting, and vice versa.

[8]This fits well with the choice of insanity filter since it is exactly observations outside $[Y - 3s, Y + 3s]$ we regard as insane.

All combination schemes in this paper were applied separately to each horizon.

To compare the WAA to another loss based algorithm we performed forecasting using inverse MSE weights with the MSE calculated from all previous forecast errors. Since the loss of a τ periods ahead forecast will not be observed until these τ periods have elapsed, one would have to initialize the weights of the loss based algorithms (WAA and MSE) in some fashion. We did this by not beginning the comparison until the first losses at the relevant horizon were realized. Consequently, for τ periods ahead forecasts the actual number of evaluation periods was 72-τ.

Results

Tables 1.3-1.8 in the Appendix report the Root Mean Square Forecast Errors (RMSFE) of each point forecast relative to the RMSFE of the linear autoregressive specification. The numbers in brackets are the RMSFE of the linear autoregressive specification. Empty sections in the tables refer to series for which data was unavailable.

Inspection of these tables reveals that no single model systematically outperforms the others. This is in line with Teräsvirta et al. (2005). One notices that for some time series there are models which have a relative RMSFE of 1 at all horizons. This indicates that the linear specification was chosen in each window for this model class. Table 1.1 summarizes Tables 1.3-1.8. It shows the RMSFE ratios as well as the rank of each model across all data sets and all horizons (Overall), all horizons (Horizon), and all types of data sets (Data).

Performance of the Universal Approximators

An aim of the paper was to consider the performance of universal approximators. Table 1.1 reveals that the Elliptic Basis Function Networks outperform the Kolmogorov-Gabor polynomials and the Artificial Neural Networks overall. We will explain this result below. All universal approximators perform well for the inflation series with relative RMSFE clearly below 1. In fact, for Denmark, France, and Italy all universal approximators yield more accurate forecasts at all horizons than the linear AR for this series. This is also the case for the forecasts of French industrial production. On the other hand there does not exist a single country/variable combination for which all the universal approximators have relative RMSFE above 1 at all horizons.

		Horizon					Data				
	Overall	1	3	6	12	24	INF	IP	I	M	U
AR	1.000(11)	1.000(7)	1.000(8)	1.000(12)	1.000(8)	1.000(12)	1.000(12)	1.000(8)	1.000(9)	1.000(1)	1.000(7)
NC	0.980(6)	0.999(5)	0.989(7)	0.977(6)	1.008(11)	0.926(7)	0.852(1)	1.118(12)	0.868(1)	1.154(12)	0.990(6)
STR	0.990(8)	1.015(9)	1.006(9)	0.992(9)	1.014(12)	0.923(5)	0.882(9)	1.064(11)	0.962(7)	1.048(8)	1.028(11)
KG	1.005(12)	1.058(12)	1.009(10)	0.991(8)	0.994(7)	0.972(11)	0.873(7)	1.008(9)	1.075(11)	1.082(10)	1.030(12)
ANN	0.998(10)	1.026(11)	1.011(11)	0.998(11)	1.003(9)	0.951(9)	0.877(8)	0.995(7)	1.080(12)	1.049(9)	1.020(10)
EBF	0.980(7)	0.999(6)	0.988(6)	0.981(7)	0.979(6)	0.954(10)	0.927(11)	0.981(5)	0.991(8)	1.017(2)	1.004(8)
EQ(All)	0.948(3)	0.984(3)	0.965(2)	0.950(3)	0.950(4)	0.893(3)	0.860(3)	0.972(3)	0.955(6)	1.027(4)	0.968(1)
EQ(UA)	0.966(5)	1.001(8)	0.977(5)	0.965(5)	0.965(5)	0.923(6)	0.864(5)	0.978(4)	1.011(10)	1.028(6)	0.986(5)
WAA(c_B)	0.948(2)	0.984(2)	0.965(3)	0.949(2)	0.946(2)	0.894(4)	0.863(4)	0.969(2)	0.951(5)	1.027(5)	0.968(2)
WAA(c_L)	0.950(4)	0.988(4)	0.971(4)	0.954(4)	0.948(3)	0.889(2)	0.869(6)	0.992(6)	0.913(2)	1.039(7)	0.980(4)
MSE	0.939(1)	0.983(1)	0.964(1)	0.942(1)	0.934(1)	0.872(1)	0.857(2)	0.965(1)	0.922(3)	1.026(3)	0.969(3)
Last	0.993(9)	1.021(10)	1.016(12)	0.995(10)	1.006(10)	0.927(8)	0.909(10)	1.057(10)	0.949(4)	1.086(11)	1.009(9)

Table 1.1: Relative RMSFE ratios and the corresponding ranks (in ascending order) for the overall, the horizonwise, and datawise performance of each forecast procedure. The ratios were calculated by taking the average over the relevant ratios from Tables 1.3 through 1.8. For example, the horizonwise performance was calculated by taking the average of all ratios at a fixed forecast horizon across all data sets. AR: Linear Autoregression, NC: No Change forecasts, STR: Smooth Transition Regression, KG: Kolmogorov-Gabor polynomial, ANN: Artificial Neural Network, EBF: Elliptic Basis Function network, EQ(All): Equal weighting of all individual models, EQ(UA): Equal weighting of the universal approximators, WAA(c_B): WAA with $c = c_B$, WAA(c_L): WAA with $c = c_L$, MSE: Weighting by inverse MSE, Last: Forecasting with the model which has the best previously realized forecast at the given horizon.

Furthermore, Table 1.1 shows that the performance of the universal approximators in general improves as the forecast horizon gets longer. For the EBF this relationship is even monotonic. This pattern can be found for all three universal approximators at the same time for, for example, the Japanese unemployment series and for the KG and EBF for Danish interest rates. The performance of the KG for the Danish interest rate series also illustrates that the relative RMSFE need not be decreasing when monotonic. In particular, the performance of the KG worsens as the forecast horizon gets longer. The fact that the most extreme forecasts (good as well as bad) are found at the longest forecast horizon is no surprise given that the forecasts are carried out recursively. This is due to the fact that the benefit/loss from a superior/inferior specification is accumulated as the forecasts are iterated forward.

Overall, the EBF is the best performing universal approximator. In fact, Table 1.1 reveals that the linear autoregression is beaten at all forecast horizons by the EBF. The EBF also outperforms the other universal approximators at all horizons except the 24 month horizon where it is outperformed by the ANN. The reason for the promising performance of the EBF is that it takes the best from two extremes. For 21 of the 47 data sets the EBF is actually just the linear AR (no hidden units are chosen)[9]. When this purely linear specification is chosen, it is often the case that the two other universal approximators perform less well compared to the linear AR, see for example the money supply series for Denmark, Germany, and Sweden. When basing the overall performance of the EBF only on those data sets in which the EBF differs from the linear AR, we get an overall relative RMSFE of .96. So for 21 data sets the EBF equals the linear AR, and for those data sets for which the two models differ, the EBF on average does 4% better.

Table 1.2 contains the lowest and highest relative RMSFE of each procedure across all datasets. The table also shows the number of times the relative RMSFE of each procedure was below .9 as well as how often it was above 1.1. It is of interest that out of $47 \times 5 = 235$ relative RMSFE (number of table entries for each procedure in Tables 1.3-1.8) the EBF never has a relative RMSFE exceeding 1.189. This value is much lower than the corresponding value for any of the other universal approximators. The Kolmogorov-Gabor polynomials, for example, have a worst case relative RMSFE of 1.91. Actually the worst case performance of the EBF is better than the worst case performance of any forecasting

[9]For the Swedish unemployment series the EBF has a relative RMSFE of 1 due to rounding error - not due to the fact that a linear model was chosen every time.

method considered - including forecast combinations. Furthermore, it only happens five times that the EBF performs more than 10% worse than the linear AR. This too is better than any other forecasting procedure and much lower than any non-combinatorial method. Hence, it is reasonable to conclude that the risk of very bad forecasts with the EBF is limited. This fact is emphasized by the above finding that the forecasts of the EBF often simply equal the forecasts of the linear AR in cases where the latter performs well. So it is not surprising that the EBF has relatively few extreme forecasts.

It is also worth noticing that the lowest relative RMSFE obtained by the EBF is lower than the corresponding number for the two other universal approximators. But then, the EBF has fewer relative RMSFEs below 0.9 than any other method. Nevertheless, our conclusion is that the EBF is a stable procedure which combines the best from linear and nonlinear forecasting procedures.

	NC	STR	KG	ANN	EBF	EQ(All)	EQ(UA)	WAA(c_B)	WAA(c_L)	MSE	Last
Lowest	0.286	0.297	0.497	0.421	0.411	0.459	0.435	0.437	0.301	0.319	0.356
Highest	1.782	1.555	1.911	1.701	1.189	1.249	1.393	1.241	1.417	1.274	1.557
# < 0.9	67	43	47	33	21	49	37	49	54	52	49
# > 1.1	48	46	43	27	5	11	11	9	13	6	41

Table 1.2: The table contains the lowest as well as the highest relative RMSFE of each procedure (except AR) and the number of relative RMSFE below 0.9 and above 1.1 (except AR) calculated from Tables 1.3 through 1.8. NC: No Change forecasts, STR: Smooth Transition Regression, KG: Kolmogorov-Gabor polynomial, ANN: Artificial Neural Network, EBF: Elliptic Basis Function network, EQ(All): Equal weighting of all individual models, EQ(UA): Equal weighting of the universal approximators, WAA(c_B): WAA with $c = c_B$, WAA(c_L): WAA with $c = c_L$, MSE: Weighting by inverse MSE, Last: Forecasting with the model which has the best previously realized forecast at the given horizon.

Smooth Transition and No Change forecasts

Table 1.1 reveals that the Smooth Transition model performs very well at the 24 month horizon. For the Italian industrial production series it obtains a relative RMSFE as low as 0.297 at this horizon. On the other hand Table 1.2 shows that the STR is the only procedure which has more relative RMSFE above 1.1 than below 0.9. In particular it has 46 values above 1.1 which indicates that it is a less stable procedure than the EBF. As opposed to Stock and Watson (1999) we find that the STR can match the ANN. We also find that the ANN in general has its lowest RMSFE at the longest forecast horizon. This too is in opposition to Stock and Watson (1999). It should be mentioned, however, that Stock and Watson (1999) carry out their forecasts directly while we do it recursively. Hence, the

results are not directly comparable. Finally, we notice that the STR performs well for the interest rate series except for Denmark, Japan, and Norway.

It is well known that the No Change forecasts are optimal in an expected square error sense if the series of interest is a martingale. Table 1.1 indicates that the No Change forecasts perform well for the inflation and interest rate series. In fact they perform better than any other procedure for these series. However, the opposite is the case for the industrial production and money supply series where they perform worse than any other method. This lack of stability is confirmed by Table 1.2 which shows that the No Change forecasts are the procedure with most relative RMSFE below 0.9 and above 1.1. In order to interpret these results we test whether each of the series is a random walk against a stationary AR(1) with intercept. Only for 1 out of 11 inflation series this hypothesis is rejected at a 5% significance level. For the interest rate series this hypothesis is never rejected. Consequently, the No Change forecasts should perform well for these data sets. For the industrial production series the random walk is rejected 5 out of 10 times while the corresponding numbers for the money supply series are 1 out of 6. These results are more mixed. While it is sensible that the random walk hypothesis is rejected rather often for the industrial production series, it is rejected less often than expected for the money supply series.

Importance of the Insanity Filter

Applying an insanity filter turned out to be very important. Forecasting recursively 24 months ahead meant that some of the bootstrap paths yielded clearly unrealistic forecasts. This was particularly pronounced for the Kolmogorov-Gabor polynomials which are of an explosive nature due to their polynomial structure. A single insane forecast can (and turned out to) make even the average over bootstrap paths a bad forecast, and so it is vital to remove insane paths. The problem was less pronounced for the other nonlinear models for which the hidden units are bounded. Still an insanity filter was important for these too - while explosive less often, they could yield very unreasonable forecasts. The number of bootstrap paths removed by the insanity filter could vary greatly. Sometimes very few (none) were removed, while in other cases several hundred were weeded out by the insanity filter.

Forecast Combinations

A general conclusion is that it pays of to combine forecasts. The forecast combination schemes in general outperform the individual methods. This is in line with Stock and Watson (1999) and Teräsvirta et al. (2005). Weighting by inverse MSE performs best overall and is also superior at each forecast horizon as seen from Table 1.1. The Weighted Average Algorithm performs roughly the same for both values of c. Table 1.2 reveals that $c = c_L$ gives more extreme values (good as well as bad) than $c = c_B$. In total these two effects balance each other and WAA(c_B) and WAA(c_L) perform equally well. The inverse MSE scheme performs slightly better than the two versions of the WAA since it has as many good outcomes (< 0.9) as WAA(c_L) and as few bad outcomes (> 1.1) as WAA(c_B), see Table 1.2.

Regarding equal weighting schemes, combinig all models is preferable to only combining the universal approximators. This is the case overall, at all horizons, and for all types of data sets. Table 1.2 indicates that the reason for this difference is more low (< 0.9) relative RMSFEs while the number of high (> 1.1) relative RMSFEs is roughly unaltered compared to only combining the universal approximators.

Finally, Figure 1.1 in the Appendix shows the development of the weights in the WAA for $c = c_L$ for the industrial production series for the UK at all horizons. This series was chosen since it illustrates some interesting features of the WAA. When inspecting Table 1.7 it is not surprising that as the forecast horizon increases more weight is given to the Kolmogorov-Gabor polynomials since these did well at the long horizons for this series. By the same token it is sensible that the linear autoregression receives less weight as the forecast horizon becomes longer. When inspecting the development of the weights at the 24-month horizon in Figure 1.1, it is interesting to see how the weights of the WAA can adapt over time as the relative performance of the individual models changes. In particular, the relative weights assigned to the Elliptic Basis Function Networks and the Kolmogorov-Gabor polynomials change around period 35.

1.6 Conclusions

In this paper we consider the forecasting performance of nonlinear models relative to linear autoregressions. Three of the model classes employed are universal approximators - the Kolmogorov-Gabor polynomials, the Artificial Neural Net-

works, and the Elliptic Basis Function Networks.

Regarding the first question posed in the introduction, we find that there are potential gains to be made from using universal approximators. In particular, the Elliptic Basis Functions turn out to deliver good forecasts. Their main merit is that they combine the best from two worlds. In the cases where the linear autoregressions tend to be superior, the EBF equals these and in the cases where there seem to be gains to be made from using hidden units, these are included into the EBF. As opposed to many other nonlinear procedures, forecasting with an EBF is not too risky either since even the worst performance observed is not too bad. Furthermore, the proposed estimation strategy is not difficult to implement since it essentially amounts to running a series of linear regressions. Hence, we believe that the EBF is a useful addition to the set of macroeconomic forecasters.

Secondly, when forecasting recursively with nonlinear models an insanity filter is definitely advisable since even a single wild bootstrap path can affect the forecast adversely.

The No Change forecasts also perform quite well but their performance is very unstable compared to the one of the EBF.

Thirdly, we find that the WAA is unable to outperform more standard methods. From a practitioners point of view it is encouraging that simple equal weighting does so well since it is easy to understand and implement.

All forecasts are carried out recursively in this paper. Hence a possible extension which we are currently working on is to investigate whether there are gains to be made by forecasting directly with the universal approximators. Other ways of selecting the hidden units could also be studied. Finally, we consider univariate procedures in this paper. Further gains in the forecast accuracy might be achievable by applying the universal approximators - in particular the EBF - in a multivariate setting.

1.7 Appendix

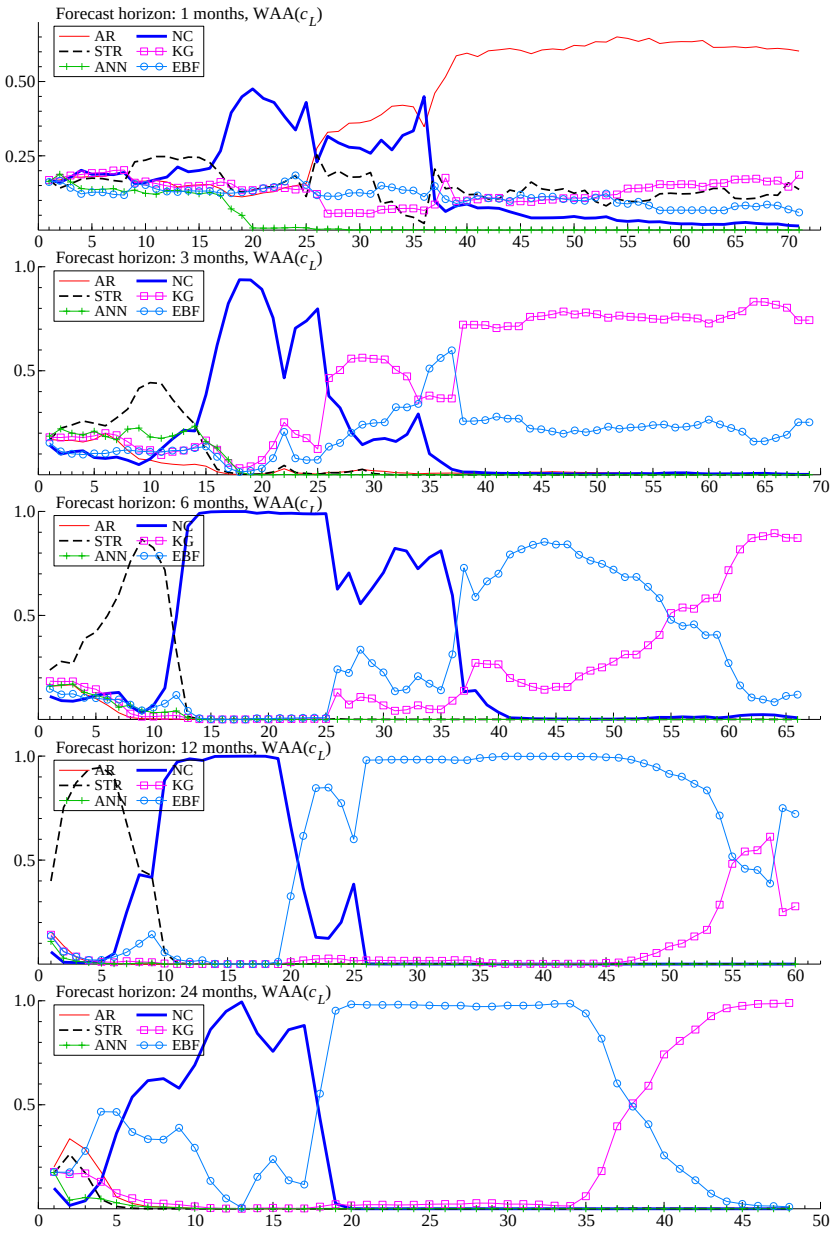

Figure 1.1: Development of the weights of the WAA applied to the Industrial Production series for the UK with $c = c_L$.

	Inflation					Industrial Production					Interest Rate					Money Supply					Unemployment				
Canada	1	3	6	12	24	1	3	6	12	24	1	3	6	12	24	1	3	6	12	24	1	3	6	12	24
AR	[0.538]	[0.940]	[1.295]	[1.644]	[1.284]	[0.519]	[0.769]	[1.165]	[1.253]	[1.377]	[0.172]	[0.274]	[0.399]	[0.672]	[1.144]	[0.923]	[1.674]	[2.263]	[2.510]	[1.823]	[0.149]	[0.248]	[0.363]	[0.475]	[0.813]
NC	0.972	0.983	0.977	0.966	0.794	1.085	1.089	1.062	1.135	0.826	0.949	0.896	0.804	0.666	0.583	1.033	1.086	1.204	1.580	1.186	0.993	0.973	0.966	0.922	0.942
STR	0.999	0.990	0.998	0.991	0.901	0.992	0.968	0.970	1.017	0.730	0.982	0.940	0.867	0.745	0.691	0.996	1.007	1.030	1.198	0.923	1.016	1.033	1.053	1.116	1.132
KG	0.985	0.962	0.940	0.885	0.659	1.044	1.180	1.271	1.308	0.988	1.010	1.025	1.070	1.112	1.078	1.027	1.031	1.106	1.370	1.074	1.030	1.059	1.079	1.080	1.079
ANN	1.013	1.000	0.993	0.966	0.843	1.021	1.111	1.125	1.116	0.881	0.993	1.000	0.995	0.921	0.816	1.021	1.010	1.034	1.189	1.020	1.000	1.000	1.000	1.000	1.000
EBF	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.038	1.067	1.075	1.189	0.931	1.000	1.000	1.000	1.000	1.000
EQ(All)	0.992	0.985	0.980	0.961	0.838	1.005	1.029	1.032	1.010	0.826	0.981	0.964	0.939	0.896	0.849	1.006	1.019	1.062	1.249	0.979	1.001	1.001	1.007	1.010	1.022
EQ(UA)	0.996	0.983	0.973	0.944	0.814	1.011	1.079	1.106	1.084	0.911	0.992	0.999	1.013	1.004	0.951	1.014	1.024	1.064	1.246	1.001	1.006	1.015	1.022	1.020	1.022
WAA(c_B)	0.992	0.985	0.980	0.960	0.834	1.005	1.029	1.031	1.005	0.825	0.981	0.964	0.939	0.894	0.842	1.006	1.019	1.062	1.241	0.977	1.001	1.001	1.007	1.010	1.023
WAA(c_L)	0.990	0.985	0.974	0.910	0.746	1.005	1.016	1.002	0.914	0.718	0.981	0.960	0.899	0.763	0.629	1.008	1.020	1.078	1.297	1.002	1.001	1.001	1.006	1.006	1.046
MSE	0.991	0.986	0.982	0.960	0.819	1.001	1.004	1.013	0.980	0.786	0.980	0.961	0.921	0.816	0.723	1.006	1.021	1.075	1.274	0.965	1.001	1.001	1.006	1.010	1.030
Last	0.975	1.000	0.985	0.977	0.930	1.060	1.063	1.019	1.158	0.882	0.977	0.950	0.916	0.730	0.588	1.022	1.079	1.119	1.371	1.018	1.002	1.001	1.037	1.046	1.072

	Inflation					Industrial Production					Interest Rate					Money Supply					Unemployment				
Denmark	1	3	6	12	24	1	3	6	12	24	1	3	6	12	24	1	3	6	12	24	1	3	6	12	24
AR	[0.256]	[0.416]	[0.634]	[1.048]	[1.472]						[0.197]	[0.394]	[0.573]	[0.920]	[1.272]	[1.307]	[2.705]	[3.655]	[5.303]	[6.379]	[0.098]	[0.197]	[0.359]	[0.774]	[1.944]
NC	0.937	0.857	0.782	0.715	0.609						0.834	0.850	0.824	0.731	0.628	1.121	1.104	1.210	1.372	1.469	1.059	1.109	1.117	1.042	0.861
STR	1.000	0.902	0.752	0.706	0.629						1.121	1.050	1.036	1.027	1.016	0.993	0.991	1.048	1.121	1.086	1.016	1.055	1.107	1.101	1.106
KG	0.993	0.826	0.752	0.632	0.534						1.006	1.076	1.434	1.727	1.911	0.964	0.914	0.958	1.218	1.202	1.024	1.042	1.020	1.020	0.938
ANN	0.920	0.858	0.734	0.675	0.610						1.022	1.317	1.646	1.701	1.629	1.015	1.045	1.054	1.155	1.149	1.004	1.016	1.003	1.001	0.978
EBF	0.977	0.905	0.847	0.771	0.681						0.929	0.900	0.881	0.805	0.723	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
EQ(All)	0.952	0.864	0.774	0.707	0.640						0.900	0.969	1.061	1.118	1.111	0.975	0.974	1.006	1.118	1.125	0.996	1.004	1.008	1.003	0.969
EQ(UA)	0.942	0.841	0.750	0.669	0.596						0.876	1.011	1.225	1.359	1.393	0.967	0.969	0.983	1.108	1.100	1.000	1.009	0.997	0.999	0.969
WAA(c_B)	0.952	0.863	0.770	0.697	0.622						0.900	0.969	1.051	1.070	1.002	0.975	0.974	1.005	1.111	1.116	0.996	1.004	1.008	1.003	0.969
WAA(c_L)	0.942	0.843	0.785	0.698	0.605						0.893	0.931	0.868	0.763	0.669	0.981	0.969	1.013	1.016	1.046	0.996	1.006	1.016	1.042	0.965
MSE	0.947	0.857	0.759	0.679	0.599						0.886	0.932	0.933	0.853	0.749	0.975	0.982	1.012	1.107	1.091	0.997	1.007	1.019	1.032	0.980
Last	0.956	0.889	0.828	0.829	0.601						0.952	1.076	0.918	0.806	0.660	1.096	1.021	1.027	1.274	1.168	0.994	1.011	1.020	1.041	1.023

Table 1.3: Relative Root Mean Square Forecast Error ratios of each model with the linear autoregression being the benchmark. The first row in each panel contains the Root Mean Square Forecast Error of the linear autoregression in parentheses. AR: Linear Autoregression, NC: No Change forecasts, STR: Smooth Transition Regression, KG: Kolmogorov-Gabor polynomial, ANN: Artificial Neural Network, EBF: Elliptic Basis Function network, EQ(All): Equal weighting of all individual models, EQ(UA): Equal weighting of the universal approximators, WAA(c_B): WAA with $c = c_B$, WAA(c_L): WAA with $c = c_L$, MSE: Weighting by inverse MSE, Last: Forecasting with the model which has the best previously realized forecast at the given horizon.

	Inflation					Industrial Production					Interest Rate					Money Supply					Unemployment				
Finland	1	3	6	12	24	1	3	6	12	24	1	3	6	12	24	1	3	6	12	24	1	3	6	12	24
AR	[0.320]	[0.544]	[0.794]	[1.277]	[1.329]	[3.489]	[4.751]	[5.154]	[5.176]	[6.058]											[0.100]	[0.224]	[0.344]	[0.446]	[0.945]
NC	0.945	0.869	0.792	0.748	1.324	0.964	1.032	1.262	1.535	1.278											0.894	0.911	0.940	1.272	1.309
STR	0.998	0.919	0.799	0.733	1.114	0.987	0.986	1.081	1.152	0.978											0.826	0.984	1.033	1.102	1.381
KG	1.014	0.946	0.815	0.734	1.095	1.001	0.991	1.015	1.045	1.115											0.901	0.900	0.865	0.899	0.846
ANN	1.000	1.022	1.016	1.036	1.300	0.956	0.966	1.026	1.095	1.015											1.004	1.064	1.187	1.444	1.494
EBF	1.000	1.000	1.000	1.000	1.000	0.950	0.939	0.976	1.027	1.002											1.000	1.000	1.000	1.000	1.000
EQ(All)	0.982	0.943	0.865	0.794	0.924	0.939	0.953	1.031	1.122	1.033											0.885	0.931	0.945	1.018	1.118
EQ(UA)	0.997	0.980	0.917	0.864	0.965	0.955	0.957	0.998	1.047	1.037											0.957	0.973	0.988	1.041	1.063
WAA(c_B)	0.982	0.942	0.861	0.795	1.161	0.939	0.954	1.029	1.102	1.033											0.885	0.931	0.944	1.018	1.117
WAA(c_L)	0.979	0.914	0.802	0.751	1.328	0.968	0.999	1.063	1.042	1.003											0.885	0.930	0.937	1.003	1.067
MSE	0.984	0.935	0.833	0.753	1.038	0.938	0.956	1.030	1.088	1.032											0.876	0.921	0.901	0.988	1.102
Last	1.015	0.944	0.887	0.707	1.138	0.948	1.024	1.208	1.096	1.185											0.880	1.029	0.887	0.928	1.067

	Inflation					Industrial Production					Interest Rate					Money Supply					Unemployment				
France	1	3	6	12	24	1	3	6	12	24	1	3	6	12	24	1	3	6	12	24	1	3	6	12	24
AR	[0.319]	[0.534]	[0.621]	[0.774]	[1.103]	[1.426]	[1.811]	[2.246]	[2.396]	[2.008]	[0.154]	[0.327]	[0.485]	[0.764]	[0.976]						[0.061]	[0.114]	[0.222]	[0.480]	[1.018]
NC	0.939	0.821	0.707	0.632	0.591	1.044	1.011	0.971	1.041	1.187	0.992	0.935	0.876	0.797	0.733						1.061	1.194	1.163	1.063	0.927
STR	0.930	0.810	0.693	0.663	0.654	1.010	0.987	0.986	0.959	0.887	0.993	0.994	0.986	0.965	0.937						1.034	1.035	1.009	0.924	0.901
KG	0.997	0.821	0.689	0.622	0.610	0.998	0.933	0.819	0.795	0.806	0.989	0.957	0.968	0.958	0.979						1.497	1.251	1.176	1.223	1.377
ANN	0.916	0.798	0.705	0.637	0.591	0.959	0.880	0.778	0.756	0.791	1.000	1.000	1.000	1.000	1.000						1.003	1.004	1.006	1.003	1.001
EBF	0.926	0.805	0.694	0.697	0.639	0.972	0.949	0.877	0.896	0.924	1.000	1.000	1.000	1.000	1.000						1.000	1.000	1.000	1.000	1.000
EQ(All)	0.936	0.825	0.726	0.685	0.656	0.967	0.907	0.840	0.816	0.809	0.991	0.975	0.965	0.945	0.921						1.014	1.008	1.007	1.000	1.018
EQ(UA)	0.930	0.797	0.683	0.643	0.610	0.958	0.901	0.805	0.794	0.811	0.993	0.980	0.984	0.981	0.982						1.063	1.033	1.036	1.057	1.113
WAA(c_B)	0.936	0.824	0.722	0.678	0.646	0.967	0.907	0.840	0.815	0.805	0.991	0.975	0.964	0.944	0.919						1.014	1.008	1.007	1.000	1.018
WAA(c_L)	0.932	0.813	0.689	0.659	0.618	0.974	0.920	0.841	0.789	0.759	0.991	0.973	0.947	0.876	0.779						1.012	1.008	1.018	1.016	1.036
MSE	0.936	0.826	0.708	0.683	0.645	0.968	0.916	0.844	0.816	0.791	0.991	0.972	0.955	0.917	0.854						1.011	1.007	1.012	1.008	1.040
Last	0.953	0.893	0.778	0.795	0.729	1.021	0.985	0.846	0.958	0.946	0.982	0.984	0.938	0.873	0.733						1.106	1.123	1.134	1.105	1.041

Table 1.4: Relative Root Mean Square Forecast Error ratios of each model with the linear autoregression being the benchmark. The first row in each panel contains the Root Mean Square Forecast Error of the linear autoregression in parentheses. AR: Linear Autoregression, NC: No Change forecasts, STR: Smooth Transition Regression, KG: Kolmogorov-Gabor polynomial, ANN: Artificial Neural Network, EBF: Elliptic Basis Function network, EQ(All): Equal weighting of all individual models, EQ(UA): Equal weighting of the universal approximators, WAA(c_B): WAA with $c = c_B$, WAA(c_L): WAA with $c = c_L$, MSE: Weighting by inverse MSE, Last: Forecasting with the model which has the best previously realized forecast at the given horizon.

	Inflation					Industrial Production					Interest Rate					Money Supply					Unemployment				
Germany	1	3	6	12	24	1	3	6	12	24	1	3	6	12	24	1	3	6	12	24	1	3	6	12	24
AR	[0.316]	[0.426]	[0.536]	[0.714]	[0.534]	[1.464]	[1.976]	[2.837]	[2.512]	[2.836]	[0.155]	[0.342]	[0.543]	[0.913]	[1.292]	[2.255]	[3.323]	[2.544]	[3.160]	[3.141]	[0.122]	[0.301]	[0.521]	[0.967]	[1.896]
NC	0.976	0.980	0.950	0.912	1.406	1.086	1.069	1.126	1.312	1.241	0.988	0.892	0.785	0.664	0.539	1.039	1.119	1.331	1.539	1.499	1.025	1.017	0.992	0.987	0.965
STR	1.011	1.021	1.015	0.987	1.187	1.015	1.043	1.055	1.379	1.118	0.982	0.936	0.861	0.755	0.638	1.035	1.100	1.113	1.141	1.001	0.943	0.997	1.015	1.022	0.982
KG	1.068	1.046	1.044	1.010	1.168	1.044	1.006	1.004	1.041	0.930	0.985	0.897	0.848	0.790	0.763	1.031	1.074	1.073	1.052	0.966	1.173	1.003	0.937	0.946	0.861
ANN	1.001	1.001	1.001	1.000	1.000	1.030	1.019	1.017	1.016	0.974	1.000	1.000	1.000	1.000	1.000	1.043	1.098	1.080	1.064	1.045	0.995	0.971	0.955	0.962	0.953
EBF	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.975	1.011	1.047	1.033	1.033
EQ(All)	1.005	1.002	0.991	0.965	1.035	0.997	0.994	1.001	1.063	1.008	0.968	0.933	0.889	0.831	0.736	1.010	1.037	1.048	1.118	1.052	0.978	0.959	0.956	0.969	0.954
EQ(UA)	1.019	1.011	1.007	0.988	0.990	1.013	0.999	1.000	1.010	0.961	0.981	0.950	0.934	0.910	0.860	1.012	1.033	1.008	1.032	0.999	1.006	0.950	0.941	0.954	0.935
WAA(c_B)	1.005	1.002	0.991	0.966	1.037	0.997	0.994	1.001	1.065	1.009	0.968	0.933	0.888	0.827	0.730	1.010	1.038	1.037	1.076	1.032	0.978	0.959	0.956	0.969	0.957
WAA(c_L)	1.003	1.001	0.997	1.010	1.334	1.001	1.035	1.077	1.354	1.001	0.968	0.926	0.829	0.697	0.564	1.015	1.068	1.000	1.000	1.000	0.979	0.963	0.963	0.985	0.988
MSE	1.005	1.001	0.992	0.974	1.103	0.996	0.991	0.992	1.073	0.987	0.963	0.925	0.865	0.767	0.626	1.012	1.026	1.030	1.065	1.008	0.981	0.963	0.959	0.972	0.964
Last	1.008	1.049	0.999	1.024	1.347	1.029	1.038	0.999	1.194	1.181	1.017	0.982	0.880	0.761	0.572	1.057	1.077	1.155	1.185	1.142	0.946	1.088	1.039	0.946	1.042

	Inflation					Industrial Production					Interest Rate					Money Supply					Unemployment				
Italy	1	3	6	12	24	1	3	6	12	24	1	3	6	12	24	1	3	6	12	24	1	3	6	12	24
AR	[0.160]	[0.281]	[0.545]	[1.010]	[1.874]	[1.261]	[2.118]	[3.026]	[2.817]	[3.6039	[0.147]	[0.321]	[0.502]	[0.806]	[0.992]										
NC	0.810	0.725	0.544	0.406	0.286	0.923	0.943	0.947	1.054	0.792	1.014	0.945	0.874	0.792	0.721										
STR	0.994	0.822	0.627	0.425	0.297	1.164	1.217	1.222	1.555	0.791	0.997	0.998	0.986	0.945	0.846										
KG	0.944	0.792	0.642	0.508	0.497	1.058	0.977	0.953	0.918	0.877	1.025	1.017	0.970	0.954	0.982										
ANN	0.949	0.863	0.693	0.491	0.421	1.026	0.950	0.956	0.921	0.831	0.996	0.993	0.993	1.007	0.997										
EBF	0.985	0.827	0.630	0.474	0.411	1.057	1.028	0.988	0.906	0.877	1.000	1.000	1.000	1.000	1.000										
EQ(All)	0.930	0.791	0.642	0.507	0.459	0.988	0.963	0.946	0.930	0.791	0.989	0.980	0.956	0.930	0.862										
EQ(UA)	0.948	0.797	0.628	0.471	0.435	1.033	0.977	0.958	0.896	0.848	1.001	0.996	0.976	0.963	0.905										
WAA(c_B)	0.930	0.791	0.641	0.499	0.437	0.988	0.962	0.947	0.939	0.788	0.989	0.980	0.956	0.930	0.862										
WAA(c_L)	0.927	0.780	0.608	0.440	0.301	0.964	0.996	1.154	1.417	0.768	0.989	0.980	0.952	0.914	0.808										
MSE	0.926	0.785	0.620	0.455	0.319	0.976	0.964	0.953	0.949	0.784	0.985	0.975	0.942	0.896	0.811										
Last	0.881	0.873	0.810	0.488	0.356	0.983	0.978	1.015	1.094	0.774	1.035	1.016	0.949	0.903	0.873										

Table 1.5: Relative Root Mean Square Forecast Error ratios of each model with the linear autoregression being the benchmark. The first row in each panel contains the Root Mean Square Forecast Error of the linear autoregression in parentheses. AR: Linear Autoregression, NC: No Change forecasts, STR: Smooth Transition Regression, KG: Kolmogorov-Gabor polynomial, ANN: Artificial Neural Network, EBF: Elliptic Basis Function network, EQ(All): Equal weighting of all individual models, EQ(UA): Equal weighting of the universal approximators, WAA(c_B): WAA with $c = c_B$, WAA(c_L): WAA with $c = c_L$, MSE: Weighting by inverse MSE, Last: Forecasting with the model which has the best previously realized forecast at the given horizon.

	Inflation					Industrial Production					Interest Rate					Money Supply					Unemployment				
Japan	1	3	6	12	24	1	3	6	12	24	1	3	6	12	24	1	3	6	12	24	1	3	6	12	24
AR	[0.257]	[0.443]	[0.601]	[0.794]	[1.161]	[1.615]	[2.870]	[4.807]	[4.804]	[3.7869	[0.132]	[0.243]	[0.337]	[0.466]	[0.3309	[2.057]	[4.723]	[7.800]	[8.905]	[7.355]	[0.129]	[0.197]	[0.295]	[0.530]	[1.083]
NC	0.969	0.912	0.837	0.812	0.551	1.064	1.180	1.186	1.422	0.808	0.985	0.991	0.982	0.973	1.117	0.996	1.032	1.064	1.207	0.692	1.009	0.928	0.873	0.706	0.629
STR	1.003	0.937	0.823	0.830	0.642	1.134	1.224	1.208	1.301	0.630	1.002	0.991	1.010	1.005	1.415	1.173	0.912	0.887	1.077	0.937	1.127	1.117	1.104	0.987	0.966
KG	1.026	0.963	1.057	0.869	0.543	1.011	1.006	1.004	1.004	0.966	1.118	1.161	0.998	0.961	0.597	1.438	1.126	0.954	0.941	1.244	1.112	1.110	1.063	0.871	0.777
ANN	1.051	1.026	0.938	0.734	0.532	1.030	1.019	1.038	1.191	0.670	1.462	1.306	1.133	1.173	1.000	1.013	1.011	1.007	1.054	1.000	0.998	1.002	0.983	0.928	0.900
EBF	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.007	1.020	1.036	1.123	0.938	1.000	1.002	1.004	1.006	1.011	1.009	1.004	0.999	0.980	0.955
EQ(All)	0.971	0.910	0.853	0.724	0.478	1.012	1.032	1.034	1.075	0.750	1.014	0.939	0.955	0.960	0.733	0.987	0.946	0.957	1.025	0.956	0.995	0.989	0.961	0.875	0.856
EQ(UA)	0.976	0.949	0.945	0.779	0.549	1.010	1.004	1.007	1.028	0.839	1.083	0.969	0.971	1.017	0.608	1.045	0.977	0.965	0.988	1.080	0.998	1.001	0.974	0.888	0.861
WAA(c_B)	0.971	0.910	0.853	0.719	0.465	1.012	1.033	1.038	1.093	0.725	1.014	0.939	0.956	0.961	0.742	0.986	0.953	0.964	1.039	1.039	0.995	0.989	0.961	0.877	0.858
WAA(c_L)	0.976	0.924	0.864	0.826	0.568	1.014	1.028	1.044	1.359	0.666	1.011	0.962	0.986	1.017	1.146	1.143	1.101	1.083	1.061	1.056	0.994	0.982	0.958	0.918	1.011
MSE	0.979	0.919	0.806	0.748	0.512	1.012	1.031	1.035	1.090	0.716	1.005	0.971	0.968	0.971	0.996	0.985	0.952	0.960	1.026	0.995	0.995	0.988	0.954	0.904	0.941
Last	1.032	0.994	0.894	0.989	0.683	1.054	1.042	1.061	1.370	0.752	1.106	1.114	1.009	1.040	1.268	1.083	0.980	0.995	1.108	0.774	1.051	1.037	0.948	0.829	0.983

	Inflation					Industrial Production					Interest Rate					Money Supply					Unemployment				
Norway	1	3	6	12	24	1	3	6	12	24	1	3	6	12	24	1	3	6	12	24	1	3	6	12	24
AR	[0.621]	[1.283]	[1.646]	[2.132]	[1.813]	[4.569]	[5.003]	[5.962]	[6.788]	[6.501]	[0.248]	[0.491]	[0.684]	[1.165]	[1.076]						[0.109]	[0.231]	[0.408]	[0.758]	[1.283]
NC	1.000	0.978	0.960	0.939	0.934	1.073	1.136	1.075	1.233	0.860	0.967	0.950	0.973	0.957	1.036						1.005	1.006	1.002	1.000	0.997
STR	1.006	1.011	0.983	0.957	0.949	1.023	1.170	1.256	1.397	0.881	1.044	1.069	1.068	1.056	1.275						0.988	0.986	0.980	0.976	0.964
KG	1.034	1.016	1.004	0.993	1.002	1.051	1.009	1.012	1.038	0.956	1.161	1.079	1.120	1.074	1.427						0.985	0.994	0.998	0.975	0.830
ANN	1.022	1.004	0.972	0.896	0.928	1.192	1.018	0.991	1.051	0.881	1.360	1.161	1.185	1.065	1.097						0.992	0.985	0.972	0.974	0.948
EBF	1.000	1.000	1.000	1.000	1.000	0.996	0.971	0.972	0.992	0.912	1.028	1.038	1.045	1.032	1.068						1.012	1.019	1.018	1.008	1.003
EQ(All)	1.001	0.993	0.973	0.942	0.882	0.971	0.978	0.979	1.058	0.817	1.025	1.009	1.030	1.008	1.103						0.989	0.989	0.986	0.981	0.952
EQ(UA)	1.012	1.001	0.986	0.955	0.944	1.051	0.985	0.983	1.020	0.904	1.088	1.044	1.075	1.033	1.162						0.984	0.987	0.983	0.975	0.921
WAA(c_B)	1.001	0.993	0.976	0.930	0.909	0.970	0.977	0.978	1.034	0.789	1.025	1.009	1.030	1.008	1.110						0.989	0.989	0.986	0.982	0.955
WAA(c_L)	1.006	1.000	0.995	0.903	0.967	1.014	1.011	1.031	1.000	0.860	1.021	1.001	1.032	1.049	1.240						0.989	0.992	0.997	1.004	1.001
MSE	1.001	0.994	0.975	0.940	0.891	0.969	0.979	0.985	1.035	0.769	1.022	1.007	1.029	1.015	1.114						0.989	0.990	0.986	0.984	0.967
Last	0.985	1.037	0.980	1.017	0.863	1.069	1.057	1.225	1.112	0.872	1.034	1.080	0.997	1.048	1.268						0.975	0.949	0.924	0.955	0.966

Table 1.6: Relative Root Mean Square Forecast Error ratios of each model with the linear autoregression being the benchmark. The first row in each panel contains the Root Mean Square Forecast Error of the linear autoregression in parentheses. AR: Linear Autoregression, NC: No Change forecasts, STR: Smooth Transition Regression, KG: Kolmogorov-Gabor polynomial, ANN: Artificial Neural Network, EBF: Elliptic Basis Function network, EQ(All): Equal weighting of all individual models, EQ(UA): Equal weighting of the universal approximators, WAA(c_B): WAA with $c = c_B$, WAA(c_L): WAA with $c = c_L$, MSE: Weighting by inverse MSE, Last: Forecasting with the model which has the best previously realized forecast at the given horizon.

	Inflation					Industrial Production					Interest Rate					Money Supply					Unemployment				
Sweden	1	3	6	12	24	1	3	6	12	24	1	3	6	12	24	1	3	6	12	24	1	3	6	12	24
AR	[0.356]	[0.617]	[0.887]	[1.369]	[1.656]	[2.588]	[2.435]	[2.437]	[2.778]	[4.086]	[0.165]	[0.351]	[0.560]	[0.9449]	[1.394]	[1.818]	[2.379]	[2.998]	[3.650]	[4.009]	[0.215]	[0.349]	[0.541]	[0.895]	[1.323]
NC	0.973	0.904	0.816	0.797	1.048	1.189	1.196	1.325	1.525	1.270	1.006	0.919	0.866	0.770	0.660	1.028	1.062	1.063	1.252	1.307	1.034	0.989	1.004	0.992	0.992
STR	1.007	0.963	0.869	0.842	1.080	1.003	1.005	1.050	1.114	1.071	0.996	0.989	0.971	0.952	0.896	1.022	1.053	1.084	1.299	1.047	1.067	1.167	1.310	1.408	1.093
KG	0.992	0.923	0.848	0.811	0.941	1.037	1.037	1.007	1.018	1.082	1.039	1.024	0.984	0.969	1.042	1.056	1.008	0.988	1.045	1.090	1.206	1.263	1.175	1.251	1.057
ANN	0.995	0.940	0.852	0.826	1.015	1.047	1.038	0.971	1.045	1.085	1.000	1.000	1.000	1.000	1.000	1.042	1.113	1.000	1.016	1.017	1.002	1.004	1.000	1.004	0.985
EBF	1.000	1.000	1.000	1.000	1.000	1.027	1.021	1.049	1.128	1.085	0.984	0.972	0.977	0.966	0.973	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
EQ(All)	0.990	0.943	0.874	0.840	0.934	1.032	1.022	1.033	1.120	1.083	0.989	0.967	0.949	0.926	0.894	1.011	1.017	0.990	1.059	1.031	0.999	0.995	0.997	0.990	0.924
EQ(UA)	0.993	0.945	0.881	0.848	0.927	1.028	1.023	0.994	1.054	1.082	0.991	0.982	0.971	0.964	0.963	1.016	1.021	0.977	1.003	1.023	1.018	1.035	1.025	1.033	0.928
WAA(c_B)	0.990	0.943	0.874	0.841	0.954	1.032	1.022	1.032	1.111	1.080	0.989	0.967	0.949	0.926	0.891	1.011	1.017	0.990	1.059	1.032	0.999	0.995	0.997	0.986	0.924
WAA(c_L)	0.990	0.944	0.872	0.824	1.110	1.018	1.013	1.004	1.022	1.052	0.989	0.967	0.940	0.886	0.707	1.011	1.017	0.996	1.072	1.060	0.999	0.987	0.982	0.989	1.090
MSE	0.988	0.944	0.874	0.820	0.994	1.030	1.020	1.026	1.076	1.067	0.990	0.967	0.941	0.923	0.773	1.011	1.018	0.993	1.060	1.068	1.001	1.000	1.007	1.002	0.887
Last	0.991	0.943	0.872	0.802	1.058	1.148	1.128	1.229	1.132	1.207	1.073	0.991	0.930	0.870	0.660	1.013	1.108	1.107	1.252	1.193	1.092	1.089	1.224	1.233	0.991

	Inflation					Industrial Production					Interest Rate					Money Supply					Unemployment				
UK	1	3	6	12	24	1	3	6	12	24	1	3	6	12	24	1	3	6	12	24	1	3	6	12	24
AR	[0.264]	[0.502]	[0.694]	[1.060]	[1.347]	[1.290]	[1.576]	[1.953]	[2.004]	[1.806]	[0.147]	[0.329]	[0.450]	[0.706]	[0.841]						[0.045]	[0.072]	[0.121]	[0.263]	[0.656]
NC	0.892	0.779	0.655	0.522	0.571	1.063	0.995	0.934	1.159	0.964	1.021	0.935	0.863	0.772	0.617						1.011	1.052	1.016	0.749	0.464
STR	0.980	0.904	0.776	0.535	0.575	1.025	1.104	1.089	1.219	0.797	0.987	0.959	0.914	0.856	0.713						0.973	0.938	0.900	0.700	0.423
KG	0.955	0.823	0.703	0.593	0.579	1.020	0.926	0.879	0.896	0.777	1.054	1.122	1.226	1.286	1.489						1.062	1.029	0.995	0.804	0.783
ANN	0.974	0.799	0.731	0.604	0.537	1.106	1.078	1.043	1.026	0.854	1.000	0.998	0.997	1.002	1.042						1.015	1.097	1.189	1.081	0.902
EBF	0.995	0.992	1.005	1.034	0.887	1.038	0.932	0.891	0.917	0.950	1.014	1.020	1.034	1.021	1.012						1.051	1.111	1.141	0.991	0.782
EQ(All)	0.941	0.843	0.730	0.584	0.604	1.007	0.948	0.913	0.948	0.788	0.993	0.988	0.975	0.947	0.905						0.988	0.940	0.892	0.741	0.637
EQ(UA)	0.949	0.837	0.745	0.642	0.603	1.027	0.947	0.921	0.929	0.843	1.009	1.035	1.070	1.083	1.158						1.022	1.015	0.994	0.828	0.719
WAA(c_B)	0.941	0.843	0.729	0.584	0.590	1.008	0.948	0.912	0.947	0.785	0.993	0.988	0.974	0.945	0.888						0.988	0.940	0.892	0.741	0.635
WAA(c_L)	0.941	0.831	0.699	0.620	0.567	1.021	0.971	0.999	1.038	0.772	0.993	0.982	0.933	0.862	0.625						0.988	0.941	0.894	0.743	0.564
MSE	0.942	0.839	0.706	0.628	0.568	1.009	0.948	0.913	0.951	0.776	0.993	0.983	0.949	0.864	0.701						0.989	0.948	0.909	0.732	0.515
Last	1.026	0.919	0.771	0.768	0.600	1.084	1.051	1.036	1.203	0.898	1.062	1.082	1.038	0.932	0.653						1.054	1.017	0.968	0.899	0.644

Table 1.7: Relative Root Mean Square Forecast Error ratios of each model with the linear autoregression being the benchmark. The first row in each panel contains the Root Mean Square Forecast Error of the linear autoregression in parentheses. AR: Linear Autoregression, NC: No Change forecasts, STR: Smooth Transition Regression, KG: Kolmogorov-Gabor polynomial, ANN: Artificial Neural Network, EBF: Elliptic Basis Function network, EQ(All): Equal weighting of all individual models, EQ(UA): Equal weighting of the universal approximators, WAA(c_B): WAA with $c = c_B$, WAA(c_L): WAA with $c = c_L$, MSE: Weighting by inverse MSE, Last: Forecasting with the model which has the best previously realized forecast at the given horizon.

	Inflation					Industrial Production					Interest Rate					Money Supply					Unemployment				
US	1	3	6	12	24	1	3	6	12	24	1	3	6	12	24	1	3	6	12	24	1	3	6	12	24
AR	[0.406]	[0.709]	[0.912]	[1.209]	[1.030]	[0.729]	[1.439]	[2.294]	[1.722]	[1.417]	[0.215]	[0.425]	[0.524]	[0.727]	[0.612]	[1.348]	[2.031]	[2.544]	[3.179]	[4.996]	[0.117]	[0.218]	[0.346]	[0.500]	[0.846]
NC	0.923	0.931	0.981	1.007	1.223	1.026	1.077	1.056	1.782	1.300	1.029	0.979	0.942	0.898	0.867	0.948	1.006	1.009	0.984	1.081	1.064	1.144	1.224	1.037	0.855
STR	1.002	0.976	1.003	1.030	1.247	0.999	0.940	0.863	1.315	1.114	1.004	0.992	0.962	0.901	0.787	1.065	1.039	1.025	1.016	1.035	1.019	1.069	1.149	1.139	0.921
KG	1.076	0.995	0.969	0.971	1.120	0.968	0.968	0.948	1.520	1.131	1.136	1.033	1.042	1.046	1.038	1.372	1.042	1.041	0.997	1.057	0.999	1.060	1.109	0.941	0.840
ANN	0.996	0.965	0.958	0.948	0.950	0.964	0.929	0.895	1.365	0.984	1.004	1.001	1.000	1.000	0.994	1.062	1.057	1.033	1.005	1.023	1.000	1.000	1.000	1.000	1.000
EBF	0.973	0.929	0.980	0.931	1.000	0.985	0.923	0.890	1.060	0.949	1.000	1.000	1.000	1.000	1.000	1.035	1.034	1.029	1.008	1.078	1.004	1.009	1.011	1.000	1.000
EQ(All)	0.986	0.958	0.974	0.971	1.062	0.970	0.933	0.885	1.195	1.002	0.999	0.993	0.977	0.953	0.866	1.020	1.007	1.006	0.986	1.035	0.990	1.006	1.036	0.964	0.930
EQ(UA)	1.004	0.955	0.964	0.945	1.012	0.966	0.922	0.887	1.295	1.001	1.017	1.006	1.008	1.009	1.007	1.097	1.027	1.019	0.993	1.050	0.989	1.007	1.025	0.967	0.943
WAA(c_B)	0.985	0.958	0.974	0.972	1.069	0.970	0.934	0.892	1.172	1.004	0.999	0.993	0.977	0.952	0.847	1.019	1.007	1.006	0.987	1.035	0.991	1.006	1.036	0.967	0.938
WAA(c_L)	0.963	0.945	0.981	0.983	1.124	0.975	0.956	0.995	1.044	0.917	0.999	0.992	0.967	0.924	0.882	1.011	1.016	1.010	1.013	1.011	0.992	1.024	1.074	1.022	1.000
MSE	0.982	0.957	0.973	0.973	1.079	0.971	0.931	0.902	1.167	0.987	0.993	0.993	0.975	0.945	0.835	1.017	1.006	1.006	0.987	1.035	0.994	1.021	1.060	0.963	0.958
Last	0.988	0.948	1.009	1.019	1.188	0.978	0.895	1.017	1.557	1.001	1.152	1.044	0.999	0.973	0.949	1.013	1.032	1.028	0.988	1.092	1.029	1.054	1.091	0.917	0.924

Table 1.8: Relative Root Mean Square Forecast Error ratios of each model with the linear autoregression being the benchmark. The first row in each panel contains the Root Mean Square Forecast Error of the linear autoregression in parentheses. AR: Linear Autoregression, NC: No Change forecasts, STR: Smooth Transition Regression, KG: Kolmogorov-Gabor polynomial, ANN: Artificial Neural Network, EBF: Elliptic Basis Function network, EQ(All): Equal weighting of all individual models, EQ(UA): Equal weighting of the universal approximators, WAA(c_B): WAA with $c = c_B$, WAA(c_L): WAA with $c = c_L$, MSE: Weighting by inverse MSE, Last: Forecasting with the model which has the best previously realized forecast at the given horizon.

1.8 Bibliography

Bates, J. and C. Granger (1969). Combination of forecasts . *Operations Research Quarterly 20*, 451–468.

Cesa-Bianchi, N. and G. Lugosi (2006). *Prediction, learning, and games*. Cambridge University Press.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems (MCSS) 2*, 303–314.

Freund, Y. and R. E. Schapire (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences 55*, 119–139.

Granger, C. W. J. and T. Teräsvirta (1993). *Modelling Nonlinear Economic Relationships*. Oxford University Press, Oxford.

Hornik, K., M. Stinchcombe, and H. White (1989). Multilayer feedforward networks are universal approximators. *Neural Networks 2*, 359–366.

Kallenberg, O. (1997). *Foundations of Modern Probability*. Springer, New York.

Kivinen, J. and M. K. Warmuth (1999). Averaging Expert Predictions. *In Paul Fischer and Hans Ulrich Simon, editors, Proceedings of the 4th European Conference on Computational Learning Theory EuroCOLT '99*, 153–167.

Kock, A. B. and T. Teräsvirta (2011). Forecasting macroeconomic variables using neural network models and three automated model selection techniques. *CREATES Research Paper 2011-27, Aarhus University*.

Littlestone, N. and M. Warmuth (1994). The Weighted Majority Algorithm. *Information and Computation 108*, 212–261.

Park, J. and I. W. Sandberg (1991). Universal approximation using radial-basis-function networks. *Neural Computation 3*, 246–257.

Park, J. and I. W. Sandberg (1993). Approximation and radial-basis-function networks. *Neural Computation 5*, 305–316.

Park, J. and I. W. Sandberg (1994). Nonlinear approximations using elliptic basis function networks. *Circuits, Systems, and Signal Processing 13*, 99–113.

Rudin, W. (1987). *Real and Complex Analysis*. McGraw-Hill Book Company.

Stock, J. H. and M. W. Watson (1999). *A comparison of linear and nonlinear univariate models for forecasting macroeconomic time series, in R.F. Engle and H. White (eds)*. Oxford: Oxford University Press.

Swanson, N. and H. White (1995). A model-selection approach to assessing the information in the term structure using linear models and artificial neural networks. *Journal of Business & Economic Statistics 13*, 265–275.

Teräsvirta, T., C. W. J. Granger, and D. Tjøstheim (2010). *Modelling Nonlinear Economic Time Series*. Oxford University Press, Oxford.

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

Timmermann, A. (2006). *Forecast Combination, in G. Elliott, C.W.J. Granger, A Timmermann (eds)*, Volume 1. Elsevier, Amsterdam.

White, H. (2006). Approximate nonlinear forecasting methods, in G. Elliott, C.W.J. Granger, A Timmermann (eds). *Handbook of Economic Forecasting 1*, 459–512.

Chapter 2

Forecasting Macroeconomic Variables using Neural Network Models and Three Automated Model Selection Techniques

Anders Bredahl Kock
Aarhus University and CREATES
Timo Teräsvirta
Aarhus University and CREATES

Financial support from CREATES, funded by the Danish National Research Foundation, is gratefully acknowledged. Part of this work was carried out when the first author was visiting the Department of Economics at the University of California, Berkeley, and the second author the Department of Economics at the European University Institute, Florence. We are thankful for the kind hospitality of these institutions during our visits. Material from this paper has been presented at the workshop in Econometric Aspects of Price Transmission Analysis, Georg-August University of Göttingen, August 2010, the 19th Symposium of the Society of Nonlinear Dynamics and Econometrics, Washington DC, March 2011, the 31st International Annual Symposium in Forecasting, Prague, June 2011, and seminars at Banque de France and the European University Institute, Florence. We thank participants of these occasions for their comments. The first author would like to thank his PhD committee consisting of Henning Bunzel, Dick van Dijk and Jurgen Doornik for their careful reading and constructive comments. The authors are solely responsible for any errors and shortcomings in this work.

Abstract

In this paper we consider the forecasting performance of a well-defined class of flexible models, the so-called single hidden-layer feedforward neural network models. A major aim of our study is to find out whether they, due to their flexibility, are as useful tools in economic forecasting as some previous studies have indicated. When forecasting with neural network models one faces several problems, all of which influence the accuracy of the forecasts. First, neural networks are often hard to estimate due to their highly nonlinear structure. In fact, their parameters are not even globally identified. Recently, White (2006) presented a solution that amounts to converting the specification and nonlinear estimation problem into a linear model selection and estimation problem. He called this procedure the QuickNet and we shall compare its performance to two other procedures which are built on the linearisation idea: the Marginal Bridge Estimator and Autometrics. Second, one must decide whether forecasting should be carried out recursively or directly. Comparisons of these two methods exist for linear models and here these comparisons are extended to neural networks.

Finally, a nonlinear model such as the neural network model is not appropriate if the data is generated by a linear mechanism. Hence, it might be appropriate to test the null of linearity prior to building a nonlinear model. We investigate whether this kind of pretesting improves the forecast accuracy compared to the case where this is not done.

2.1 Introduction

Artificial Neural Networks (ANN) have been quite popular in many areas of science for describing various phenomena and forecasting them. They have also been used in forecasting macroeconomic time series and financial series, see Kuan and Liu (1995) for a successful example on exchange rate forecasting, and Zhang et al. (1998) and Rech (2002) for more mixed results. The main argument in their favour is that ANNs are universal approximators, which means that they are capable of approximating arbitrarily accurately functions satisfying only mild regularity conditions. The ANN models thus have a strong nonparametric flavour. One may therefore expect them to be a versatile tool in economic forecasting and adapt quickly to rapidly changing forecasting situations. Recently, Ahmed et al. (2010) conducted an extensive forecasting study comprising more than 1000 economic time series from the M3 competition Makridakis and Hibon

(2000), and a large number of what they called machine learning tools. They concluded that the ANN model that we are going to consider, the single hidden-layer feedforward ANN model or multi-layer perceptron with one hidden layer, was one of the best or even the best performer in their study. A single hidden-layer ANN model is already a universal approximator; see Cybenko (1989) and Hornik et al. (1989).

A major problem in the application of ANN models is the specification and estimation of these models. A large number of modelling strategies have been developed for the purpose. It is possible to begin with a small model and increase its size ("specific-to-general", "bottom up", or "growing the network"). Conversely, one can specify a network with a large number of variables and hidden units or "neurons" and then reduce its size ("general-to-specific", "top down" or "pruning the network"). Since the ANN model is nonlinear in parameters, its parameters have to be estimated numerically, which may be a demanding task if the number of parameters in the model is large. Recently, White (2006) devised a clever strategy for modelling ANNs that converts the specification and ensuing nonlinear estimation problem into a linear model selection problem. This greatly simplifies the estimation stage and alleviates the computational effort. It is therefore of interest to investigate how well this strategy, called QuickNet, performs in macroeconomic forecasting. A natural benchmark in that case is a linear autoregressive model.

Quite often, application of White's strategy leads to a situation in which the number of variables in the set of candidate variables exceeds the number of observations. The strategy handles these cases without problems, because it essentially works from specific to general and then back again. We shall also consider a one-way variant from specific to general in this study. One may want to set a maximum limit for variables to be included in the model to control its size.

There exist other modelling strategies that can also be applied to selecting the variables. In fact, White (2006) encouraged comparisons between his method and other alternatives, and here we shall follow his suggestion. In this work, we consider two additional specification techniques. One is Autometrics by Doornik (2009), see also Krolzig and Hendry (2001) and Hendry and Krolzig (2005), and the other one is the Marginal Bridge Estimator (MBE), see Huang et al. (2008). The former is designed for econometric modelling, whereas the latter one has its origins in statistics. Autometrics works from general to specific, and the same may be said about MBE. We shall compare the performance of these

three methods when applying White's idea of converting the specification and estimation problem into a linear model selection problem and selecting hidden units for our ANN models. That is one of the main objectives of this paper.

The focus in this study is on multiperiod forecasting. There are two ways of generating multiperiod forecasts. One consists of building a single model and generating the forecasts for more than one period ahead recursively. The other one, called direct forecasting, implies that a separate model is built for each forecasting horizons, and no recursions are involved. For discussion, see for example Teräsvirta (2006), Teräsvirta et al. (2010, Chapter 14), or Kock and Teräsvirta (2011). In nonlinear forecasting, the latter method appears to be more common, see for example Stock and Watson (1999) and Marcellino (2002), whereas Teräsvirta et al. (2005) constitutes an example of the former alternative. A systematic comparison of the performance of the two methods exists, see Marcellino et al. (2006), but it is restricted to linear autoregressive models. Our aim is to extend these comparisons to nonlinear ANN models.

Nonlinear models can sometimes generate obviously insane forecasts. One way of alleviating this problem is to use insanity filters as in Swanson and White (1995, 1997a,b) who discuss this issue. We will compare two filters to the unfiltered forecasts and see how they impact on the forecasting performance of the neural networks.

In this work the ANN models are augmented by including lags of the variable to be forecast in them. As a result, the augmented models nest a linear autoregressive model. It is well known that if the data-generating process is linear, the augmented ANN model is not even locally identified; see for example Lee et al. (1993), Teräsvirta et al. (1993) or Teräsvirta et al. (2010, Chapter 5) for discussion. A general discussion of identification problems in ANN models can be found in Hwang and Ding (1997). It may then be advisable to first test linearity of each series under consideration before applying any ANN modelling strategy to it. But then, it may also be argued that linearity tests are unnecessary, because the set of candidate variables can be (and in our case is) defined to include both linear lags and hidden units. The modelling technique can then choose among all of them and find the combination that is superior to the others. We shall compare these two arguments. This is done by carrying out pretesting and only fitting an ANN model to the series if linearity is rejected. Forecasts are generated from models specified this way and compared with forecasts from the ANN models obtained using White's method and the three automatic modelling techniques.

The main criterion of comparing forecasts is the Root Mean Square Forecast Error (RMSFE), which implies a quadratic loss function. Other alternatives are possible, but the RMSFE is commonly used and thus even applied here. We rank the methods, which makes some comparisons possible. Furthermore, we also carry out Wilcoxon signed rank tests but principally for descriptive purposes, so the tests are not used as an ex post model selection criterion; see Costantini and Kunst (2011) for a discussion.

It might be desirable to compare White's method with modelling strategies which are not based on linearising the problem but in which statistical methods such as hypothesis testing and nonlinear maximum likelihood estimation are applied. Examples of these include Swanson and White (1995, 1997a,b), Anders and Korn (1999) and Medeiros et al. (2006). These approaches do, however, require plenty of human resources, unless the number of time series under consideration and forecasts generated from them are small. This is because nonlinear iterative estimation is hard to automate. Each estimation needs a nonnegligible amount of tender loving care, and when the number of time series to be considered is large, ANN model building and forecasting tend to require a substantial amount of resources.

In this paper we investigate the forecasting performance of the above techniques. We first conduct a small simulation study to see how well these techniques perform when the data are generated by a known nonlinear model. The economic data sets consist of the monthly unemployment and consumer price index series from the 1960's until 2009.

The plan of the paper is as follows. The neural network model is presented in Section 2.2 and estimation techniques in Section 2.3. The recursive and direct forecasting methods are discussed in Section 2.4 and the results are summarized in Section 2.5, while Section 2.6 concludes.

2.2 The Model

We begin by briefly introducing the Artificial Neural Network (ANN) model and reviewing some of its properties. The techniques for specifying the structure of the model and estimating the parameters will be considered in the next section. Our model is the so-called single-hidden-layer feedforward autoregressive neu-

ral network model or single-hidden-layer perceptron

$$y_t = \beta_0' \mathbf{z}_t + \sum_{j=1}^{q} \beta_j (1 + \exp\{\gamma_j' \mathbf{z}_t\})^{-1} + \varepsilon_t \tag{2.1}$$

where $\mathbf{z}_t = (1, y_{t-1}, ..., y_{t-p})'$, $\beta_0 = (\beta_{00}, \beta_{01}, ..., \beta_{0p})'$, $\gamma_j = (\gamma_{j0}, \gamma_{j1}, ..., \gamma_{jp})$ and $\varepsilon_t \sim$ iid$\mathcal{N}(0, \sigma^2)$. The weak stationarity condition of (2.1) is the same as that of the corresponding linear AR(p) model. The ANN model is a so-called universal approximator in the following sense. Suppose there is a functional relationship between y and $\mathbf{z}$: $y = H(\mathbf{z})$. Then under appropriate regularity conditions for any $\delta > 0$ there exists a positive integer $q < \infty$ such that $\left\| H(\mathbf{z}) - \sum_{j=1}^{q} \beta_j (1 + \exp\{\gamma_j' \mathbf{z}\})^{-1} \right\| < \delta$ where $\|\cdot\|$ is an appropriate norm. This indicates that (2.1) is a very flexible functional form and thus in principle capable of satisfactorily approximating various nonlinear processes.

Before forecasting with the model (2.1), the number of logistic functions or hidden units q has to be specified and its parameters estimated. Various specification techniques have been proposed in the literature. One possibility is to begin with a large model (large q) and reduce the size of the model, that is, to prune the network. Another possibility is to begin with a small model and add hidden units, which is called 'growing the network'. Either way, one also has to estimate the parameters of the model which, given that it is heavily nonlinear, may be numerically demanding, in particular when q is large. For discussion, see for example Fine (1999, Chapter 6), Goffe, Ferrier, and Rogers (1994), or Simon (1999).

Nevertheless, if the parameter vectors γ_j, $j = 1, ..., q$, are known, the model is linear in parameters. This opens up the possibility to combine specification and estimation into a single linear model selection problem. White (2006) suggested this technique for specifying and estimating artificial neural network models. The linear model selection problem encountered is the one of choosing a subset of variables from the set

$$S = \{y_{t-i}, i = 1, ..., p; (1 + \exp\{\gamma_j' \mathbf{z}_t\})^{-1}, j = 1, ..., M\} \tag{2.2}$$

where M is large. Since the quality of the estimates depends on the size of S, the number of variables in a typical macroeconomic application is likely to exceed the number of observations. Model selection techniques that can handle such a situation are discussed in the next section.

The neural network model (2.1) is not the only possible universal approximator for this application. White (2006) mentions ridgelets, Candès (1998, 2003),

as an alternative. Polynomials would probably in this context not be the best possible class of universal approximators. The fit of the estimated polynomials often deteriorates at both ends of the series they describe, which is not a desirable feature in forecasting economic variables such as growth rates. Another universal approximator, the Fourier Flexible Form (FFF), is discussed in Gallant (1984). In applying the FFF, the problem of constructing the variables would have two aspects. One would have to choose the linear combinations $\gamma_j' \mathbf{z}_t$, but one would also have to decide the number of frequencies in the sum of trigonometric components. We settle for the ANN model, because it is, alongside the polynomials, probably the most commonly used universal approximator, and because QuickNet was originally designed to solve the specification and estimation problem for this model.

2.3 Modeling with three Automatic Model Selection Algorithms

We consider three model selection algorithms that apply to our modelling problem, in which the number of variables exceeds the number of observations. They are Autometrics, constructed by Doornik (2009), Marginal Bridge Estimator (MBE), see Huang et al. (2008), and QuickNet, White (2006). Autometrics is built on the principle of moving from general to specific, which means beginning with a large model and gradually reducing its size. QuickNet may be characterised as a specific-to-general-to specific procedure, although we shall also report results on a simplified specific-to-general version. The starting-point of MBE also involves all variables, but the process of selecting the final model is very different from Autometrics. We shall now describe these three techniques in more detail, beginning with Autometrics.

Autometrics

Modelling begins with a linear model called the General Unrestricted Model (GUM). When the number of variables is less than the number of observations the GUM contains all candidate variables. The model is subjected to significance tests. If all variables have statistically significant coefficient estimates, the GUM is the final model. Otherwise, because there is no unique way of going from general to specific, the algorithm searches simpler models using different search

paths. It does that by removing variables with insignificant coefficients. When the model cannot be reduced any more, it is subjected to diagnostic tests. If it passes the tests, it is called a terminal model. Since there are many search paths, there will in general be several terminal models as well.

After reaching this stage, Autometrics forms the union of the terminal models and tests the terminal models against it. The union of the models that pass the tests form a new GUM. The general-to-specific testing procedure is then repeated and a new set of terminal models obtained. If all models in this set are rejected against the new union model, the union will be the final model. Otherwise, modelling restarts with yet another GUM and continues until a final model has been reached.

In our case, the number of variables exceeds the number of observations. We follow Hendry and Krolzig (2005) and divide the variables into subsets, each of which contains fewer variables than observations. This implies that at the outset there exists more than one GUM. Each of these GUMs now forms a starting-point for Autometrics and the algorithm yields a set of terminal models for each GUM. The terminal models derived from all subsets of variables or all GUMs are merged to form a single union model. If the number of variables in this model is less than the number of observations, which happens in our application, model selection proceeds from this union model as described above.

Autometrics is partly a black box. The user can, however, affect the outcomes by selecting a number of settings, such as the significance level of the tests the algorithm relies on. We carry out tests at p-values of 0.001 and with the pre-search lag reduction switched on.

Marginal Bridge Estimator

MBE is designed for situations often occurring in statistical and genomic applications in which there is a large number of candidate variables but only a small subset of these may belong to the model. Following Huang et al. (2008), consider first the Bridge estimator (BE). This is a shrinkage estimator for a linear regression model

$$y_i = \alpha + \beta' \mathbf{x}_i + \varepsilon_i,\ i = 1, ..., n \tag{2.3}$$

where $\mathbf{x}_i = (x_{i1}, ..., x_{ip_n})'$ is a $p_n \times 1$ observation vector (p_n may increase in n but $p_n < n$,) and $\alpha = 0$ without loss of generality. Furthermore, $\varepsilon_i \sim \text{iid}(0, \sigma^2)$. BE estimates β by minimizing

$$L(\beta) = \sum_{i=1}^{n} \left(y_i - \beta' \mathbf{x}_i\right)^2 + \lambda_n \sum_{k=1}^{p_n} |\beta_k|^\gamma \tag{2.4}$$

where $\gamma > 0$ and $\lambda_n > 0$ determines the size of the penalty. Let the true parameter vector be $\beta_0 = (\beta_{10}', \beta_{20}')'$ with β_{10} having no zero entries, $\beta_{20} = \mathbf{0}$, and let $\hat{\beta}_n = (\hat{\beta}_{1n}', \hat{\beta}_{2n}')'$ be the corresponding estimator from (2.4). BE minimizes the OLS objective function plus a penalty for parameters different from zero. Hence, it shrinks estimates towards zero . Huang et al. (2008) showed that under regularity conditions parameters are i) estimated consistently ($\hat{\beta}_n \to \beta_n$ in probability), ii) the truly zero parameters are set to zero $\left(P(\hat{\beta}_{2n} = 0) \to 1\right)$ and iii) the asymptotic distribution of the estimators of nonzero parameters is the same as if only these had been included in the model. This means that the parameters of the nonzero coefficients are estimated (asymptotically) as efficiently as if only the relevant variables had been included in the model from the outset.

For BE to possess this property one needs $p_n < n$. When this condition no longer holds, MBE is applicable. The idea is to run a series of 'mini' or 'marginal' regressions, with a penalty on parameters that differ from zero. The function to be minimized equals

$$Q_n(\beta) = \sum_{k=1}^{p_n} \sum_{i=1}^{n} (y_i - \beta_k x_{ik})^2 + \lambda_n \sum_{k=1}^{p_n} |\beta_k|^\gamma \tag{2.5}$$

Let $\widetilde{\beta}_n = (\widetilde{\beta}_{1n}', \widetilde{\beta}_{2n}')'$ be the estimator of β_0 from (2.5). Under regularity conditions and $0 < \gamma < 1$, (a) the estimator $\widetilde{\beta}_{2n} = \mathbf{0}$ with probability converging to one, and (b) $P(\widetilde{\beta}_{1nk} \neq 0, \widetilde{\beta}_{1nk} \in \widetilde{\beta}_{1n}) \to 1$, as $n \to \infty$. Property (a) is similar to ii) for the BE. According to (b), the elements of $\widetilde{\beta}_{1n}$ converge to nonzero values. Thus, (a) and (b) jointly can be expected to efficiently separate the relevant variables from the rest.

Of the conditions underlying the above result the so-called partial orthogonality condition is problematic in a time series context. It states that the correlation between the relevant and irrelevant variables is not allowed to be too high. This condition can be violated if the explanatory variables are lags and functions of lags of the dependent variable as in our case. However, as we shall see in Section 2.5, MBE works quite well even in our context.

QuickNet

QuickNet (QN) resembles an earlier modelling device called RETINA, see Perez-Amaral, Gallo, and White (2003). The idea of RETINA is to find the explanatory variables that in absolute terms are most strongly correlated with y_t. The most correlated variable is selected first, and the following ones one by one thereafter. QuickNet differs from RETINA in that the set of candidate variables is different, as is the model selection criterion used for final selection. QuickNet works as follows. First, the set of candidate variables S, see (2.2), is constructed. The variables have to be such that they show sufficient variation in the sample and are not perfectly linearly correlated; see White (2006) for details. This set of candidate variables is also used when Autometrics and MBE are applied. Once this has been done, a predetermined number of variables, $\overline{q}$, are added to the model from the set S, according to the rule that selects the variable with the strongest (positive or negative) correlation with the residuals of the previously estimated model. Then a model selection criterion is applied to choose a subset of the $\overline{q}$ variables. We used 10-fold cross validation as suggested by Hastie, Tibshirani, and Friedman (2009).

We also experiment with a simplified unidirectional version of this method. The variables are selected one at a time as before, but the significance of the the added variable is tested at each step. Parsimony is appreciated, so the significance level of the tests is decreased as the number of variables n the model increases. Adding variables is terminated at the first non-rejection of the null hypothesis, so this is a pure specific-to-general strategy. In the empirical section, we apply this method such that the significance level of the first test in the sequence equals 0.2. Beginning with this value, the significance level is then halved at each step. In reporting results in Section 2.5, this method is called QN-SG. To compare the forecasts of the neural network models to genuinely nonparametric ones, direct Nadaraya-Watson kernel regression forecasts (NP) are generated. Finally, no change forecasts (NC), which forecast that the variable of interest takes the same value at any future point in time as it does at the time of forecasting, are computed and compared with the others.

2.4 Forecasting

Two Ways of Generating Multiperiod Forecasts

There are two main ways of creating multiperiod forecasts. One can either generate the forecasts recursively, or one may apply direct forecasting. In the former case, one and the same model is used for all forecast horizons. Direct forecasting implies that a separate model is built for each forecast horizon. In the empirical section of the paper we shall compare results from these two approaches. A brief discussion of these two techniques follows next.

Recursive Forecasts

In order to illuminate recursive forecasting, consider the model (2.1) with $p = q = 1$. These restrictions are for notational simplicity only. Assuming the information set $\mathscr{F}_{T-1} = \{y_{T-j}, j \geq 1\}$ is independent of future error terms, the one-period-ahead forecast made at time T equals

$$y_{T+1|T} = E\left(y_{T+1}|\mathscr{F}_T\right) = \beta_{00} + \beta_{01} y_T + \beta_1 (1 + \exp\{\gamma_0 + \gamma_1 y_T\})^{-1}.$$

The corresponding conditional mean $y_{T+2|T}$, that is, the two-period forecast, becomes

$$\begin{aligned} y_{T+2|T} &= E\left(\beta_{00} + \beta_{01} y_{T+1} + \beta_1 (1 + \exp(\gamma_0 + \gamma_1 y_{T+1}))^{-1} + \varepsilon_{T+2}|\mathscr{F}_T\right) \\ &= \beta_{00} + \beta_{01} y_{T+1|T} + \beta_1 E\left(1 + \exp(\gamma_0 + \gamma_1 (y_{T+1|T} + \varepsilon_{T+1}))^{-1}|\mathscr{F}_T\right) \\ &= \beta_{00} + \beta_{01} y_{T+1|T} + \beta_1 \int_{-\infty}^{\infty} (1 + \exp(\gamma_0 + \gamma_1 (y_{T+1|T} + z)))^{-1} \phi(z) dz \end{aligned} \tag{2.6}$$

where $\phi(z)$ is the density of the $\mathscr{N}(0, \sigma^2)$ random variable. The integral in (2.6) can be computed by numerical integration. Note that it becomes a multiple integral when the forecast horizon $h > 2$. It is therefore better to calculate its value by simulation or by bootstrapping the residuals of the model, because this remains a computationally feasible method even when $h > 2$. Some authors bypass this complication altogether by setting $\varepsilon_{T+1} = 0$ in the logistic function, and as a result their forecasts are biased estimates of the conditional mean.

In this work we apply the same bootstrap as in Chapter 1. The bootstrap has the advantage over simulation that unconditional heteroskedasticity of unknown form is allowed in the error process. More discussion about recursive forecasting

can be found in Teräsvirta (2006), Kock and Teräsvirta (2011) or Teräsvirta et al. (2010, Chapter 14) among others.

Direct forecasts

In direct forecasting, the conditional mean estimate arises from a different model for each time horizon. Given the information set $\mathscr{F}_T$, the forecast for $T+h$ made at T equals

$$y^{\mathrm{D}}_{T+h|T} = g_h(y_T, y_{T-1}, ..., y_{T-p+1})$$

where g_h is a function of y_T and its lags. In our case, model selection is made using the three aforementioned techniques, but there is a 'gap' in the model in that $y_{T+h-1}, ..., y_{T+1}$ do not enter the equation. The advantage of the direct method lies in its computational simplicity: no recursions are needed. But then, a separate model has to be specified for each forecast horizon.

Forecasts based on differences and forecast errors

The forecasts based on differences are obtained in the following way. When forecasting recursively first differences $\Delta y_t = y_t - y_{t-1}$ are being modelled and forecast. The p lags of the left hand side variable are thus $\Delta y_{t-1}, ..., \Delta y_{t-p}$ To get an h-periods-ahead forecast, which is of y_{T+h}, the first-difference forecasts have to be cumulated[1]:

$$E\left(y_{T+h}|\mathscr{F}_T\right) = \sum_{j=1}^{h} E\left(\Delta y_{T+j}|\mathscr{F}_T\right) + y_T. \tag{2.7}$$

The corresponding forecast error is $e_{T+h|T} = y_{T+h} - E\left(y_{T+h}|\mathscr{F}_T\right)$.

In direct h-periods-ahead forecasting, the variable to be modeled is $\Delta_h y_t = y_t - y_{t-h}$. The p lags of the left-hand side variable are thus $\Delta_h y_{t-h}, ..., \Delta_h y_{t-h-p+1}$ and the corresponding forecast of y_{T+h} is $E\left(\Delta_h y_{T+h}|\mathscr{F}_T\right) + y_T$. The estimated model yields direct estimates of the conditional mean.

The measure of performance in this work is the root mean square forecast error (RMSFE). It is calculated for each time series from out-of-sample forecasts for the forecasting period beginning at T_0 and ending at $T - h_{\max}$, where T is the last available observation and h_{max} is the maximum forecast horizon. Thus,

$$\mathrm{RMSFE}_h = \{(T - h_{max} - T_0 + 1)^{-1} \sum_{t=T_0}^{T-h_{\max}} e^2_{t+h|t}\}^{1/2}.$$

[1]The unknown $E\left(\Delta y_{T+j}|\mathscr{F}_T\right)$ are of course replaced by their bootstrapped counterparts.

Insanity Filters

Nonlinear models may sometimes generate forecasts that are deemed unrealistic in the light of the hitherto observed values of the time series. This has prompted forecasters to introduce precautions in order to avoid excessive forecast errors. The idea is to replace an unrealistic forecast with a more conventional and believable one. It has been applied, among others, by Swanson and White (1995, 1997a,b) who call the procedure the insanity filter, Stock and Watson (1999) and Teräsvirta et al. (2005). We shall make use of two insanity filters. The first one works as follows: If the h-step ahead predicted change exceeds the maximum h-step change observed during the estimation period, the most recently observed value of the variable to be predicted is the forecast. Hence, in the words of Swanson and White (1995) we "replace craziness by ignorance". We shall call this filter the Swanson and White (SW) filter. In the second filter, the extreme predicted change is replaced by a forecast from our benchmark linear autoregressive model: craziness is replaced by linearity.

2.5 Results

The above techniques are applied to the monthly Consumer Price Index (CPI) and unemployment series for the G7 countries as well as the four Scandinavian countries. Before considering these macroeconomic series a small Monte Carlo experiment is conducted. As mentioned in the introduction, the purpose of this exercise is to see how the three modelling procedures perform under controlled circumstances when the data generating process is known and contained in the linear span of S and thus is possible to select.

General methodology and data

The technique for generating the potential hidden units for the ANN model (2.1) is described in the Appendix. We have modified the original White (2006) technique somewhat to make it more suitable to our modelling problem. For QuickNet and MBE we used 10-fold cross validation as in Hastie et al. (2009) to determine the number of hidden units to be included. We also used the hv-Cross Validation procedure of Racine (2000) but this did not improve the results, so they are omitted. Following the suggestion of White (2006), the maximum number of variables in the ANN models was set to ten.

The macroeconomic series are obtained from the OECD Main Economic Indicators. Most series begin in the 1960s and end in December 2009 or January 2010. The CPI series were transformed to logarithms before modelling them, and the forecast errors discussed in the paper are errors in forecasting the transformed series.

Monte Carlo

For our simulation study we chose a strongly nonlinear model from Medeiros et al. (2006). These authors took the well-known annual Wolf's sunspot number series and, after transforming the observations using the Box-Cox transformation as in Ghaddar and Tong (1981), fitted an ANN model (2.1) with two hidden units to the transformed series. The model is:

$$\begin{aligned} y_t = &-0.17 + 0.85y_{t-1} + 0.14y_{t-2} - 0.31y_{t-3} + 0.08y_{t-7} \\ &+ 12.8G_1(\mathbf{y}_{t-1}) + 2.44G_2(\mathbf{y}_{t-1}) + \varepsilon_t \end{aligned} \tag{2.8}$$

where the two hidden units are

$$G_1(\mathbf{y}_{t-1}) = \left(1 + \exp\left(-0.46\left(0.29y_{t-1} - 0.87y_{t-2} + .40y_{t-7} - 6.68\right)\right)\right)^{-1}$$

and

$$G_2(\mathbf{y}_{t-1}) = \left(1 + \exp\left(-1.17 \times 10^3\left(0.83y_{t-1} - 0.53y_{t-2} - 0.18y_{t-7} + 0.38\right)\right)\right)^{-1}$$

and $\varepsilon_t \sim$i.i.d.N(0,1). We generate 500 time series of 600 observations from this model. The set of potential variables consists of G_1, G_2, 1000 other hidden units, and ten lags of y_t. The number of variables thus greatly exceeds the number of observations. The forecast horizons are one, two, and five years, and the maximum number of variables per each selected model equals ten. We report RMSFE ratios such that the denominator is the RMSFE of forecasts from (2.8), computed from the 500 replications.

Table 2.1 contains these ratios for the recursive forecasts. The first three entries in the column named DGP contain the RMSFE for the forecasts from the true model (2.8). As expected, all RMSFE ratios exceed unity. Autometrics-selected models generate by far the most accurate forecasts of the alternatives

Recursive	Hor.	DGP	AR	QN	MBE	Autom.	QN-SG
	1	1.82	1.456	1.343	1.730	1.105	1.805
NF	2	2.739	1.536	3.282	1.659	1.073	1.568
	5	4.172	1.337	$9 \cdot 10^4$	1.394	4023	1.283
	1	1	1.456	1.513	1.730	1.105	1.855
SW	2	1.001	1.536	1.532	1.658	1.074	1.552
	5	1.001	1.392	1.218	1.395	1.028	1.269
	1	1	1.456	1.322	1.730	1.105	1.776
AR	2	1.001	1.536	1.366	1.658	1.074	1.552
	5	1.001	1.337	1.214	1.395	1.028	1.269

Table 2.1: Average root mean square forecast error ratios for the recursive forecasts of the simulated sunspot series. DGP: Data generating process, AR: Autoregression, QN: QuickNet, MBE: Marginal Bridge Estimator, Autom.: Autometrics, QN-SG: Quick-Net specific to general. NF: No Filter (for the DGP the NF subcolumn contains the actual root mean square forecast error from forecasting with the DGP), SW: Swanson-White filter, AR: Insane forecasts replaced by linear autoregressive ones.

Direct	Hor.	AR	QN	MBE	Autom.	QN-SG	NP	NC
	1	1.456	1.343	1.730	1.105	1.805	1.546	3.560
NF	2	1.518	9.575	1.549	1.652	1.436	1.332	4.226
	5	1.306	1.353	1.241	1.359	1.293	1.124	3.984
	1	1.456	1.513	1.730	1.105	1.855	1.658	
SW	2	1.518	1.52	1.549	1.532	1.733	1.424	
	5	1.363	1.326	1.241	1.322	1.293	1.124	
	1	1.456	1.322	1.730	1.105	1.776	1.555	
AR	2	1.518	1.35	1.549	1.355	1.444	1.335	
	5	1.306	1.219	1.241	1.246	1.293	1.124	

Table 2.2: Average root mean square forecast error ratios for the direct forecasts of the simulated sunspot series. NP: Non-parametric, NC: No Change forecasts. NF: No Filter, SW: Swanson-White filter, AR: Insane forecasts replaced by linear autoregressive ones.

to the DGP, indicating that the method works well when there is a true model that can be selected from the set of variables available for the purpose. The other methods lead to models whose forecasts are of more or less the same quality. The forecasts from MBE-selected models do not need filtering but are nevetheless slightly more inaccurate than the other (filtered) ones.

The performance of direct models is reported in Table 2.2. Models selected by Autometrics no longer generate more accurate forecasts than the other non-linear models. Every possible direct model is misspecified by definition because the shortest lag (two-year model) or lags (five-year model) of y_t cannot be used, and Autometrics clearly suffers from this. Note the good performance of the nonparametric model forecasting five years ahead. The kernel autoregression

Rec	Hor.	DGP	AR	QN	MBE	Autom.	QN-SG
	1	3.35	4.53	4.03	4.92	3.68	4.6
AR	2	5.18	7.28	6.13	7.68	5.77	7.01
	5	5.5	7.28	6.24	7.16	5.49	6.6

Dir	Hor.	AR	QN	MBE	Autom.	QN-SG	NP	NC
	1	4.53	4.03	4.92	3.68	4.6	4.5	6.23
AR	2	7.19	6.65	7.54	6.2	6.93	6.67	10.4
	5	7.1	6.6	6.79	6.47	7.07	6.15	10.4

Table 2.3: Average ranks based on the absolute forecasts errors. For each procedure for which forecasts are carried out recursively as well as directly the forecasts from the two alternatives are identical at the 1-month horizon. Hence, the comparison is only made across the DGP forecasts and the direct forecasts at the 1-month horizon and by construction the ranks are the same for the recursive counterparts.

seems to make most of the available information, and the forecasts hardly need filtering. In fact, the SW filter has a negative effect on the accuracy of the forecasts from this model. As may be expected, the No Change forecast does not perform well in predicting these strongly cyclical realisations.

We also compare the methods by calculating the average ranks of the absolute forecast errors. Only the results for the AR filtered forecasts are reported since the ranks obtained from the SW filtered ones are similar.

The ranks can be found in Table 2.3. As can be expected from the RMSFE results, the forecasts from the DGP have the lowest ranks. However, the ranks of the recursive forecasts by Autometrics are not much higher and even as low as the DGP ones at the five year horizon. Of the remaining neural network procedures MBE forecasts have the highest ranks while the No Change forecasts are by far the least accurate overall. This is not surprising due to the cyclical nature of the series to be forecast. The nonparametric forecasts perform about as well as the ANN-based procedures at the shortest horizons and better than them at the five year horizon.

Another robust way of considering the results is to use Wilcoxon's signed-rank test (Wilcoxon (1945)) for comparing forecasts from the DGP with the others. The null hypothesis is that the absolute forecast error of the DGP and that of the other model have the same mean whereas the alternative is that the alternative model has a lower mean absolute forecast error. The tests are carried out separately for each horizon. The results are reported in Table 2.4. A normal approximation has been used in calculating the p-values. This is appropriate due to the large number of forecasts (500). Small p-values indicate that the alternative model produces more accurate forecasts than the DGP. If the alternative

Recursive	Hor.	AR	QN	MBE	Autom.	QN-SG
	1	1	1	1	1	1
AR	3	1	1	1	1	1
	5	1	1	1	0.852	1

Direct	Hor.	AR	QN	MBE	Autom.	QN-SG	NP	NC
	1	1	1	1	1	1	1	1
AR	3	1	1	1	1	1	1	1
	5	1	1	1	1	1	1	1

Table 2.4: *p*-values of the Wilcoxon signed-rank test for testing the null of the mean of the forecast errors from the DGP being equal to the mean of the corresponding forecast error from the model in each column of the table. The tests are carried out separately at each horizon and the alternative hypothesis is that the model in the table has a lower mean. Top panel: Recursive forecasts. Bottom panel: Direct forecasts.

Recursive	Total	Linear	Nonlinear	DGP units
QN	9.55	0.348	9.21	1.64
MBE	9.22	0.756	8.47	0.77
Autom	11	1.5	9.51	3.47
QN-GS	5.3	0.324	4.98	1.12

Table 2.5: Average number of variables selected for the recursive forecasts of the CPI based on differences. "Total" indicates total number of variables included, "Linear" indicates the number of linear units included, "Nonlinear" gives the number of hidden units included, and DGP units gives the number of units included from the data generating process.

hypothesis is that the forecasts from the DGP have the lowest mean, one simply subtracts the reported *p*-values from one and obtains the *p*-values of this test. All tests are based on the AR-filtered forecasts[2].

As can be seen from Table 2.4, the results in Tables 2.1 and 2.2 accord with those from the Wilcoxon test. It is not possible to reject the hypothesis that the absolute forecast errors of the DGP forecasts and those from the alternative model have the same mean if the alternative hypothesis is that the alternative model has a lower mean. If the alternative hypothesis is that the DGP forecast errors have a lower mean, the null of equal means is rejected at a five percent significance level with a single exception: the recursive five-year forecasts from the Autometrics-selected ANN model.

Table 2.5 offers some background to the results in Tables 2.1 and 2.2. It con-

[2]Alternatively, one could consider the Giacomini-White test (Giacomini and White (2006)) which includes the Diebold-Mariano test (Diebold and Mariano (1995)) as a special case. The Giacomini-White test, however, relies on a rolling window. The Giacomini-White test was also carried out but most often the conclusions were the same as for the Wilcoxon test and so the results are not reported here.

tains information about the size and variable types in the nonlinear models for recursive forecasting. The average number of variables in every type of model is larger than the size of (2.8) which is six variables as the intercept is not counted. It is worth noting that Autometrics, while selecting the largest models, picks up elements of the true model more frequently than the other model selection techniques. This is probably the most important factor in explaining its success in forecasting. Moreover, Autometrics on average chooses more linear lags than the other models, although fewer than their number in the true model. The average number of linear lags in the other models is rather small. The specific-to-general QN-SG is clearly more parsimonious than QuickNet, but this result is not invariant to the choice of significance levels in the test sequence. QuickNet-based recursive forecasts are somewhat more accurate than QN-SG ones at one-and five-year horizons.

Macroeconomic forecasts

The CPI and unemployment series are forecast at the 1, 3, 6, and 12-month horizons. The CPI series are transformed into logarithms, and 240 forecasts based on an expanding window are generated for each horizon[3]. Forecasts from models of differenced series are formed as described in Section 2.4. The pool of variables contains 600 hidden units with $p = 6$ in (2.1) and the first six linear lags of the dependent variable.

The models are respecified every six months. This is because Autometrics is quite slow: otherwise respecification could easily be done every month. Pretesting linearity and letting the nonlinear model selection operate only if the linearity hypothesis was rejected did not on average improve the performance of the nonlinear models. This may be due the fact that linear lags are included into the pool of hidden units which makes it possible to select a linear model anyway.

Consumer Price Index

The RMSFE ratios for recursive CPI forecasts from models of differenced series can be found in Table 2.6. The denominator in the RMSFE ratio is now the RMSFE of the recursive linear AR forecasts. It is seen that filtering the forecasts is necessary. All four model selection techniques lead to ANN models that

[3]For some of the shorter data sets the number of forecasts is less than 240, because the first window was set to include at least 200 observations.

Recursive	Hor.	AR	QN	MBE	Autom.	QN-SG
	1	1	16.82	1.02	257.9	1.043
NF	3	1	$5 \cdot 10^4$	$2 \cdot 10^6$	$2 \cdot 10^9$	1.052
	6	1	$4 \cdot 10^5$	$1 \cdot 10^6$	$6 \cdot 10^9$	2.411
	12	1	$1 \cdot 10^6$	$1 \cdot 10^6$	$1 \cdot 10^{10}$	$3 \cdot 10^5$
	1	1	1.040	1.020	1.074	1.047
SW	3	1.004	1.033	1.020	1.075	1.061
	6	1.003	1.055	1.020	1.085	1.076
	12	1.011	1.107	1.034	1.172	1.091
	1	1	1.042	1.019	1.072	1.044
AR	3	1	1.025	1.014	1.058	1.052
	6	1	1.036	1.017	1.047	1.071
	12	1	1.066	1.032	1.105	1.088

Table 2.6: Average root mean square forecast error ratios for the recursive forecasts of the CPI series based on differences. NF: No Filter, SW: Swanson-White filter, AR: Insane forecasts replaced by linear autoregressive ones.

generate some very inaccurate forecasts. This is the case already for one-month forecasts and is due to the fact that some models contain very strongly correlated variables. A pair of them typically has large (in absolute value) coefficients with opposite signs. Forecasting with such a model yields inaccurate forecasts and cumulating them in forecasting more than one month ahead makes the situation even worse. This is clearly seen from the table. Furthermore, all ratios exceed one, which means that on average no ANN model, not even after filtering, generates more accurate recursive forecasts than the linear AR model. Models selected by MBE perform slightly better than the other nonlinear models.

These results may be compared with the ones in Table 2.7. This table contains the RMSFE ratios for direct forecasts from models built using differenced series. Models built using QuickNet and Autometrics still generate a few forecasts that require filtering, whereas MBE-based forecasts do not. After filtering the six- and 12-month forecasts from the ANN models are more accurate than the benchmark ones. This is also the case for forecasts from direct linear AR models. Their RMSFE ratios are comparable to those obtained from models built by MBE which is the best-performing model selection technique. The forecasting performance of the nonparametric model is below average, and the 'no change' forecasts are less accurate than even the corresponding recursive ones.

The RMSFE ratios in Table 2.8 refer to recursive forecasts from models built on CPI levels. Filtered forecasts are more accurate on average than the corresponding forecasts in Table 2.6. MBE-based forecasts are the most accurate

Direct	Hor.	AR	QN	MBE	Autom.	QN-SG	NP	NC
	1	1	16.82	1.02	257.9	1.043	1.148	1.133
NF	3	0.976	2.699	0.9893	2464	1.02	1.074	1.169
	6	0.8123	20.77	0.8239	1869	0.8362	0.9335	1.159
	12	0.7336	3.286	0.7284	20.08	0.7436	0.8203	1.134
	1	1	1.040	1.020	1.074	1.047	1.150	
SW	3	0.976	1.039	0.9893	1.059	1.030	1.081	
	6	0.8123	0.8452	0.8239	0.8987	0.836	0.9335	
	12	0.7336	0.7584	0.7284	0.8355	0.7397	0.8203	
	1	1	1.042	1.019	1.072	1.044	1.147	
AR	3	0.976	1.020	0.9893	1.042	1.019	1.075	
	6	0.8123	0.840	0.8239	0.8819	0.835	0.9335	
	12	0.7336	0.7591	0.7284	0.8371	0.7395	0.8203	

Table 2.7: Average root mean square forecast error ratios for the direct forecasts of the CPI series based on differences. NF: No Filter, SW: Swanson-White filter, AR: Insane forecasts replaced by linear autoregressive ones.

Recursive	Hor.	AR	QN	MBE	Autom.	QN-SG
	1	1.011	1.013	0.977	1.062	1.139
NF	3	1.001	20.39	0.9315	6311	1.195
	6	0.9728	$3 \cdot 10^5$	0.8535	$1 \cdot 10^7$	1.223
	12	0.9372	$3 \cdot 10^6$	0.787	$3 \cdot 10^8$	1.309
	1	1.011	1.013	0.977	1.062	1.139
SW	3	1.001	0.9685	0.9314	1.003	1.184
	6	0.9728	0.896	0.8532	0.9299	1.187
	12	0.9372	0.823	0.7871	0.8489	1.185
	1	1.011	1.013	0.977	1.062	1.139
AR	3	1.001	0.9661	0.9314	1.003	1.181
	6	0.9728	0.8923	0.8532	0.9299	1.187
	12	0.9372	0.8143	0.7871	0.8489	1.167

Table 2.8: Average root mean square forecast error ratios for the recursive forecasts of the CPI series based on levels. NF: No Filter, SW: Swanson-White filter, AR: Insane forecasts replaced by linear autoregressive ones.

ones and models built by QN-SG generate the least accurate recursive forecasts: all ratios remain above one. Recursive linear AR models built on levels are somewhat superior to the ones built on differences. The RMSFE ratios lie below one for the two longest horizons but are greater than the corresponding ratios for forecasts from models obtained by MBE, QuickNet and Autometrics.

Table 2.9 contains the RMSFE ratios for direct forecasts from models specified and estimated from the level series. It appears that MBE is the best model-building method when the criterion is the RMSFE. The ratios are even smaller

Direct	Hor.	AR	QN	MBE	Autom.	QN-SG	NP	NC
	1	1.011	1.013	0.977	1.062	1.139	16.77	1.133
NF	3	0.9661	0.9418	0.9057	0.9761	1.198	8.037	1.169
	6	0.9053	3.401	0.8114	0.9982	1.204	5.072	1.159
	12	0.7771	0.7205	0.6928	0.9416	1.173	3.119	1.134
	1	1.011	1.013	0.977	1.062	1.139	3.783	
SW	3	0.9661	0.9418	0.9057	0.9761	1.198	5.172	
	6	0.9053	0.8305	0.8114	0.954	1.204	4.907	
	12	0.7771	0.7205	0.6928	0.9416	1.173	3.119	
	1	1.011	1.013	0.977	1.062	1.139	3.675	
AR	3	0.9661	0.9418	0.9057	0.9761	1.198	5.136	
	6	0.9053	0.8303	0.8114	0.9564	1.204	4.904	
	12	0.7771	0.7205	0.6928	0.9416	1.173	3.119	

Table 2.9: Average root mean square forecast error ratios for the direct forecasts of the CPI series based on levels. NF: No Filter, SW: Swanson-White filter, AR: Insane forecasts replaced by linear autoregressive ones.

than the ones found in Tables 2.6–2.8. Direct models selected by QuickNet also perform better than the recursive ones, whereas the same cannot be said of models based on Autometrics or QN-SG. In the light of these results, going from specific to general and back again (QuickNet) is a better idea than going from specific to general only (QN-SG), but this finding cannot be generalized. It may be noted that the nonparametric model built on levels generates much less accurate forecasts than the same model estimated from differenced series. Its RMSFE ratios are remarkably larger than any other ratio. Summing up, it seems that direct forecasts are on average more accurate than the recursive ones. Exceptions do exist: compare Autometrics-based six- and 12-month RMSFE ratios in Tables 2.8 and 2.9. It should be pointed out that these results are aggregate ones and do not necessarily hold for all 11 countries.

As was the case for the simulation study we also compare the forecast performance of the methods applied by considering their ranks. This is done for all countries and forecast horizons. Furthermore, forecasts from models built on differences and the ones based on levels are included in the same comparison.

The results can be found in Table 2.10. At the 1-month horizon absolute forecast errors from the ANN procedures have ranks very close to each other, which is in accordance with the findings from Tables 2.6-2.9. The nonparametric and No Change forecasts have considerably higher ranks than the other procedures. This is true for the forecasts based on differences as well as the ones based on levels. In particular the high ranks for the nonparametric forecasts based on levels are no surprise in the light of the high relative RMSFE in

Rec Diff	Hor.	AR	QN	MBE	Autom.	QN-SG
	1	6.53	6.62	6.57	6.89	6.71
AR	3	11.8	11.9	11.7	12	12.4
	6	12.9	12.8	12.7	12.8	13.5
	12	14.5	14.7	14.4	15.1	15.1

Dir Diff	Hor.	AR	QN	MBE	Autom.	QN-SG	NP	NC
	1	6.53	6.62	6.57	6.89	6.71	8.2	7.38
AR	3	11	11.2	11.2	11.1	11.3	13.8	13.6
	6	8.95	9.25	9.05	9.85	9.37	12.5	14.2
	12	8.9	8.66	8.65	9.44	8.52	11.4	14.4

Rec Level	Hor.	AR	QN	MBE	Autom.	QN-SG
	1	6.62	6.32	6.21	6.49	7.68
AR	3	11.7	10.6	10.7	10.8	14.1
	6	12.1	10.8	10.7	10.5	14.9
	12	12.3	10.2	10.5	9.89	14.9

Dir Level	Hor.	AR	QN	MBE	Autom.	QN-SG	NP	NC
	1	6.62	6.32	6.21	6.49	7.68	7.97	7.38
AR	3	10.5	10.3	10.1	10.7	14.3	17.7	13.6
	6	10.5	9.95	9.73	10.8	15	22	14.2
	12	8.91	9.32	8.94	9.89	14.5	21.6	14.4

Table 2.10: Average ranks based on the absolute forecasts errors. For each procedure for which forecasts are carried out recursively as well as directly the forecasts from the two alternatives are identical at the 1-month horizon. Hence, the comparison is only made across the direct forecasts at the 1-month horizon and by construction the ranks are the same for the recursive counterparts.

Table 2.9. In general, the direct methods have the lowest average ranks. This is the case in particular for the forecasts based on the differences of the series. The overall winner at the 12-month horizon is the QN-SG method based on differences while the MBE-based forecasts come in second. This again agrees with the results reported in Tables 2.6-2.9.

Similarly to the simulated example, we conduct Wilcoxon's signed-rank test for pairs of absolute forecast error series. The benchmark, the recursive linear AR forecast based on differences, is always one of the forecasts in the pair. As already discussed, the null hypothesis of the test is that the means of the absolute forecast errors are equal, and the alternative is that the absolute forecast errors of the 'other model' have the smaller mean of the two. The upper panel of Table 2.11 contains p-values of the test for recursive forecasts from differenced models. Most of them are close to one, which means that the null hypothesis is rejected in the opposite direction. This accords with the information in Table 2.6, where all RMSFE ratios were greater than one. The MBE-based forecasts

Recursive	Hor.	QN	MBE	Autom.	QN-SG
	1	0.998	0.778	1	1
AR	3	0.963	0.0582	0.982	1
	6	0.949	0.266	0.995	1
	12	1	1	1	1

Direct	Hor.	AR	QN	MBE	Autom.	QN-SG	NP	NC
	1		0.998	0.778	1	1	1	1
AR	3	0	0.011	$1 \cdot 10^{-5}$	$1 \cdot 10^{-5}$	0.005	1	1
	6	0	0	0	0	0	0	1
	12	0	0	0	0	0	0	0.968

Table 2.11: p-values of the Wilcoxon signed-rank test for testing the null of the mean of the forecast errors from recursive forecasts of the CPI series from the linear AR estimated on differences is equal to the mean of the corresponding forecast error from the model in each column of the table. The tests are carried out separately at each horizon and the alternative hypothesis is that the model in the table has a lower mean. Top panel: Models estimated recursively on differences. Bottom panel: Models estimated directly on differences.

at horizons up to six months constitute the only exception.

The lower panel contains the p-values for direct forecasts. They are mostly close to zero for long forecasting horizons. The no change forecast is the only exception: all p-values are large. Direct forecasts can thus be deemed superior to recursive ones when the models are built on differenced CPI-series. This strengthens conclusions that emerge from Tables 2.6 and 2.7.

Table 2.12 contains p-values of the same test and null hypothesis when the forecasts are obtained using models built on CPI series in levels. Results in the upper panel show that the null hypothesis is rejected in favour of the recursive linear AR forecasts based on differences when compared to the model selected by QN-SG. The other methods generate ANN models that yield more accurate recursive forecasts than the linear AR model (p-values are close to zero) or forecasts for which the null hypothesis is not rejected (QuickNet and Autometrics one-month forecasts). The lower panel shows that QN-SG-based direct models do not perform well either. The same can be said about the nonparametric model and the 'no change' forecasts. Considering all four horizons at once, MBE emerges as the best-performing model selection criterion for direct models when Wilcoxon's test is used as the yardstick.

As in the simulated example, it is interesting to see whether the size of the model and the accuracy of the forecasts from it have to do with each other. Table 2.13 contains information about the size and composition of models based on differenced series. When forecasting recursively, it is seen from the left panel

Recursive	Hor.	QN	MBE	Autom.	QN-SG
	1	0.160	$1 \cdot 10^{-4}$	0.171	1
AR	3	$1 \cdot 10^{-8}$	0	$4 \cdot 10^{-6}$	1
	6	0	0	0	1
	12	0	0	0	1

Direct	Hor.	AR	QN	MBE	Autom.	QN-SG	NP	NC
	1	1	0.160	$1 \cdot 10^{-4}$	0.171	1	1	1
AR	3	0	0	0	$4 \cdot 10^{-8}$	1	1	1
	6	0	0	0	0	1	1	1
	12	0	0	0	0	1	1	0.968

Table 2.12: p-values of the Wilcoxon signed-rank test for testing the null of the mean of the forecast errors from recursive forecasts of the CPI series from the linear AR estimated on differences is equal to the mean of the corresponding forecast error from the model in each column of the table. The tests are carried out separately at each horizon and the alternative hypothesis is that the model in the table has a lower mean. Top panel: Models estimated recursively on levels. Bottom panel: Models estimated directly on levels.

Recursive	Total	Linear	Nonlinear
QN	6.35	0.298	6.05
MBE	5.51	0.818	4.69
AM	15.5	0.393	15.1
QN-SG	4.03	0.195	3.83

Direct MBE	Total	Linear	Nonlinear
1 mth	5.51	0.818	4.69
3 mths	5.48	2.45	3.03
6 mths	5.29	3.55	1.74
12 mths	2.69	1.72	0.964

Table 2.13: Left panel: Average number of variables selected for the models generating recursive forecasts of the CPI based on differences. "Total" indicates total number of variables included, "Linear" indicates the number of linear units included, and "Nonlinear" gives the number of hidden units included. Right panel: Average number of variables selected for the direct forecasts of the CPI based on differences by MBE.

that QN-SG selects the most parsimonious models which do not, however, yield the most accurate forecasts. MBE selects somewhat less parsimonious models that on average yield the most accurate recursive forecasts. It also chooses the largest fraction of linear lags, although their average number remains below one. Models selected by Autometrics are by far the largest ones. There does not seem to be a clear connection between the model size and forecast accuracy.

The right-hand panel of Table 2.13 contains the average size and composition of models based on differenced series and selected by MBE for direct forecasting. The average number of variables is halved when one moves from six- to 12-month models, whereas the share of linear lags of the total increases up to six-month models and remains about the same for 12-month ones.

Table 2.14 contains the same information for models built on levels. All methods now select more linear variables than in the previous case. QN-SG is still the most parsimonious technique, and even QuickNet selects fewer vari-

Recursive	Total	Linear	Nonlinear
QN	5.35	1.09	4.27
MBE	7.19	5.64	1.55
AM	19.1	1.34	17.7
QN-SG	1.39	1	0.386

Direct MBE	Total	Linear	Nonlinear
1 mth	7.19	5.64	1.55
3 mths	7.24	5.74	1.49
6 mths	7.42	6	1.42
12 mths	7.21	6	1.21

Table 2.14: Left panel: Average number of variables selected for the models generating recursive forecasts of the CPI based on levels. "Total" indicates total number of variables included, "Linear" indicates the number of linear units included, and "Nonlinear" gives the number of hidden units included. Right panel: Average number of variables selected for the direct forecasts of the CPI based on levels by MBE.

ables than MBE. As Tables 2.9 and 2.12 indicate, forecasts from MBE are still the most accurate ones on average. The use of Autometrics leads to largest models. They perform better than QN-SG-selected models but less well than ones specified using MBE. The right panel of table 2.14 shows that MBE select a large number of linear lags for all direct models. In fact, every MBE-model built for the two longest horizons contains all six lags and only a small number of hidden units. A comparison of the RMSFE ratios in Tables 2.6 and 2.8 on the one hand and Tables 2.7 and 2.9 on the other (indirectly) suggests that direct models based on level data and selected by MBE may be slightly superior to the same type of model, selected by the same technique, but based on differenced series. Whether or not this is due to the larger amount of linear lags in the former models is not clear, however.

Individual countries

To shed light on some of the cross-country variation in the results that cannot be seen in the summary tables we now consider results for some individual countries, Italy, Japan, and the US. They are selected because there are interesting differences between them. The remaining country-specific RMSFE are available at `http://econ.au.dk/research/research-centres/creates/research/research-papers/supplementary-downloads/rp-2011-27/`.

Tables 2.15 and 2.16 show the RMSFE ratios for the US CPI forecasts based on differences (only the results for the AR-filter are presented). It is seen that it is indeed possible to improve upon the linear AR model even when forecasting recursively, although this is not true for all three methods. In fact, only MBE outperforms the linear autoregression at all horizons, which again indicates it may be superior to QuickNet and Autometrics in forecasting the CPI series. The

US Recursive	Hor.	AR	QN	MBE	Autom.	QN-SG
	1	1	1.03	0.9803	1.044	1.011
AR	3	1	0.9811	0.9722	1.003	1.034
	6	1	0.9836	0.9388	0.9769	1.006
	12	1	1.017	0.9484	1.033	1.058

Table 2.15: Average root mean square forecast error ratios for the recursive forecasts of the US CPI series based on differences. AR: Insane forecasts replaced by linear autoregressive ones.

US Direct	Hor.	AR	QN	MBE	Autom.	QN-SG	NP	NC
	1	1	1.03	0.9803	1.044	1.011	1.044	1.221
AR	3	0.9952	0.9742	0.9988	0.9298	0.9403	0.9417	1.179
	6	0.8312	0.8749	0.8332	0.8656	0.8825	0.8098	1.224
	12	0.9483	1.014	0.9145	0.9856	0.9501	0.8423	1.598

Table 2.16: Average root mean square forecast error ratios for the direct forecasts of the US CPI series based on differences. AR: Insane forecasts replaced by linear autoregressive ones.

Wilcoxon tests were also carried out on the individual countries. Based on these, the above findings are significant since at no horizon does one observe a higher p-value than 0.044 when testing the AR forecasts against the recursive MBE ones.

On average MBE selects seven variables of which 3.38 are linear lags when forecasting recursively. It is more parsimonious and includes a higher number of linear lags than the other procedures.

A comparison of Tables 2.15 and 2.16 shows that for the US the the recursive forecasts are less accurate than the corresponding direct ones. The differences in the RMSFE are, however, less pronounced than in Tables 2.6 and 2.7. The finding that the direct forecasts are more accurate than the recursive ones is uniform across all countries. All ANN models, independent of the variable selection procedure, work well in direct forecasting. However, for the US they are at the longest horizons outperformed by the nonparametric model and perform less well than they do in general. For the direct forecasts MBE is again the most parsimonious procedure whereas Autometrics on average selects the largest models. MBE-based models also contain the largest number of linear units.

The averaged results for the forecasts based on levels also sometimes hide differences between the individual countries. To illustrate this, Tables 2.17 and 2.18 present RMSFE ratios for Italy, Japan, and the US (only the results based on the AR-filter are shown). Table 2.17 shows that there can be considerable variation in the performance of the variable selection procedures. MBE is the

ITA Recursive	Hor.	AR	QN	MBE	Autom.	QN-SG
	1	1.002	1.165	0.9263	1.603	1.757
AR	3	1.055	1.188	0.8965	1.629	1.955
	6	1.135	1.252	0.9073	1.677	2.067
	12	1.195	1.264	0.8355	1.715	1.855
JP Recursive	Hor.	AR	QN	MBE	Autom.	QN-SG
	1	0.9667	1.02	0.9885	1.069	1.021
AR	3	0.9188	0.9652	0.9612	0.8725	0.9721
	6	0.7846	0.8702	0.8419	0.6776	0.8807
	12	0.6717	0.755	0.7242	0.5016	0.7644
US Recursive	Hor.	AR	QN	MBE	Autom.	QN-SG
	1	0.9999	1.008	0.9477	0.9458	1.088
AR	3	0.9961	0.87	0.9152	0.9035	0.9962
	6	0.9877	0.7469	0.8597	0.8102	0.9509
	12	0.9633	0.6898	0.9324	0.8573	1.125

Table 2.17: Average root mean square forecast error ratios for the recursive forecasts of the CPI series based on levels. AR: Insane forecasts replaced by linear autoregressive ones.

most stable procedure and the only one which has RMSFE ratios below unity for all three countries at all horizons[4], but for each country a different variable selection procedure is dominant. The relative stability of MBE is most likely due to the fact that for all three countries this procedure selects the largest number of linear units. For Italy and the US it includes all six linear units and for Japan it chooses a purely linear model every time (though not the AR(6)). Nevertheless, MBE is outperformed by Autometrics which generally chooses only a small fraction of linear lags. However, MBE includes unusually few linear units (3.6) for Japan, so it may still argued that models with a high number of linear units combined with a few nonlinear ones perform well on average.

A comparison of the results in Table 2.17 with the ones in Table 2.18 indicates that that the direct forecasts are superior to their recursive counterparts. This accords with the overall results. Moreover, the nonparametric model generates very inaccurate forecasts for these three countries, which is also in line with the general results. The direct MBE forecasts again have RMSFE ratios below unity. The performance of Autometrics varies quite remarkably. In forecasting

[4]Recall, however, that the benchmark is the AR(6) model forecasted recursively based on differences of the time series. But even then, it still illustrates the rather stable performance of MBE. Its relative RMSFE are actually below one for the recursive level based forecasts for all countries at all horizons except for the UK and Denmark for which the 1-month forecasts have ratios above one.

ITA Direct	Hor.	AR	QN	MBE	Autom.	QN-SG	NP	NC
	1	1.002	1.165	0.9263	1.603	1.757	0.9978	1.689
AR	3	0.9942	1.091	0.875	1.423	1.977	13.64	1.943
	6	0.9135	1.01	0.8358	1.428	1.949	8.529	1.942
	12	0.7215	0.8067	0.7564	1.571	1.772	4.438	1.733
JP Direct	Hor.	AR	QN	MBE	Autom.	QN-SG	NP	NC
	1	0.9667	1.02	0.9885	1.069	1.021	4.768	0.9762
AR	3	0.883	0.925	0.9675	0.8878	0.9711	4.708	0.872
	6	0.7352	0.8424	0.7352	0.7064	0.8715	2.999	0.662
	12	0.5334	0.6727	0.5334	0.5289	0.699	1.591	0.4879
US Direct	Hor.	AR	QN	MBE	Autom.	QN-SG	NP	NC
	1	0.9999	1.008	0.9477	0.9458	1.088	0.9979	1.221
AR	3	0.9994	0.8962	0.9212	0.9275	1.014	1.883	1.179
	6	1.007	0.7585	0.8741	0.9568	0.9696	4.698	1.224
	12	1.032	0.7579	0.9107	0.9179	1.147	4.062	1.598

Table 2.18: Average root mean square forecast error ratios for the direct forecasts of the CPI series based on levels. AR: Insane forecasts replaced by linear autoregressive ones. The NC forecasts are not filtered.

12 months ahead, Autometrics-based forecasts are an excellent choice for Japan, a mediocre one for the US, and are definitely not to be recommended for forecasting the Italian CPI. The situation is the same if the recursive forecasts in Table 2.17 are considered. Autometrics-based forecasts are much better than the recursive AR ones for Japan, except at the one-month horizon, still acceptable for the US, and very inaccurate for Italy.

Results on forecasting the CPI series sugest that forecasts based on levels are superior to their counterparts based on differences. Furthermore, direct forecasting is preferable to recursive forecasts and MBE is the most stable forecasting procedure. This last observation may be attributed to the high number of linear units MBE includes and which it supplements with a few relevant nonlinear units.

Unemployment

A common feature of results on forecasting unemployment rate series with those on forecasting CPI series is the appearance of some vastly inaccurate forecasts and the consequent need for filtering. This is first seen from Table 2.19 that contains the RMSFE ratios for recursive forecasts when the models are built on differenced series. For filtered foreacsts, all ratios still lie above one. Models

Recursive	Hor.	AR	QN	MBE	Autom.	QN-SG
	1	1	1.124	1.005	$2 \cdot 10^9$	1.045
NF	3	1	97.78	1.001	$7 \cdot 10^9$	1.054
	6	1	3333	1.003	$1 \cdot 10^{10}$	1.051
	12	1	$5 \cdot 10^4$	1.006	$1 \cdot 10^{10}$	1.026
	1	1	1.090	1.006	1.216	1.079
SW	3	1.004	1.081	1.007	1.239	1.073
	6	1.004	1.058	1.008	1.26	1.056
	12	1.001	1.026	1.008	1.221	1.026
	1	1	1.068	1.005	1.161	1.049
AR	3	1	1.07	1.002	1.152	1.058
	6	1	1.05	1.004	1.142	1.051
	12	1	1.021	1.006	1.092	1.025

Table 2.19: Average root mean square forecast error ratios for the recursive forecasts of the unemployment series based on differences. NF: No Filter, SW: Swanson-White filter, AR: Insane forecasts replaced by linear autoregressive ones.

Direct	Hor.	AR	QN	MBE	Autom.	QN-SG	NP	NC
	1	1	1.124	1.005	$2 \cdot 10^9$	1.045	0.9999	1.109
NF	3	0.9979	59.88	1.024	$7 \cdot 10^6$	1.063	1.024	1.167
	6	1.002	1.133	1.031	$2 \cdot 10^9$	1.133	1.046	1.148
	12	1.031	250.1	1.054	$2 \cdot 10^8$	1.197	1.091	1.055
	1	1	1.090	1.006	1.216	1.079	1.013	
SW	3	1.001	1.060	1.028	1.197	1.062	1.022	
	6	1.002	1.104	1.030	1.196	1.116	1.046	
	12	1.030	1.101	1.049	1.223	1.128	1.080	
	1	1	1.068	1.005	1.161	1.049	0.9999	
AR	3	0.9979	1.053	1.025	1.184	1.058	1.023	
	6	1.002	1.101	1.028	1.195	1.113	1.042	
	12	1.031	1.109	1.047	1.215	1.138	1.082	

Table 2.20: Average root mean square forecast error ratios for the direct forecasts of the unemployment series based on differences. NF: No Filter, SW: Swanson-White filter, AR: Insane forecasts replaced by linear autoregressive ones.

selected by MBE appear to lead to the most accurate nonlinear forecasts, and they do not need filtering. Autometrics-selected models are, even after filtering, the most inaccurate ones. Table 2.20 indicates that on average, direct forecast are not superior to recursive ones. This is true for both linear and nonlinear forecasts. Nonparametric forecasts do not require much filtering but are less accurate than the ones from the MBE-forecasts.

In the case of unemployment series, the models built on levels do not produce forecasts superior to their counterparts from models based on differences. Table

Recursive	Hor.	AR	QN	MBE	Autom.	QN-SG
	1	0.9994	$2 \cdot 10^5$	1.007	1.302	1.048
NF	3	1.010	$3 \cdot 10^5$	1.028	$1 \cdot 10^8$	1.076
	6	1.024	$5 \cdot 10^6$	1.041	$8 \cdot 10^8$	1.087
	12	1.016	$9 \cdot 10^6$	1.036	$1 \cdot 10^9$	1.065
	1	0.9994	1.064	1.007	1.148	1.048
SW	3	1.01	1.079	1.027	1.147	1.076
	6	1.018	1.106	1.036	1.135	1.087
	12	1.011	1.085	1.030	1.098	1.061
	1	0.9994	1.067	1.007	1.146	1.048
AR	3	1.010	1.080	1.027	1.150	1.076
	6	1.024	1.108	1.042	1.135	1.087
	12	1.016	1.093	1.034	1.104	1.061

Table 2.21: Average root mean square forecast error ratios for the recursive forecasts of the unemployment series based on levels. NF: No Filter, SW: Swanson-White filter, AR: Insane forecasts replaced by linear autoregressive ones.

2.21 contains the RMSFE ratios for recursive forecasts. Again, MBE-selected models seem to generate more accurate forecasts than the others, whereas model selection using Autometrics leads to the most inaccurate forecasts. The most striking feature of Table 2.22 is that the nonparametric forecasts, which need no filtering, are nevertheless on average distinctly more inaccurate than forecasts generated by any other model or method. It can also be noted that direct linear forecasts from linear AR models built on untransformed series have RMSFE ratios close to one, while no filtering has been necessary. The 'no change' forecasts are somewhat less accurate than the ones generated by QuickNet-selected models but better than Autometrics-ones.

Table 2.23 contains the average ranks of the models based on the absolute forecast errors. In this comparison forecasts from Autometrics-selected models have the highest ranks among the ANN models. In accordance with Tables 2.19-2.22, MBE is the best performing ANN selection method. However, none of them has a lower average rank than the linear AR model. The performance of the nonparametric forecasts is highly dependent on whether the models are built on differences or levels. In the former case the ranks are much lower than in the latter, which are by far the highest overall. This accords with the RMSFE results in Table 2.22.

Table 2.24 contains p-values of Wilcoxon's signed-rank test of absolute forecast errors for AR-filtered forecasts from models based on differenced unemployment series. The null hypothesis of equal means is mostly rejected (the

Direct	Hor.	AR	QN	MBE	Autom.	QN-SG	NP	NC
	1	0.9994	$2 \cdot 10^5$	1.007	1.302	1.048	1.514	1.109
NF	3	1.006	8.759	1.033	2760	1.068	1.507	1.167
	6	1.005	1.327	1.045	1.995	1.084	1.499	1.148
	12	1.008	29.72	1.056	3.201	1.040	1.354	1.055
	1	0.9994	1.064	1.007	1.148	1.048	1.514	
SW	3	1.006	1.08	1.033	1.195	1.068	1.503	
	6	0.999	1.138	1.039	1.261	1.084	1.495	
	12	1.005	1.063	1.051	1.202	1.04	1.332	
	1	0.9994	1.067	1.007	1.146	1.048	1.514	
AR	3	1.006	1.072	1.033	1.196	1.068	1.503	
	6	1.005	1.133	1.044	1.257	1.084	1.495	
	12	1.008	1.055	1.054	1.203	1.04	1.333	

Table 2.22: Average root mean square forecast error ratios for the direct forecasts of the unemployment series based on levels. NF: No Filter, SW: Swanson-White filter, AR: Insane forecasts replaced by linear autoregressive ones.

Rec Diff	Hor.	AR	QN	MBE	Autom.	QN-SG
	1	6.58	6.89	6.77	7.17	6.84
AR	3	11.1	11.9	11.2	12.2	12
	6	10.9	11.8	11	12.2	12
	12	11.1	11.7	11.3	12.2	11.9

Dir Diff	Hor.	AR	QN	MBE	Autom.	QN-SG	NP	NC
	1	6.58	6.89	6.77	7.17	6.84	6.52	5.98
AR	3	11	11.9	11.6	12.4	11.7	11.6	12.6
	6	10.8	12	11.3	12.3	11.9	11.2	12.9
	12	11.4	12	11.3	12.7	12.1	12.2	12.4

Rec Level	Hor.	AR	QN	MBE	Autom.	QN-SG
	1	6.36	6.91	6.45	7.22	6.89
AR	3	11	11.9	11.3	12.7	12.2
	6	11	12.2	11.3	12.8	12.2
	12	10.8	12.1	11.2	12.7	12.1

Dir Level	Hor.	AR	QN	MBE	Autom.	QN-SG	NP	NC
	1	6.36	6.91	6.45	7.22	6.89	9.23	5.98
AR	3	10.8	11.7	11.2	13.1	11.8	15.5	12.6
	6	10.5	11.9	11	13.6	12	15.7	12.9
	12	10.7	11.6	11.3	13.1	11.3	15	12.4

Table 2.23: Average ranks based on the absolute forecast errors. For each procedure for which forecasts are carried out recursively as well as directly the forecasts from the two alternatives are identical at the 1-month horizon. Hence, the comparison is only made across the direct forecasts at the 1-month horizon and by construction the ranks are the same for the recursive counterparts.

Recursive	Hor.	QN	MBE	Autom.	QN-SG
	1	1	0.99	1	1
AR	3	1	0.645	1	1
	6	1	0.962	1	1
	12	1	0.999	1	1

Direct	Hor.	AR	QN	MBE	Autom.	QN-SG	NP	NC
	1		1	0.990	1	1	0.332	1
AR	3	0.226	1	0.999	1	1	1	1
	6	0.367	1	0.999	1	1	0.991	1
	12	1	1	1	1	1	1	1

Table 2.24: p-values of the Wilcoxon signed-rank test for testing the null of the mean of the forecast errors from recursive forecasts of the unemployment series from the linear AR estimated on differences is equal to the mean of the corresponding forecast error from the model in each column of the table. The tests are carried out separately at each horizon and the alternative hypothesis is that the model in the table has a lower mean. Top panel: Models estimated recursively on differences. Bottom panel: Models estimated directly on differences.

p-value is close to unity) in favour of the linear recursive AR absolute forecast errors having a smaller mean than the ones from the other model. MBE forecasts three months ahead are the only ANN-exception. For direct linear three- and six-month forecasts the null hypothesis of equal means is not rejected either.

Table 2.25 contains the same information with the difference that the alternative model is built on levels instead of differences. In this case, the null hypothesis is never rejected for MBE-based direct models, but it turns out that at longest horizons, direct linear AR forecasts have smaller absolute errors than the recursive AR ones (the corresponding p-values in Table 2.25 are close to zero). The result for 12-month forecasts requires an explanation. In Table 2.22, the 12-month RMSFE ratio of the direct linear AR forecasts equals 1.008, which does not indicate superiority of these forecasts over the recursive linear ones. Even after filtering, the direct linear 12-month model generates, however, a couple of large absolute forecast errors. This has a considerable effect on the RMSFE but a lesser one on the signed-rank statistic, in which the size of a particular error weighs less than in the RMSFE. The direct 12-month AR forecasts do have a smaller RMSFE ratio than the other methods in Table 2.22, which is in accord with the information in Table 2.25.

Statistics on the size and composition of the ANN models for forecasting based on the differenced unemployment series can be found in Table 2.26. When forecasting recursively MBE generates the smallest models and Autometrics the largest ones. QuickNet and QN-SG lie in between. Most of the selected variables

Recursive	Hor.	QN	MBE	Autom.	QN-SG
	1	1	0.823	1	1
AR	3	1	0.999	1	1
	6	1	0.999	1	1
	12	1	0.723	1	1

Direct	Hor.	AR	QN	MBE	Autom.	QN-SG	NP	NC
	1	0.333	1	0.823	1	1	1	1
AR	3	0.204	1	0.764	1	1	1	1
	6	0.003	1	0.404	1	1	1	1
	12	$3 \cdot 10^{-6}$	0.990	0.148	1	0.828	1	1

Table 2.25: p-values of the Wilcoxon signed-rank test for testing the null of the mean of the forecast errors from recursive forecasts of the unemployment series from the linear AR estimated on levels is equal to the mean of the corresponding forecast error from the model in each column of the table. The tests are carried out separately at each horizon and the alternative hypothesis is that the model in the table has a lower mean. Top panel: Models estimated recursively on levels. Bottom panel: Models estimated directly on levels.

Recursive	Total	Linear	Nonlinear
QN	6.62	0.0569	6.56
MBE	3.66	0.115	3.55
AM	13.7	0.172	13.5
QN-SG	6.12	0.0569	6.06

MBE	Total	Linear	Nonlinear
1 mth	3.66	0.115	3.55
3 mths	2.88	0.295	2.59
6 mths	2.58	0.227	2.35
12 mths	2.49	0.0556	2.43

Table 2.26: Left panel: Average number of variables selected for the models gnerating recursive forecasts of the unemployment series based on differences. "Total" indicates total number of variables included, "Linear" indicates the number of linear units included, and "Nonlinear" gives the number of hidden units included. Right panel: Average number of variables selected for the direct forecasts of the unemployment based on differences by MBE.

are hidden units. The average size of the MBE-based direct models decreases slightly with the forecasting horizon. It appears that there is positive correlation with the size of the model and its forecasting performance. In Table 2.19 and 2.20 models selected by MBE have the smallest RMSFE ratios, whereas Autometrics-based models have the largest ones. Models chosen using QuickNet and QN-SG lie in the middle.

Table 2.27 contains the same statistics for models based on levels. QN-SG now produces the most parsimonious models when forecasting recursively, and even QuickNet-based models have a smaller average size than the ones chosen by MBE. The share of linear lags is now appreciably greater than in Table 2.26, and this is the case for all four procedures. Autometrics selects the largest direct models, whose average size is practically the same as it is in the models for recursive forecasting. It appears that in this case, leaving out lags does not affect

Recursive	Total	Linear	Nonlinear
QN	5.02	1	4.02
MBE	6.08	5.23	0.848
AM	13.6	1.42	12.2
QN-SG	2.98	1	1.98

MBE	Total	Linear	Nonlinear
1 mth	6.08	5.23	0.848
3 mths	6.03	5.31	0.723
6 mths	5.89	5.29	0.6
12 mths	5.12	4.34	0.782

Table 2.27: Left panel: Average number of variables selected for the models generating recursive forecasts of the unemployment series based on levels. "Total" indicates total number of variables included, "Linear" indicates the number of linear units included, and "Nonlinear" gives the number of hidden units included. Right panel: Average number of variables selected for the direct forecasts of the unemployment series based on levels by MBE.

the average size of the model.

The correlation between the size of the model and the accuracy of the forecasts is weaker than in the previous case. A look at Table 2.21 shows that MBE-selected models still have the smallest RMSFE ratios, although they do not have the smallest size. Note, however, that they contain the largest number of linear lags, which may have affected the outcome. The position of Autometrics is unchanged: largest models and largest RMSFE ratios.

All models for direct forecasting contain more linear terms when the models are in levels than when they are in differences. They share this feature with the corresponding models built for forecasting the CPI.

Individual countries

The average results for the unemployment series are indicative of the results for the individual countries. No large differences can be found on the country level. However, to illustrate that not all relative RMSFE ratios are close to unity we discuss a few individual country results. The tables with the relative RMSFE for each individual country can be found at `http://econ.au.dk/research/research-centres/creates/research/research-papers/supplementary-downloads/rp-2011-27/`.

For example the direct forecasts on the differences of the Italian unemployment series are around ten percent more precise than the recursive forecasts from the linear autoregression. Similarly to the average results MBE delivers the most accurate forecasts of the ANN procedures while Autometrics is the most imprecise in particular at the short horizons. In fact these two observations are both very stable across all eleven countries emphasizing the fact that the average results reflect the results for the individual countries rather well.

The direct forecasts on the levels of the German unemployment series are

another example of a country for which it is possible to outperform the recursive linear autoregressive model on the differences. These forecasts are also an instance of a series for which the SW-filter produces more accurate forecasts than the AR-filter. MBE is again the most successful nonlinear model.

2.6 Conclusions

In this paper we consider macroeconomic forecasting with a flexible nonlinear model, the single-hidden layer feedforward neural network model that is a universal approximator. We apply the idea of White (2006) of transforming the specification and estimation problem of this model to a linear model specification problem. This leads to a situation in which the number of candidate variables to choose among vastly exceeds the number of observations. Three modelling techniques, White's QuickNet among them, that can handle this difficulty are compared and the models selected are used for forecasting.

The benchmark in our forecast comparisons is, with one exception, the linear AR model with recursive forecasts. It turns out to be difficult to improve upon its forecasting precision using recursive forecasting, while the direct method seems to be a more successful approach. It appears that the Marginal Bridge Estimator of Huang et al. (2008) yields the best performing ANN models overall, but the results do vary from one country to the other. Autometrics of Doornik (2009) selects models with excellent forecasting performance when there is a well-fitting model to be discovered but does poorly when no potential model fits the data well. QuickNet selects models whose average forecasting performance lies between that of the two others. Parsimony plays a role since MBE often selects models with the fewest variables of the available alternatives. The purely nonparametric model generates relatively accurate forecasts for inflation series but is much less successful in forecasting unemployment rates. The performance of the models may also vary as a function of the forecasting horizon.

All three techniques often produce models that yield some very erroneous or 'insane' forecasts, which makes filtering them necessary. The two insanity filters considered in this paper perform almost equally well, although the AR filter may have a slight edge over the filter that Swanson and White (1995) introduced. Multicollinearity is the main reason for insane forecasts, and it might be a good idea to develop all three modelling strategies further in order to reduce the probability of the outcomes in which the final model contains very strongly linearly correlated variables.

We find that that testing linearity before variable selection does not help in choosing useful models. It may do so for certain countries and variables but may lead to weakened forecasting performance in some others. For this reason it cannot be recommended as a part of any of the three modelling strategies under consideration.

Forecasts are generated using both the recursive and the direct method. Overall, direct forecasting is somewhat superior to the recursive technique, but it does not dominate the latter. The results vary from one country and variable to the other. This is true also in comparing the accuracy of recursive and direct forecasts from linear AR models: on average direct forecasts are more accurate than the recusive ones.

When it comes to choosing between models based on first differences of the series and ones specified and estimated using levels it turns out that in forecasting the CPI, models built on levels tend to generate more accurate forecasts on average than the corresponding models constructed using first differences. It is not clear why that is the case. In forecasting unemployment rates the outcome is less clear: the models based on levels cannot be viewed as superior to models built on first differences.

A general conclusion is that the ANN model can be useful in macroeconomic forecasting, but that the linear AR model is a serious competitor. In practice, the forecaster may experiment with several models and methods before settling for one, if the final goal is to find a model with the best performance for a given country and variable. Another possibility left for further work would be to combine recursive and direct forcasts obtained with various linear AR and ANN models.

Finally, the purpose of this work has not been to compare the forecasting performance of different nonlinear models. Doing so in a satisfactory fashion would require a vast amount of resources. It would also shift the focus away from our main aim: comparing different modelling techniques for the single-hidden layer ANN model made possible by the work of White (2006), and has not been attempted here.

2.7 Appendix: Creating the Pool of Hidden Units

First we shall consider the procedure of White (2006). It consists of three steps.

1. Rewrite the argument of the logistic function in (2.1) as follows:

$$\gamma' \mathbf{z}_t = \gamma_0 + \gamma_1(\gamma_2' \widetilde{\mathbf{z}}_t)$$

using the notation: $\mathbf{z}_t = (1, \widetilde{\mathbf{z}}_t')'$. For convenience, assume that each element of $\widetilde{\mathbf{z}}_t$ has mean zero. The vector γ_2 is the direction vector whose length equals one, and it is selected first. This is done as follows. Let the random vector $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$. Then set $\gamma_2 = \mathbf{x}(\mathbf{x}'\mathbf{x})^{-1/2}$ which is uniformly distributed on the unit sphere $\mathcal{S}^{p-1}$ in $\mathbb{R}^p$.

2. Given γ_2, choose $\gamma_1 > 0$ such that it is at least of the magnitude of $\widehat{\sigma}_z = \text{std}(\gamma_2' \widetilde{\mathbf{z}}_t)$ with the range spanning modest multiples of $\widehat{\sigma}_z$. Draw γ_1 at random from this range. The scalar γ_1 gives the length of the vector γ_2 and controls the slope of the hidden unit as a function of $\gamma_2' \widetilde{\mathbf{z}}_t$.

3. Choose γ_0 such that it has mean zero and standard deviation comparable to $\text{std}(\gamma_1(\gamma_2' \widetilde{\mathbf{z}}_t))$. Draw γ_0 at random from this distribution. This scalar controls the location of the hidden unit.

In our experiments, selecting γ_1 as in step (2) above frequently led to values of this parameter that were too small in the sense that the hidden unit did not display sufficient variation in the sample. This had adverse consequences to the forecasts. To avoid them, we constructed a modification with the following structure:

1. Rewrite the argument of the logistic function in (2.1) as follows:

$$\gamma' \mathbf{z}_t = \gamma_1 / \hat{\sigma}_z(\gamma_2' \tilde{\mathbf{z}}_t - \gamma_0) \tag{2.9}$$

Choose γ_2' as described above in step (1) above.

2. Next obtain γ_0. Consider the values $x_t = \gamma_2' \tilde{\mathbf{z}}_t$, $t = 1, ..., T,$. Let $x_{\min}$ and $x_{\max}$ denote the minimum and maximum values of this sequence. Let $d = x_{\max} - x_{\min}$. Now draw γ_0 from a uniform distribution on $[x_{\min} + \delta d, x_{\max} - \delta d]$ for $\delta \in [0, 0.5]$. We choose $\delta = 0.1$. In this way we make sure that the hidden units are not centered at very small or large values of $\gamma_2' \tilde{\mathbf{z}}_t$. As a result of the parameterization (2.9), demeaning $\widetilde{\mathbf{z}}_t$ is not necessary.

3. Finally, the slope parameter γ_1 is chosen uniformly at random from the set $\left\{1.25^j : \ j = 0, 1, ..., 20\right\}$. Hence, the smallest possible value of γ_1 is 1 while the largest possible value is 87. The set is deliberately constructed to be denser for small values since the slope of the logistic function changes more for changes in γ_1 when γ_1 is small than when γ_1 is big. For large values of γ_1 changes in γ_1 will not affect the slope of the logistic function much and so it is less important to have a dense grid here.

The decisive difference between the two strategies lies in choosing γ_1. In the strategy of White (2006), γ_1 is not a scale-free parameter. That is, a change of units in $\tilde{\mathbf{z}}_t$ affects the set of possible slopes that can be selected, which is a disadvantage. In (2.9), γ_1 is a scale-free parameter due to the division by $\widehat{\sigma}_z$, for discussion, see for example Teräsvirta (1998). This makes it possible to define a reasonable range for this parameter. The minimum value of the scale-free γ_1 is set to unity in order to avoid logistic functions with too little sample variation.

2.8 Bibliography

Ahmed, N. K., A. F. Atiya, N. El Gayer, and H. El-Shishiny (2010). An empirical comparison of machine learning tools for time series forecasting. *Econometric Reviews 29*, 594–621.

Anders, U. and O. Korn (1999). Model selection in neural networks. *Neural Networks 12*, 309–323.

Candès, E. J. (1998). *Ridgelets: theory and applications*. PhD thesis, Dept. of Statistics, Stanford University.

Candès, E. J. (2003). Ridgelets: estimating with ridge functions. *The Annals of Statistics 31*, 1561–1599.

Costantini, M. and R. M. Kunst (2011). On the usefulness of the Diebold-Mariano test in the selection of prediction models: some Monte Carlo evidence. Working paper, University of Vienna.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems 2*, 303–314.

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

Doornik, J. A. (2009). Autometrics. In J. L. Castle and N. Shephard (Eds.), *The Methodology and Practice of Econometrics*, pp. 88–122. Oxford University Press, Oxford.

Fine, T. L. (1999). *Feedforward neural network methodology*. Springer Verlag, New York.

Gallant, A. R. (1984). The Fourier flexible form. *American Journal of Agricultural Economics 66*, 204–208.

Ghaddar, D. K. and H. Tong (1981). Data transformation and self-exciting threshold autoregression. *Applied Statistics 30*, 238–248.

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

Goffe, W. L., G. D. Ferrier, and J. Rogers (1994). Global optimization of statistical functions with simulated annealing. *Journal of Econometrics 60*, 65–99.

Hastie, T., R. Tibshirani, and J. H. Friedman (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Verlag, New York.

Hendry, D. F. and H. M. Krolzig (2005). The properties of automatic Gets modelling. *Economic Journal 115*, 32–61.

Hornik, K., M. Stinchcombe, and H. White (1989). Multilayer feedforward networks are universal approximators. *Neural Networks 2*, 359–366.

Huang, J., J. L. Horowitz, and S. Ma (2008). Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Annals of Statistics 36*, 587–613.

Hwang, J. T. G. and A. A. Ding (1997). Prediction intervals for artificial neural networks. *Journal of the American Statistical Association 92*, 748–757.

Kock, A. B. and T. Teräsvirta (2011). Forecasting with nonlinear time series models. In M. P. Clements and D. F. Hendry (Eds.), *Oxford Handbook of Economic Forecasting*, pp. 61–87. Oxford University Press, Oxford.

Krolzig, H. M. and D. F. Hendry (2001). Computer automation of general-to-specific model selection procedures. *Journal of Economic Dynamics and Control 25*, 831–866.

Kuan, C.-M. and T. Liu (1995). Forecasting exchange rates using feedforward and recurrent neural networks. *Journal of Applied Econometrics 10*, 347–364.

Lee, T.-H., H. White, and C. W. J. Granger (1993). Testing for neglected non-linearity in time series models: A comparison of neural network methods and alternative tests. *Journal of Econometrics 56*, 269–290.

Makridakis, S. and M. Hibon (2000). The M3-Competition: results, conclusions and implications. *International Journal of Forecasting 16*, 451–476.

Marcellino, M. (2002). Instability and non-linearity in the EMU. Discussion Paper No. 3312, Centre for Economic Policy Research.

Marcellino, M., J. H. Stock, and M. W. Watson (2006). A comparison of direct and iterated multistep AR methods for forecasting macroeconomic time series. *Journal of Econometrics 135*, 499–526.

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

Perez-Amaral, T., G. M. Gallo, and H. White (2003). A flexible tool for model building: the relevant transformation of the inputs network approach (RETINA). *Oxford Bulletin of Economics and Statistics 65*, 821–838.

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

Rech, G. (2002). Forecasting with artificial neural network models. SSE/EFI Working Paper Series in Economics and Finance 491, Stockholm School of Economics.

Simon, H. (1999). *Neural networks: a comprehensive foundation*. Prentice Hall, Upper Saddle River, NJ.

Stock, J. H. and M. W. Watson (1999). A comparison of linear and nonlinear univariate models for forecasting macroeconomic time series. In R. F. Engle

and H. White (Eds.), *Cointegration, Causality, and Forecasting: A Festschrift in Honour of Clive W.J. Granger*, pp. 1–44. Oxford University Press, Oxford.

Swanson, N. R. and H. White (1995). A model-selection approach to assessing the information in the term structure using linear models and artificial neural networks. *Journal of Business & Economic Statistics 13*, 265–275.

Swanson, N. R. and H. White (1997a). A model selection approach to real-time macroeconomic forecasting using linear models and artificial neural networks. *Review of Economics and Statistics 79*, 540–550.

Swanson, N. R. and H. White (1997b). Forecasting economic time series using flexible versus fixed specification and linear versus nonlinear econometric models. *International Journal of Forecasting 13*, 439–461.

Teräsvirta, T. (1998). Modeling economic relationships with smooth transition regressions. In A. Ullah and D. E. A. Giles (Eds.), *Handbook of Applied Economic Statistics*, pp. 507–552. Dekker, New York.

Teräsvirta, T. (2006). Forecasting economic variables with nonlinear models. In G. Elliott, C. W. J. Granger, and A. Timmermann (Eds.), *Handbook of Economic Forecasting*, Volume 1, pp. 413–457. Elsevier, North-Holland.

Teräsvirta, T., C. W. J. Granger, and D. Tjøstheim (2010). *Modelling Nonlinear Economic Time Series*. Oxford University Press, Oxford.

Teräsvirta, T., C. F. Lin, and C. W. J. Granger (1993). Power of the neural network linearity test. *Journal of Time Series Analysis 14*, 209–220.

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

White, H. (2006). Approximate nonlinear forecasting methods. In G. Elliott, C. W. J. Granger, and A. Timmermann (Eds.), *Handbook of Economic Forecasting*, Volume 1, pp. 459–512. Elsevier, Amsterdam.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin 1*, 80–83.

Zhang, G., B. E. Patuwo, and M. Y. Hu (1998). Forecasting with artificial neural networks: The state of the art. *International Journal of Forecasting 14*, 35–62.

Chapter 3

Oracle Efficient Variable Selection in Random and Fixed Effects Panel Data Models

Anders Bredahl Kock
Aarhus University and CREATES

Conditionally accepted in *Econometric Theory*

Abstract

This paper generalizes the results for the Bridge estimator of Huang et al. (2008) to linear random and fixed effects panel data models which are allowed to grow in both dimensions. In particular, we show that the Bridge estimator is oracle efficient. It can correctly distinguish between relevant and irrelevant variables and the asymptotic distribution of the estimators of the coefficients of the relevant variables is the same as if only these had been included in the model, i.e. as if an oracle had revealed the true model prior to estimation.

The author wishes to thank Svend Erik Graversen, Niels Haldrup, Michael Jansson, Jørgen Hoffmann-Jørgensen, Adrian Pagan, Timo Teräsvirta, Allan Würtz, the co-editor and an anonymous referee for help, comments and suggestions. Also thanks to Joel Horowitz for responding quickly to my emails. Finally, I would like to thank the committee consisting of Henning Bunzel, Dick van Dijk and Jurgen Doornik for their careful reading and constructive comments. All errors and shortcomings are my responsibility. Financial support from CREATES funded by the Danish National Research Foundation is gratefully acknowledged.

In the case of more explanatory variables than observations we prove that the Marginal Bridge estimator can asymptotically correctly distinguish between relevant and irrelevant explanatory variables if the error terms are Gaussian. Furthermore, a partial orthogonality condition of the same type as in Huang et al. (2008) is needed to restrict the dependence between relevant and irrelevant variables.

3.1 Introduction

When building a model one of the first steps is to decide which variables to include. Sometimes theory can guide the researcher towards a set of potential explanatory variables but which variables in this set are relevant and which are to be left out? Huang et al. (2008) showed that the Bridge estimator is able to discriminate between relevant and irrelevant explanatory variables in a cross section setting with fixed covariates whose number is allowed to increase with the sample size. In fact, oracle efficient estimation has received quite some attention in the statistics literature in the recent years, see (among others) Zou (2006), Candes and Tao (2007), Fan and Lv (2008), and Meinshausen and Yu (2009). However, we are not aware of any similar results for panel data models. For the case of fewer explanatory variables than observations we show that the oracle efficiency of the Bridge estimator carries over to linear panel data models with random regressors in the random and fixed effects settings. More precisely, it suffices that either the number of cross sectional units (N) or the number of observations within each cross sectional unit (T_N) goes to infinity in order to establish consistency and correct elimination of irrelevant variables. To obtain the oracle efficient asymptotic distribution (the distribution obtained by only including the relevant covariates) of the estimators of the nonzero coefficients, further restrictions are needed. In the classical setting of fixed T_N and large N these restrictions are satisfied. Further sufficient conditions for oracle efficiency are given. By fixing T_N and the number of covariates we obtain as a corollary that the asymptotic distribution of the estimators of the non-zero coefficients is exactly the classical fixed effects or random effects limit law.

If the set of potential explanatory variables is larger than the number of observations we show that the Marginal Bridge estimator of Huang et al. (2008) can be used to distinguish between relevant and irrelevant variables in random and fixed effects panel data models. A partial orthogonality condition restricting the dependence between the relevant and the irrelevant variables of the same type as

in Huang et al. (2008) is imposed. Furthermore, the error terms must be Gaussian – a price paid for letting the covariates be random. The random covariates also rendered the maximum inequalities based on exponential Orlicz norms used in Huang et al. (2008) inapplicable. However, more simple maximum inequalities in L^q spaces can still be applied but the result is that the number of irrelevant variables must be $o(N^{q/2})$ for some $q \geq 1$ (this is for fixed T_N for comparability to the known cross sectional results) as opposed to $\exp(o(N))$ (a subexponential rate). Since q is arbitrary this still allows the number of irrelevant variables to increase at any polynomial rate. The number of relevant variables may still be $o(N^{1/2})$ (again T_N is considered fixed for comparison).

Furthermore, the Marginal Bridge estimator is very fast to implement which also makes it useful as an initial screening device to weed out the most irrelevant variables before initiating the actual modeling stage.

Since cross section data can be viewed as panel data with only one observation per individual, all our results are also valid for cross section data and hence generalize the results for these.

The plan of the paper is as follows. Section 3.2 puts forward the general framework. Section 3.3 introduces the Bridge estimator and its properties while Section 3.4 discusses the Marginal Bridge estimator. Section 3.5 illustrates the results by simulation and Section 3.6 concludes. Section 3.7 contains proofs of the propositions.

3.2 Setup and Assumptions

Consider the following linear panel data model.

$$\tilde{y}_{it} = \tilde{\mathbf{x}}_{it}'\beta_0 + c_i + \tilde{\varepsilon}_{it},\ i = 1,...,N,\ t = 1,...,T_N \tag{3.1}$$

$\tilde{\mathbf{x}}_{it}$ is a $p_N \times 1$ vector of covariates indicating that the number of covariates is allowed to increase with the sample size. The interpretation of (3.1) is that N individuals are observed in T_N time periods, totaling NT_N observations. The c_i indicate the unobserved heterogeneity, i.e. unobserved time invariant variables such as intelligence of an individual or start up capital of a firm. The $\tilde{\varepsilon}_{it}$ are the idiosyncratic error terms. Some of the elements of β_0 may be zero. It is our objective to locate these while still estimating the nonzero coefficients consistently.

N as well as T_N are allowed to tend to infinity. However, all results are valid as long as N tends to infinity. Hence, the traditional large N, fixed T_N setting is

covered. Notice that T_N is indexed by N. Some of our results put no restrictions on how T_N depends on N.

Equation (3.1) can equivalently be written as

$$\tilde{\mathbf{Y}}_{iN} = \tilde{\mathbf{X}}_{iN}\beta_0 + \mathbf{c}_{iN} + \tilde{\boldsymbol{\varepsilon}}_{iN},\ i = 1,...,N, \tag{3.2}$$

where $\tilde{\mathbf{Y}}_{iN} = (\tilde{y}_{i1},...,\tilde{y}_{iT_N})'$, $\tilde{\mathbf{X}}_{iN} = (\tilde{\mathbf{x}}_{i1},...,\tilde{\mathbf{x}}_{iT_N})'$, $\tilde{\boldsymbol{\varepsilon}}_{iN} = (\tilde{\varepsilon}_{i1},...,\tilde{\varepsilon}_{iT_N})'$, $\mathbf{c}_{iN} = c_{iN}\iota_{T_N}$, $\iota'_{T_N} = (1,...,1)$, $i = 1,...,N$.

Fixed Effects.

In the fixed effects setting one assumes:

(FE1) Random sampling: $(\tilde{\mathbf{X}}_{iN}, \mathbf{c}_{iN}, \tilde{\boldsymbol{\varepsilon}}_{iN})_{i=1}^N$ is i.i.d.

(FE2) $E\left(\tilde{x}_{itl}^4\right),\ E\left(\tilde{\varepsilon}_{it}^4\right) < \infty,\ i = 1,...,N,\ t = 1,...,T_N,\ l = 1,...,p_N$

(FE3) a) $E(\tilde{\boldsymbol{\varepsilon}}_{iN}|\tilde{\mathbf{X}}_{iN}, \mathbf{c}_{iN}) = 0$ and b) $E(\tilde{\boldsymbol{\varepsilon}}_{iN}\tilde{\boldsymbol{\varepsilon}}'_{iN}|\tilde{\mathbf{X}}_{iN}, \mathbf{c}_{iN}) = \sigma^2\mathbf{I}_{T_N},\ i = 1,...,N$

where $\tilde{x}_{itl}$ is the l'th covariate of individual i in period t. For our proofs we may replace (FE3) by $E(\tilde{\boldsymbol{\varepsilon}}_{iN}|\tilde{\mathbf{X}}_{iN}) = 0$ and $E(\tilde{\boldsymbol{\varepsilon}}_{iN}\tilde{\boldsymbol{\varepsilon}}'_{iN}|\tilde{\mathbf{X}}_{iN}) = \sigma^2\mathbf{I}_{T_N}$ which is less restrictive but since (FE3) is standard in the literature we stick to this. Next, we carry out the forward orthogonal deviations transform of Arellano (2003), (page 17). This transformation removes the unobserved heterogeneity while keeping the error terms uncorrelated. In particular, define the $(T_N - 1) \times T_N$ matrix

$$\mathbf{D} = \begin{pmatrix} -1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & & 0 & 0 \\ \vdots & & & \ddots & & \vdots \\ 0 & 0 & 0 & \cdots & -1 & 1 \end{pmatrix}$$

and multiply (3.2) through by $\left(\mathbf{D}\mathbf{D}'\right)^{-1/2}\mathbf{D}$ to get

$$\mathbf{Y}_{iN} = \mathbf{X}_{iN}\beta_0 + \boldsymbol{\varepsilon}_{iN},\ i = 1,...,N, \tag{3.3}$$

where $\mathbf{Y}_{iN} = \left(\mathbf{D}\mathbf{D}'\right)^{-1/2}\mathbf{D}\tilde{\mathbf{Y}}_{iN}$, $\mathbf{X}_{iN} = \left(\mathbf{D}\mathbf{D}'\right)^{-1/2}\mathbf{D}\tilde{\mathbf{X}}_{iN}$ and $\boldsymbol{\varepsilon}_{iN} = \left(\mathbf{D}\mathbf{D}'\right)^{-1/2}\mathbf{D}\tilde{\boldsymbol{\varepsilon}}_{iN}$. Clearly, (FE3) implies

(FE3') a) $E(\boldsymbol{\varepsilon}_{iN}|\mathbf{X}_{iN}) = 0$ and b) $E(\boldsymbol{\varepsilon}_{iN}\boldsymbol{\varepsilon}'_{iN}|\mathbf{X}_{iN}) = \sigma^2\mathbf{I}_{T_N-1}$

which is what will be used in the proofs. Arellano (2003) gives the specific form of the entries of $\left(\mathbf{DD}'\right)^{-1/2}\mathbf{D}$. The number of time series observations for each individual is reduced from T_N to $T_N - 1$ by the forward orthogonal deviations transform. However, for notational convenience, we will keep using T_N for the number of time series observations in the transformed model. In a cross section setting this transform does not need to be carried out.

Assumption (FE3) b) $E(\tilde{\varepsilon}_{iN}\tilde{\varepsilon}_{iN}'|\tilde{\mathbf{X}}_{iN}, \mathbf{c}_{iN}) = \sigma^2\mathbf{I}_{T_N}$ restricts the $\tilde{\varepsilon}_{it}$ to be uncorrelated. This may be relaxed to $E(\tilde{\varepsilon}_{iN}\tilde{\varepsilon}_{iN}'|\tilde{\mathbf{X}}_{iN}, \mathbf{c}_{iN}) = \mathbf{S}$ where $\mathbf{S}$ is a known covariance matrix. In this case the forward orthogonal deviations transform is replaced by $\left(\mathbf{DD}'\right)^{-1/2}\mathbf{DS}^{-1/2}$. So heteroskedasticity can be handled as long as the structure of it is known – the transformation applied simply changes accordingly. If the heteroskedasticity is ignored the situation is more subtle and we will discuss this in more detail in Section 3.3.

Random Effects.

In the random effects setting (FE1)-(FE3) are maintained while

(RE4) a) $E\left(\mathbf{c}_{iN}|\tilde{\mathbf{X}}_{iN}\right) = 0$, b) $E\left(\mathbf{c}_{iN}\mathbf{c}_{iN}'|\tilde{\mathbf{X}}_{iN}\right) = \sigma_c^2\iota_{T_N}\iota_{T_N}'$ and c) σ and σ_c are known and finite[1]

are added. Part a) of this extra assumption restricts the dependence between $\mathbf{X}_{iN}$ and $\mathbf{c}_{iN}$ sufficiently in order to allow merging the latter with the error term while still being able to prove the desired results. Part b) specifies the conditional covariance structure of $\mathbf{c}_i$. RE4c) is needed to enable us to carry out the GLS transform below. The gain from these stronger assumptions is that they (as opposed to fixed effects) allow for the inclusion of a covariate which is constant over time and only varies over individuals. Defining $\mathbf{v}_{iN} = \mathbf{c}_{iN} + \tilde{\varepsilon}_{iN}$, (FE3) and (RE4) imply $E\left(\mathbf{v}_{iN}|\tilde{\mathbf{X}}_{iN}\right) = 0$ and

[1] In principle it suffices for most purposes that ratio σ_c/σ is known but it is hard to imagine a situation where the ratio is known while σ_c and σ are not. I wish to thank the co-editor for pointing this out.

$$E(\mathbf{v}_{iN}\mathbf{v}_{iN}'|\tilde{\mathbf{X}}_{iN}) = E([\mathbf{c}_{iN}+\tilde{\boldsymbol{\varepsilon}}_{iN}][\mathbf{c}_{iN}+\tilde{\boldsymbol{\varepsilon}}_{iN}]'|\tilde{\mathbf{X}}_{iN})$$
$$= \begin{pmatrix} \sigma_c^2+\sigma^2 & \sigma_c^2 & \cdots & \sigma_c^2 \\ \sigma_c^2 & \sigma_c^2+\sigma^2 & \cdots & \vdots \\ \vdots & & \ddots & \sigma_c^2 \\ \sigma_c^2 & & & \sigma_c^2+\sigma^2 \end{pmatrix} = \Omega$$

The presence of the unobserved heterogeneity renders the error terms correlated. Since the structure of the correlation is known, the correlation is easily removed by premultiplying (3.2) by[2] $\sigma\Omega^{-1/2}$ (GLS transform). This yields

$$\mathbf{Y}_{iN} = \mathbf{X}_{iN}\beta_0 + \boldsymbol{\varepsilon}_{iN},\ i = 1,...,N, \tag{3.4}$$

where $\mathbf{Y}_{iN} = \sigma\Omega^{-1/2}\tilde{\mathbf{Y}}_{iN}$, $\mathbf{X}_{iN} = \sigma\Omega^{-1/2}\tilde{\mathbf{X}}_{iN}$ and $\boldsymbol{\varepsilon}_{iN} = \sigma\Omega^{-1/2}\mathbf{v}_{iN}$. Hence,

(RE3') a) $E(\boldsymbol{\varepsilon}_{iN}|\mathbf{X}_{iN}) = 0$ and b) $E(\boldsymbol{\varepsilon}_{iN}\boldsymbol{\varepsilon}_{iN}'|\mathbf{X}_{iN}) = \sigma^2\mathbf{I}_{T_N}$

which is what will be used in the proofs. In a cross section setting the random effects transform does not need to be carried out. As was the case in the fixed effects setting, any known heteroskedasticity structure can be handled in the random effects setting as well, as long as Ω is known.

3.3 The Bridge estimator

The Bridge estimator estimates β_0 by minimizing

$$L_N(\beta) = \sum_{i=1}^{N}\sum_{t=1}^{T_N}\left(y_{it} - \mathbf{x}_{it}'\beta\right)^2 + \lambda_N\sum_{k=1}^{p_N}|\beta_k|^\gamma \tag{3.5}$$
$$= \sum_{j=1}^{NT_N}\left(y_j - \mathbf{x}_j'\beta\right)^2 + \lambda_N\sum_{k=1}^{p_N}|\beta_k|^\gamma, \quad \gamma > 0 \tag{3.6}$$

where summation from 1 to NT_N indicates summation over all time periods for each individual (So the first T_N terms in the sum correspond to all T_N observation

[2]The sole reason for multiplying $\Omega^{-1/2}$ by σ is that (FE3') and (RE3') become identical except for the dimension of the covariance matrix. Since (FE3') and (RE3') are the assumptions used in the proofs this indicates that the proofs only have to be carried out in either the fixed or the random effects setting.

on individual 1, the next T_N terms to all observations on individual 2 and so on. This convention is adopted in the sequel.). The Bridge estimator, denoted $\hat{\beta}_N$, may hence be seen as a sort of penalized/regularized least squares. The objective function consists of two parts; the first part being the least squares objective function and the second part penalizing parameters different from 0. The larger λ_N, the larger the penalty. For $\gamma = 1$ the minimizer of (3.6) could be called the LASSO panel estimator, (Tibshirani (1996)). For $\gamma = 2$ it could be called the ridge regression estimator (Tikhonov regularization) for panel data models. In a cross sectional setting the ridge regression is frequently used to deal with multicollinearity. The Tikhonov regularization is more generally used to solve ill-conditioned (singular) overdetermined systems of linear equations.

Let β_0 denote the true value of β where the dependence on N is suppressed as in Huang et al. (2008). Partition β_0 as $\beta_0 = (\beta_{10}', \beta_{20}')'$ where $\beta_{10} \neq 0$ is $k_N \times 1$ and $\beta_{20} = 0$ is $m_N \times 1$. Hence, the β_{10} are the coefficients corresponding to the relevant variables denoted $\mathbf{w}_{it}$. β_{20} are the coefficients of the irrelevant variables denoted $\mathbf{z}_{it}$. So $\mathbf{x}_{it}$ is partitioned as $\mathbf{x}_{it} = (\mathbf{w}_{it}', \mathbf{z}_{it}')'$. Accordingly, we define $\mathbf{X}_N = \left(\mathbf{x}_{11}, ..., \mathbf{x}_{NT_N}\right)'$, $\mathbf{W}_N = \left(\mathbf{w}_{11}, ..., \mathbf{w}_{NT_N}\right)'$ and $\mathbf{Z}_N = \left(\mathbf{z}_{11}, ..., \mathbf{z}_{NT_N}\right)'$. $\Sigma_N = (NT_N)^{-1}\mathbf{X}_N'\mathbf{X}_N$ as well as $\Sigma_{1N} = (NT_N)^{-1}\mathbf{W}_N'\mathbf{W}_N$ are the scaled Gram matrices of $\mathbf{X}_N$ and $\mathbf{W}_N$, respectively. Let ρ_{1N} and ρ_{2N} be the smallest and the largest eigenvalue of Σ_N. Similarly, define τ_{1N} and τ_{2N} as the smallest and the largest eigenvalue of Σ_{1N}. Set $\mathbf{W}_{iN} = \left(\mathbf{w}_{i1}, ..., \mathbf{w}_{iT_N}\right)'$ and for $\mathbf{x} \in \mathbf{R}^p$ $\|\mathbf{x}\| = \sqrt{\sum_{k=1}^p x_k^2}$ denotes the Euclidean norm on $\mathbf{R}^p$ stemming from the dot product. Finally, $\mathbf{x}_k = (x_{1,k}, ..., x_{NT_N,k})'$ is the vector containing all observations of the k'th explanatory variable.

Next, we state and discuss the assumptions needed to establish consistency and oracle efficiency of Bridge estimators in random and fixed effects panel data models. Notice how N and T_N enter symmetrically indicating that what matters is their product, i.e. the total number of observations, and not whether it is N or T_N which gets large (however, some theorems require further assumptions restricting the rate at which T_N increases relative to N).

(A1) $\frac{1}{NT_Np_N}\sum_{i=1}^N\sum_{t=1}^{T_N}\sum_{k=1}^{p_N} x_{itk}^2$ is bounded in L^1, i.e.,

$$\sup_{1\leq N<\infty} E\left(\frac{1}{NT_Np_N}\sum_{i=1}^N\sum_{t=1}^{T_N}\sum_{k=1}^{p_N} x_{itk}^2\right) = \sup_{1\leq N<\infty}\frac{1}{T_Np_N}\sum_{t=1}^{T_N}\sum_{k=1}^{p_N} E\left(x_{1tk}^2\right) = K < \infty$$

(A2) There exist constants $0 < \tau_1 < \tau_2 < \infty$ such that $\tau_1 \leq \tau_{1N} \leq \tau_{2N} \leq \tau_2$ *almost surely*

(A3) $\lambda_N(k_N/(NT_N))^{1/2} \to 0$

(A4) $\lambda_N \rho_{1N}^{2-\gamma}(NT_N)^{-\gamma/2} p_N^{\gamma/2-1} \to \infty$ *almost surely*

(A5) There exist constants $0 < b_0 < b_1 < \infty$ such that $b_0 \leq \min\left\{\left|\beta_{10j}\right| \ |1 \leq j \leq k_N\right\} \leq \max\left\{\left|\beta_{10j}\right| \ |1 \leq j \leq k_N\right\} \leq b_1$

(A6) $(p_N + \lambda_N k_N)/(NT_N\rho_{1N}) \to 0$ *almost surely*

(A7) $\frac{\rho_{1N}\rho_{2N}^{1/2}}{p_N^{1/2}} \in O_p(1)$

Assumption (A1) may be dropped altogether if the covariates are normalized as $\frac{1}{NT_N}\sum_{j=1}^{NT_N} x_{jk}^2 = \frac{1}{NT_N}\sum_{i=1}^{N}\sum_{t=1}^{T_N} x_{itk}^2 = 1$ for all $1 \leq k \leq p_N$. Alternatively, (A1) is satisfied if $\{x_{itk}\}$ is bounded in L^2 – this is in turns satisfied if, e.g, the covariates are uniformly bounded or if T_N and p_N are fixed constants. If the covariates are identically distributed over time, then the assumption reduces to boundedness of the Cesàro sum $\frac{1}{p_N}\sum_{k=1}^{p_N} E\left(x_{11k}^2\right)$. Finiteness of p_N or convergence of $\left\{E\left(x_{11k}^2\right)\right\}_{k=1}^{\infty}$ are sufficient for this. Finally, it may be noted that convergence of $\frac{1}{T_N p_N}\sum_{t=1}^{T_N}\sum_{k=1}^{p_N} E\left(x_{1tk}^2\right)$ is also sufficient for the desired boundedness in L^1.

Huang et al. (2008) mention that assumption (A2) is likely to be satisfied in sparse systems, where k_N is relatively small.

Regarding condition (A3) one notices that if the number of relevant covariates k_N stays fixed $\lambda_N/(NT_N)^{1/2} \to 0$. Hence, $\lambda_N \in o((NT_N)^{1/2})$.

Assumption (A4): Assume $0 < a_1 < \rho_{1N} \leq \rho_{2N} < a_2 < \infty$ for some constants a_1 and a_2 and that the number of covariates stays constant. Then it must be the case that $\lambda_N(NT_N)^{-\gamma/2} \to \infty$. This excludes $\gamma \geq 1$ by (A3). Hence, $0 < \gamma < 1$ and $\lambda_N \in o((NT_N)^{1/2}) \cap \omega((NT_N)^{\gamma/2})$ where $\omega(g(N))$ is the set of functions that diverge to infinity when divided by $g(N)$ as $N \to \infty$.

Assumption (A5) requires that the non-zero coefficients are uniformly bounded away from 0 and infinity. This is trivially satisfied if the number of covariates is finite. Also note that all results remain valid (with slight modifications) if b_1 is replaced by a sequence b_{1N} which is allowed to tend to infinity.

By assumption (A3), assumption (A6) is satisfied if $0 < a_1 < \rho_{1N} < \rho_{2N} < a_2 < \infty$ for some constants a_1 and a_2 and the number of covariates is finite. Since the Gramian Σ_N is positive semidefinite (A6) also implies that $\rho_{1N} > 0$ in order for the condition to be well defined. This excludes $p_N > NT_N$ since the rank of Σ_N can be no larger than NT_N.

Assumption (A7) is satisfied if $0 < a_1 < \rho_{1N} < \rho_{2N} < a_2 < \infty$ for some constants a_1 and a_2.

Assumptions (A2)-(A6) are identical[3] to assumptions made in Huang et al. (2008). (A1) and (A7) are not made by Huang et al. (2008) but both these assumptions are redundant if the covariates are normalized as $\frac{1}{NT_N}\sum_{j=1}^{NT_N} x_{jk}^2 = 1$ for all $1 \leq k \leq p_N$ as done by these authors.

Our first theorem states that the Bridge estimator is consistent in the random as well as the fixed effects setting. Throughout we will assume that (FE1)-(FE3) (fixed effects setting) or (FE1)-(FE3) and (RE4) (random effects setting) are satisfied.

Theorem 1. *Let $\hat{\beta}_N$ denote the minimizer of (3.6). Suppose that $\gamma > 0$ and that conditions (A1), (A3), (A5), and (A6) hold. Then $||\hat{\beta}_N - \beta_N|| \in O_p(\min(h_N, h'_N))$ where $h_N = \rho_{1N}^{-1}(p_N/(NT_N))^{1/2}$ and $h'_N = \left[(p_N + \lambda_N k_N)/(NT_N\rho_{1N})\right]^{1/2}$.*

Theorem 1 shows the consistency of the Bridge estimator by assumption (A6). By considering h_N it follows that if there exists a constant a_1 such that $0 < a_1 < \rho_{1N}$ and p_N is constant the Bridge estimator converges at the same rate as the least squares estimator. The faster the arrival rate of new explanatory variables (p_N increases) the slower the rate of convergence of the Bridge estimator since h_N as well as h'_N are increasing in p_N. If ρ_{1N} tends to 0 (approaching a singular design) the convergence rate is also slowed down. It is also seen that N and T_N enter symmetrically. This is not immediate on the outset since only independence of $\{\mathbf{X}_{iN}\}_{i=1}^{\infty}$ has been assumed while the T_N rows of each $\mathbf{X}_{iN}$ may have any dependence structure between them. What provides the result is that $E(\varepsilon_{iN}\varepsilon'_{iN}|\mathbf{X}_{iN}) = \sigma^2\mathbf{I}_{T_N}$, i.e. the conditional uncorrelatedness of the error terms of each individual. This underscores the importance of orthogonalizing (in L^2) the error terms prior to applying the Bridge estimator. In the presence of unknown arbitrary heteroskedasticity, $E(\varepsilon_{iN}\varepsilon'_{iN}|\mathbf{X}_{iN}) = \mathbf{S}$, (which can not be orthogonalized) we were only able to prove that $||\hat{\beta}_N - \beta_N|| \in O_p(h'_N + (T_N - 1)/\rho_{1N})$. So even in the situation of fixed T_N, consistency requires $\rho_{1N} \to \infty$ which is impossible.

The next theorem reveals that the Bridge estimator performs variable selection and gives the limiting law of the estimator of the nonzero coefficients.

Let $U_{1N} = \alpha' \left(E\left[\frac{1}{T_N}\mathbf{W}'_{1N}\mathbf{W}_{1N}\right]\right)^{-1} T_N^{-1/2}\mathbf{W}'_{1N}\varepsilon_{1N}$.

[3]Since we allow for random covariates some of our assumptions must hold in an almost sure sense while the equivalent assumptions in Huang et al. (2008) must hold surely.

Theorem 2. *Assume* $0 < \gamma < 1$. *Then under (A1)-(A7),*

(i) $\hat{\beta}_{2N} = 0$ *with probability converging to 1.*

(ii) Let k_N *be a fixed number k,* α *be a* $k \times 1$ *vector, and* $s_N = \sqrt{\sigma^2 \alpha' \Sigma_{1N}^{-1} \alpha}$. *If* $\left\{U_{1N}^2\right\}_{N=1}^{\infty}$ *is uniformly integrable,*

$$\frac{\max_{1 \le t \le T_N} Var\left(w_{1tl} w_{1tm}\right)}{N} \to 0 \text{ for all } 1 \le l, m \le k$$

and

$$\lim_{N \to \infty} \frac{1}{NT_N} \sum_{j=1}^{NT_N} E\left(\mathbf{w}_j \mathbf{w}_j'\right) = \lim_{N \to \infty} E\left(\frac{1}{T_N} \mathbf{W}_{1N}' \mathbf{W}_{1N}\right)$$

exists then,

$$(NT_N)^{1/2} s_N^{-1} \alpha' \left(\hat{\beta}_{1N} - \beta_{10}\right) \xrightarrow{d} N(0,1)$$

Part (i) states that not only does $\hat{\beta}_{2N} \to 0$ in probability (Theorem 1), the Bridge estimator actually sets $\hat{\beta}_{2N} = 0$ with probability converging to 1. The latter of course implies the former while the converse is not true. The fact that $\hat{\beta}_{2N}$ is set exactly equal to 0 with probability converging to 1 means that the Bridge estimator performs variable selection.

Part (ii) of the theorem states that the asymptotic distribution of the estimators of the non zero coefficients is the same as if the true model had been known in advance – i.e. as if an oracle had revealed which variables to include and which to exclude. This is a very useful result in practice. One simply includes the whole set of potential explanatory variables. The irrelevant ones will be kicked out ($\hat{\beta}_{2N} = 0$ with probability converging to 1) while the relevant ones are estimated with the same asymptotic efficiency as if the irrelevant ones had been left out from the outset. However, notice that the price paid for letting the covariates be random is that k_N must be fixed. Alternatively, one may continue to let k_N increase in N while conditioning on the covariates and establish the limiting law along the lines of Huang et al. (2008).

Next we discuss conditions under which the requirements of part (ii) of Theorem 2 hold. The following Theorem gives sufficient conditions under which $\left\{U_{1N}^2\right\}_{N=1}^{\infty}$ is uniformly integrable.

Theorem 3. $\left\{U_{1N}^2\right\}_{N=1}^{\infty}$ *is uniformly integrable if either of the following conditions is satisfied.*

(i) $T_N = T$ *for a fixed* T

(ii) The rows in $\mathbf{W}_{1N}$ *are identically distributed and* $\mathbf{W}_{iN} \perp\!\!\!\perp \varepsilon_{iN},\ i = 1,...,N.$

(iii) $\mathbf{W}_{1N}$ *and* ε_{1N} *are uniformly bounded in* N.

The assumption $\max_{1\leq t\leq T_N} Var\left(w_{1tl}w_{1tm}\right)/N \to 0$ for all $1 \leq l,m \leq k$ in part (ii) of Theorem 2 is not restrictive. It is clearly satisfied if T_N is fixed. It is also satisfied if $\max_{1\leq t\leq T_N} E\left(\left[w_{1tl}w_{1tm}\right]^2\right) \leq M < \infty$ for all T_N and $1 \leq l,m \leq k$ (second moments uniformly bounded in t) which in turn is satisfied if the variables themselves are uniformly bounded in t. The assumption is also satisfied if $\mathbf{w}_{1t}$ are identically distributed across t. If the variances are linearly increasing, i.e. $Var\left(w_{1tl}w_{1tm}\right) = a_{lm}t$ for some $a_{lm} > 0$, it suffices that $T_N/N \to 0$.[4]

If T_N is fixed, $\lim_{N\to\infty} \frac{1}{NT_N}\sum_{j=1}^{NT_N} E\left(\mathbf{w}_j\mathbf{w}_j'\right) = \lim_{N\to\infty} E\left(\frac{1}{T_N}\mathbf{W}_{1N}'\mathbf{W}_{1N}\right)$ exists. The same is true if $\mathbf{w}_{1t}$ is identically distributed across t.

Part (ii) of Theorem 2 is made more precise in the following corollary which considers the classical situation of fixed T_N. Let $\tilde{\mathbf{W}}_1$ denote the matrix containing the k untransformed relevant variables of individual 1 in all time periods and $\ddot{\tilde{\mathbf{W}}}_1$ its column demeaned version.

Corollary 1. *Under the assumptions of Theorem 2,* T_N *fixed*

(i) and (FE1)-(FE3) and the forward orthogonal deviations transform

$$N^{1/2}\left(\hat{\beta}_{1N} - \beta_{10}\right) \xrightarrow{d} N\left(0, \sigma^2\left[E\left(\ddot{\tilde{\mathbf{W}}}_1'\ddot{\tilde{\mathbf{W}}}_1\right)\right]^{-1}\right) \tag{3.7}$$

(ii) and (FE1)-(FE3), (RE4) and the GLS transform

$$N^{1/2}\left(\hat{\beta}_{1N} - \beta_{10}\right) \xrightarrow{d} N\left(0, \sigma^2\left[E\left(\tilde{\mathbf{W}}_1'\Omega^{-1}\tilde{\mathbf{W}}_1\right)\right]^{-1}\right) \tag{3.8}$$

Notice that the asymptotic distribution in (3.7) is the same as for a fixed effects estimator with *known* sparsity pattern of β_0. This underscores the oracle property of the panel Bridge estimator. Similarly, (3.8) is the asymptotic distribution of the random effects estimator with *known* sparsity pattern of β_0.

[4]More generally, if $Var\left(w_{1tl}w_{1tm}\right) \in O\left(g(t)\right)$ for all $1 \leq l,m \leq k$ for some positive increasing function g it suffices that $\frac{g(T_N)}{N} \to 0$.

3.4 The Marginal Bridge estimator

Since the Bridge estimator is not applicable when $p_N > NT_N$ (though it does allow $p_N \to \infty$) a different approach is needed for this situation. As in Huang et al. (2008) we will employ the Marginal Bridge estimator which estimates β_0 by minimizing

$$U_N(\beta) = \sum_{k=1}^{p_N} \sum_{j=1}^{NT_N} \left(y_j - x_{jk}\beta_k\right)^2 + \lambda_N \sum_{k=1}^{p_N} |\beta_k|^\gamma \tag{3.9}$$

$$= \sum_{k=1}^{p_N} \left(\sum_{j=1}^{NT_N} \left(y_j - x_{jk}\beta_k\right)^2 + \lambda_N |\beta_k|^\gamma \right) \tag{3.10}$$

From (3.10) it is clear that the objective function is nothing else than the sum of the marginal objective functions for each variable – hence the name Marginal Bridge estimator. Let $\tilde{\beta}_N$ denote the minimizer of (3.10). We show that the Marginal Bridge estimator is able to correctly distinguish between relevant and irrelevant variables even when there are more explanatory variables than observations ($p_N > NT_N$). Maintain (FE1), assume $\tilde{\varepsilon}_{iN}$ is normally distributed for all $1 \leq i \leq N$ and replace (FE2) and (FE3) by

(FE2MB) $\left(\tilde{\mathbf{X}}_{iN}, \mathbf{c}_{iN}\right) \perp\!\!\!\perp \tilde{\varepsilon}_{iN}$, with $E\left(\tilde{\varepsilon}_{it}\right) = 0$ and $E\left(\tilde{\varepsilon}_{iN}\tilde{\varepsilon}_{iN}'\right) = \sigma^2 \mathbf{I}_{T_N}$[5].

(FE2MB) clearly implies (FE3) while the reverse need not be the case (see e.g. Stoyanov (1997) for an example). However, this strengthening is not likely to be of any practical importance since it is hard to imagine *practical* examples where (FE3) is satisfied while (FE2MB) is not. After carrying out either the fixed effects or the random effects transform (FE2MB) implies that $\mathbf{X}_{iN} \perp\!\!\!\perp \varepsilon_{iN}$, $\varepsilon_{i1}, ..., \varepsilon_{iT_N}$ is i.i.d. gaussian for all $1 \leq i \leq N$, $E\left(\varepsilon_{it}\right) = 0$, and $E\left(\varepsilon_{it}^2\right) = \sigma^2$. The gaussianity assumption on the error terms is a price we must pay for working with L^q-norms instead of the exponential Orlicz-norms in Huang et al. (2008) in the proofs. Working with exponential Orlicz-norms did not turn out to be fruitful

[5]It is sufficient to assume $\tilde{\mathbf{X}}_{iN} \perp\!\!\!\perp \tilde{\varepsilon}_{iN}$ for all $1 \leq i \leq N$ but for comparison with (FE3) we refrain from this (see also the comment after (FE3)). Furthermore, as was the case in the $p_N < NT_N$ setting in Sections 3.2 and 3.3 $(\tilde{\varepsilon}_{i1}, ..., \tilde{\varepsilon}_{iT_N})$ may have any covariance structure as long as it is known so that it can be handled by an appropriate transformation. The most common choice is a diagonal matrix.

due to the presence of random covariates which rendered some otherwise useful maximal inequalities inapplicable.

Let $K_N = (1, ..., k_N)$ denote the active set, i.e. the set of indices of the relevant variables, and $J_N = (k_N + 1, ..., p_N)$ the inactive set, i.e. the set of indices of the irrelevant variables. Standardize the covariates such that $\frac{1}{NT_N}\sum_{j=1}^{NT_N} x_{jk}^2 = 1$ for all $k = 1, ..., p_N$. This implies that the covariates have moments of any order since $|x_{jk}| \leq \sqrt{\sum_{j=1}^{NT_N} x_{jk}^2} = (NT_T)^{1/2}$. Since the gaussian error terms also have moments of any order, this is the real reason enabling us to refrain from any moment assumptions like (FE2). Finally, define $\xi_{Nk} = \frac{1}{NT_N}\sum_{j=1}^{NT_N} \mathbf{w}_j' \beta_{10} x_{jk}$. Assume

(B1) There exists a constant $\xi_0 > 0$ such that $\min_{k \in K_N} |\xi_{Nk}| > \xi_0$ with probability approaching 1.

(B2) $\lambda_N / (NT_N) \to 0$.

(B3) $\frac{k_N}{\left(\lambda_N (NT_N)^{-\gamma/2}\right)^{1/(2-\gamma)}} \to 0$.

(B4) $\frac{m_N}{\left(\lambda_N (NT_N)^{-\gamma/2}\right)^{q/(2-\gamma)}} \to 0$ for some $q \geq 1$.

(B5) For all $\delta > 0$ there exists a $c_0 > 0$ and a $N_0 \in \mathbf{N}$ such that
$P\left(\frac{\sum_{j=1}^{NT_N} x_{jk} x_{jl}}{(NT_N)^{1/2}} \leq c_0,\ k \in K_N,\ l \in J_N\right) \geq 1 - \delta$ for $N \geq N_0$.

(B6) There exists a constant $0 < b_1 < \infty$ such that $\max_{k \in K_N} |\beta_{10k}| \leq b_1$.

Assumption (B1) is a technical assumption needed to prove that no variables from the active set will be discarded by the Marginal Bridge. In a fixed regressor setting it is similar to assuming that the covariance between the left hand side variable and the relevant covariates is bounded away from 0.

Assumption (B2) requires that $\lambda_N \in o(NT_N)$.

Assumption (B3) combined with assumption (B2) implies that $k_N \in o((NT_N)^{1/2})$. In the classical case of fixed T_N this amounts to $k_N \in o(N^{1/2})$. This is in line with the results of Huang et al. (2008).

For $0 < \gamma < 2$, (B3) also implies that $\lambda_N (NT_N)^{-\gamma/2} \to \infty$. Together with (B2) this yields that $\lambda_N \in o(NT_N) \cap \omega\left((NT_N)^{\gamma/2}\right)$.

Using (B2) in (B4) implies $m_N \in o\left((NT_N)^{q/2}\right)$. Since q is arbitrary this says that the number of irrelevant variables must be asymptotically dominated

by some polynomial in in the sample size. Notice that the number of irrelevant variables can not tend to infinity as fast as in Huang et al. (2008) where $m_N \in \exp(o(N))$. As indicated above the reason is that the exponential Orlicz-norms did not carry over straightforwardly to the random covariate setting and so we had to settle with maximal inequalities based on L^q norms which don't give us as sharp results. However, $m_N \in o(N^{q/2})$ (T_N fixed) is not very restrictive in practice since it still allows the number of irrelevant variables to increase at a much higher rate than the sample size as long as this rate is polynomial.

Assumption (B5) is a partial orthogonality assumption limiting the dependence between the variables in the active and the inactive set. It rules out correlations of -1 or 1[6]. However, it is not too restrictive and as will be seen from the Monte Carlo simulations in Section 3.5 the Marginal Bridge also works quite well even when the covariates in the active and inactive set are highly correlated.

Assumption (B6) is a uniform bound on the size of the coefficients belonging to the relevant variables. This assumption may be relaxed in the same way as assumption (A5) for the Bridge estimator at the price of a lower growth rate of the number of relevant variables.

Assumption (B1)-(B6) are similar to the assumptions made in Huang et al. (2008). However, we must assume gaussianity of the error terms instead of sub-Gaussianity in Huang et al. (2008)[7]. As indicated above, this is a price we pay for letting the covariates be random. The properties of the Marginal Bridge are given in the following Theorem.

Theorem 4. *Under assumption (B1)-(B6) and if* $0 < \gamma < 1$,

$$P\left(\tilde{\beta}_{2N} = 0\right) \to 1 \ \textit{and} \ P\left(\tilde{\beta}_{1Nk} = 0,\ k \in K_N\right) \to 0 \tag{3.11}$$

Hence, the Marginal Bridge estimator is able to screen out the irrelevant variables while retaining the relevant ones.

The nonzero coefficients are not estimated consistently. In order to obtain consistent estimates the same two step procedure as in Huang et al. (2008) can be applied. In the first step the Marginal Bridge estimator is applied to distinguish between the relevant and irrelevant variables. In the second step, where only the

[6] If x_{j1} and x_{j2} are perfectly correlated and (assume for simplicity) have an empirical mean of zero $x_{j2} = bx_{j1}$ *a.s.* for some constant b. Then $\frac{\sum_{j=1}^{NT_N} x_{j1}x_{j2}}{(NT_N)^{1/2}} = b(NT_N)^{1/2}$ which violates (B5).

[7] A random variable X is said to be sub-Gaussian if there exist positive constants C and K such that $P(|X| \geq x) \leq C\exp(-Kx^2)$

relevant variables are left, these may be estimated by any consistent estimator (e.g. least squares or the Bridge estimator).

3.5 Simulations

In this section the finite sample properties of the proposed estimators will be investigated. The Bridge estimator will be implemented by means of the MM-algorithm of Hunter and Li (2005) which in the present case reduces to a series of ridge regressions.

Deciding whether a variable is to be included or excluded using the Marginal Bridge is very fast. Since $\sum_{j=1}^{NT_N} x_{jk}^2 = NT_N$ for $k = 1, ..., p_N$ it follows from Lemma A in Knight and Fu (2000) that $\beta_k = 0$ iff

$$\frac{\lambda_N}{NT_N} > c_\gamma \left| \frac{\sum_{j=1}^{NT_N} y_j x_{jk}}{NT_N} \right|^{2-\gamma} \tag{3.12}$$

where $c_\gamma = \left(\frac{2}{2-\gamma}\right)\left(\frac{2(1-\gamma)}{2-\gamma}\right)^{1-\gamma}$. Hence, variable selection is extremely fast[8] even in vast dimensional models, since the inclusion of a variable is solely based on the criterion (3.12) which roughly amounts to checking whether the correlation between the left hand side variable and the covariate is sufficiently high to deem the latter relevant. Notice how only marginal information is used to decide whether a variable is to be included or not. Having decided on the sparsity pattern the second step estimates of β_{10} are found by means of least squares[9].

The following issues will be investigated

1. How often do the Bridge and the Marginal Bridge estimator select the correct sparsity pattern, i.e. how good are they at distinguishing the active from the inactive set? This is highly relevant in applied work investigating which variables help explaining the left hand side variable.

2. The median number of variables included, i.e how well do the Bridge and the Marginal Bridge reduce the dimension of the problem? This median is ideally equal to the cardinality of the active set.

[8] A model with 100 observations and 2500 potential explanatory variables takes between 0.2 and 0.3 seconds to estimate on a 2.66 GHz i7 processor.

[9] The Bridge estimator was also tried in the second step but did not outperform least squares while being considerably slower.

3. The explanatory power of the Bridge and the Marginal Bridge. To investigate this, the estimated parameters are used to fit values on a validation data set drawn from the same distribution as the training set.

4. In connection to the explanatory power it is investigated how often the procedures retain all relevant explanatory variables. As can be expected, retention of all relevant explanatory variables is important for achieving a good fit. It is also highly desirable if the procedures are to be used as initial screening devices in vast dimensional data sets.

5. The precision of the parameter estimates using the mean square error of $\hat{\beta}$.

6. The asymptotic distribution of the estimator of the non-zero β_0's. This is done by comparing the standard deviation of $\hat{\beta}_1$ to the corresponding quantities for the least squares estimator with only the active set included. The latter (in practice infeasible) estimator will be called the OLS Oracle henceforth.

7. In the $p_N < NT_N$ setting the coverage probabilities of 95% confidence intervals are reported for the Bridge estimator and the OLS Oracle to assess theorem 2 [10].

The Bridge and the Marginal Bridge estimators will be compared to the LASSO estimated by pathwise coordinate descent (see Friedman, Hastie, Höfling, and Tibshirani (2007)), the Schwarz information criterion (BIC), the OLS Oracle, and OLS on the system including all covariates. Only the Marginal Bridge, the LASSO and the OLS Oracle are applied when $p_N > NT_N$. To limit the computational burden, BIC is only applied for the designs with 15 or fewer covariates which implies a maximum of $2^{15} - 1 = 32,767$ regressions per Monte Carlo replication. All experiments are carried out with $1,000$ replications.

The data is generated from equation (3.2). In all experiments $T_N = 10$. Initial experiments indicated that $\gamma = 0.5$ works quite well for the Bridge as well as the Marginal Bridge estimator and this value will be used throughout. Larger values of γ resulted in larger models. In light of the fact that the LASSO turns out to select larger models than the Bridge Estimator, it is no surprise that as γ approaches 1 the median number of variables included by the Bridge Estimator increases too. $\tilde{\varepsilon}_{it}$ and c_i are $N(0,1)$ in all experiments.

[10] The author wishes to thank an anonymous referee for suggesting this.

The regularization parameter λ_N is usually chosen by 10-fold cross validation. Here we try this as well as the significantly faster BIC to determine λ_N for the Bridge, the Marginal Bridge and the LASSO.

The experiments

(A) N=10, $\beta_0 = (1,1,1,1,1,0,0,0,0,0,0,0,0,0,0)'$ and the covariates are independent N(0,1) variables.

(B) β_0 is as in (A). The correlation between the k'th and l'th covariate is $\rho^{|k-l|}$ with $\rho = 0.50$.

(C) As (B) but with $\rho = 0.95$.

(D) As (A) but with N=100.

(E) As (B) but with N=100.

(F) As (C) but with N=100.

(G) N=10 and 5 relevant explanatory variables with a coefficient of 1. 245 irrelevant variables. All covariates are independent.

(H) N=10 and 5 relevant explanatory variables with a coefficient of 1. 495 irrelevant variables. All covariates are independent.

(I) N=10 and 5 relevant explanatory variables with a coefficient of 1. 2495 irrelevant variables. All covariates are independent.

Note that even though the covariates are independent in Experiments G-I the maximum spurious sample correlation, i.e. the maximum observed sample correlation between covariates, may still be very high (see Fan and Lv (2008) for examples). In particular, if a relevant and irrelevant covariate are highly correlated it will be difficult to distinguish between these.

Results

Table 3.1 holds the results for experiments A-F, where $p_N < NT_N$.

Experiment A reveals that the Bridge, Marginal Bridge and Schwarz information criterion all perform quite well in the independent covariates setting.

		Cross Validation			BIC					
		Bridge	Marg Bridge	LASSO	Bridge	Marg Bridge	LASSO	BIC	OLS Oracle	OLS All
Experiment A	Sparsity pattern	0.6040	0.5190	0.0130	0.6700	0.7830	0.0990	0.6590	1.0000	0.0000
	Median #Var	5.0000	5.0000	10.0000	5.0000	5.0000	8.0000	5.0000	5.0000	15.0000
	Loss	2.1128	2.1062	2.1544	2.0947	2.0794	2.1643	2.1007	2.0667	2.2125
	Relevant retained	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	Median Beta MSE	0.0764	0.0725	0.0950	0.0706	0.0647	0.0971	0.0712	0.0584	0.1131
	Stdv	0.1188	0.1145	0.1206	0.1172	0.1145	0.1236	0.1148	0.1135	0.1198
	Cov. Prob	0.9140			0.9180				0.9390	
Experiment B	Sparsity pattern	0.6560	0.6520	0.0540	0.7170	0.9220	0.2910	0.6870	1.0000	0.0000
	Median #Var	5.0000	5.0000	8.0000	5.0000	5.0000	6.0000	5.0000	5.0000	15.0000
	Loss	2.1048	2.1009	2.1288	2.0898	2.0740	2.1374	2.0959	2.0667	2.2125
	Relevant retained	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	Median Beta MSE	0.0863	0.0843	0.0952	0.0827	0.0740	0.0927	0.0841	0.0710	0.1441
	Stdv	0.1387	0.1327	0.1362	0.1351	0.1306	0.1393	0.1314	0.1301	0.1365
	Cov. Prob	0.9300			0.9420				0.9520	
Experiment C	Sparsity pattern	0.0760	0.5600	0.1370	0.0180	0.6400	0.2750	0.0070	1.0000	0.0000
	Median #Var	5.0000	5.0000	7.0000	4.0000	5.0000	6.0000	4.0000	5.0000	15.0000
	Loss	2.1566	2.1106	2.1048	2.1664	2.1011	2.0984	2.1828	2.0667	2.2125
	Relevant retained	0.3230	0.8920	0.8960	0.0290	0.6870	0.8840	0.0110	1.0000	1.0000
	Median Beta MSE	0.4038	0.2854	0.2625	0.4809	0.2840	0.2496	0.4862	0.2189	0.4799
	Stdv	0.5042	0.4080	0.3548	0.6046	0.5138	0.3551	0.6123	0.3497	0.3680
	Cov. Prob	0.7350			0.5710				0.9540	

		Cross Validation			BIC					
		Bridge	Marg Bridge	LASSO	Bridge	Marg Bridge	LASSO	BIC	OLS Oracle	OLS All
Experiment D	Sparsity pattern	0.7570	0.6590	0.0190	0.9350	0.9860	0.4570	0.9110	1.0000	0.0000
	Median #Var	5.0000	5.0000	10.0000	5.0000	5.0000	6.0000	5.0000	5.0000	15.0000
	Loss	2.0089	2.0092	2.0135	2.0073	2.0066	2.0195	2.0074	2.0065	2.0176
	Relevant retained	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	Median Beta MSE	0.0210	0.0209	0.0279	0.0191	0.0179	0.0335	0.0186	0.0178	0.0326
	Stdv	0.0334	0.0331	0.0340	0.0333	0.0330	0.0354	0.0330	0.0330	0.0331
	Cov. Prob	0.9470			0.9470				0.9480	
Experiment E	Sparsity pattern	0.7660	0.7100	0.1180	0.9340	0.9930	0.6230	0.9110	1.0000	0.0000
	Median #Var	5.0000	5.0000	8.0000	5.0000	5.0000	5.0000	5.0000	5.0000	15.0000
	Loss	2.0085	2.0084	2.0114	2.0070	2.0066	2.0136	2.0073	2.0065	2.0176
	Relevant retained	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	Median Beta MSE	0.0239	0.0247	0.0278	0.0222	0.0215	0.0276	0.0225	0.0214	0.0410
	Stdv	0.0391	0.0387	0.0391	0.0390	0.0386	0.0395	0.0387	0.0386	0.0389
	Cov. Prob	0.9440			0.9420				0.9510	
Experiment F	Sparsity pattern	0.6960	0.7020	0.1570	0.9310	0.9930	0.4690	0.9240	1.0000	0.0000
	Median #Var	5.0000	5.0000	7.0000	5.0000	5.0000	6.0000	5.0000	5.0000	15.0000
	Loss	2.0092	2.0088	2.0094	2.0075	2.0066	2.0096	2.0072	2.0065	2.0176
	Relevant retained	1.0000	1.0000	1.0000	0.9990	1.0000	1.0000	0.9990	1.0000	1.0000
	Median Beta MSE	0.0802	0.0795	0.0775	0.0721	0.0662	0.0725	0.0684	0.0659	0.1362
	Stdv	0.1134	0.1067	0.1069	0.1120	0.1066	0.1079	0.1066	0.1066	0.1078
	Cov. Prob	0.9340			0.9380				0.9510	

Table 3.1: Top panel: Experiments A-C (N=10). Bottom panel: Experiments D-F (N=100). Cross Validation and BIC indicate which procedure was used to determine λ_N. Sparsity pattern: The fraction of times the correct sparsity pattern is detected. Median #Var: The median number of variables included. Loss: The MSE when using the estimated parameters on a validation data set drawn from the same distribution as the training set. Relevant retained: The fraction of relevant variables retained in the model. Median Beta MSE: Calculated as explained in the main text. Stdv: Standard deviation of the estimated coefficient of the first variable (which is always in the active set). Cov. Prob: Coverage probability of 95% confidence interval of the estimated coefficient of the first variable.

They all detect the correct sparsity pattern in more than half of the cases irrespective of whether cross validation or BIC is used to determine λ_N. In all respects their performance is comparable to the OLS Oracle.

As seen from Experiment B making the covariates moderately correlated does not deteriorate the performance of the procedures with respect to the fraction of times the right sparsity pattern is chosen or the fraction of times all relevant covariates are retained. However, all procedures get more imprecise. Since this is also the case for the OLS Oracle this is not a particular artefact of the Bridge class of estimators.

Experiment C reveals that as the correlation gets very high the performance of the Bridge and BIC deteriorate. On the other hand the Marginal Bridge continues to detect the right sparsity pattern in more than half of the cases. However, even the latter fails to retain all relevant variables in all cases. The coverage probabilities of the Bridge confidence intervals are significantly below 95% – this is the case in particular when BIC is used to select λ_N. Taking into account that it only retains all relevant variables in 3% of the simulations, this result is not too surprising.

Experiments D-F illuminate the asymptotic properties of the Bridge and the Marginal Bridge. In particular the Marginal Bridge with BIC used to determine λ_N detects the correct sparsity pattern in almost all cases irrespective of the correlation structure imposed on the covariates. The performance of the Bridge also gets significantly better as the sample size is increased while the LASSO only improves moderately. The Loss of all procedures is reduced and the parameters are estimated more precisely. Finally, the coverage probabilities of the Bridge are now close to the nominal rate of 95%.

Notice that the Marginal Bridge performs quite well even in the high correlation experiments C and F indicating that the partial orthogonality assumption (B5) is not overly restrictive.

It is seen that in general the BIC is a better way of determining λ_N than cross validation. BIC detects the correct sparsity pattern more often and only in Experiment C one finds that cross validation is superior with respect to the number of relevant variables retained as well as coverage probabilities.

Table 3.2 contains the results for the Experiments G-I which investigate the performance of the Marginal Bridge in the $p_N > NT_N$ case. As can be expected the correct sparsity pattern is detected less frequently. However, all relevant variables are retained very often while only few irrelevant variables are kept in the model. Hence, the Marginal Bridge is still a very effective tool for dimension

		Cross Validation			BIC					
		Bridge	Marg Bridge	LASSO	Bridge	Marg Bridge	LASSO	BIC	OLS Oracle	OLS All
Experiment G	Sparsity pattern		0.1380	0.0020		0.3380	0.0420		1.0000	
	Median #Var		9.0000	20.0000		6.0000	9.0000		5.0000	
	Loss		2.3302	2.4744		2.2263	2.6162		2.0784	
	Relevant retained		0.9820	1.0000		0.9390	0.9990		1.0000	
	Median Beta MSE		0.0299	0.0413		0.0204	0.0470		0.0141	
	Stdv		0.1465	0.1358		0.1612	0.1448		0.1081	
Experiment H	Sparsity pattern		0.0710	0.0010		0.2300	0.0290		1.0000	
	Median #Var		12.0000	23.5000		7.0000	10.0000		5.0000	
	Loss		2.5162	2.5763		2.3593	2.7642		2.0891	
	Relevant retained		0.9350	1.0000		0.8830	0.9980		1.0000	
	Median Beta MSE		0.0257	0.0320		0.0177	0.0368		0.0101	
	Stdv		0.1714	0.1414		0.2045	0.1553		0.1100	
Experiment I	Sparsity pattern		0.0100	0.0000		0.0540	0.0130		1.0000	
	Median #Var		19.0000	35.0000		9.0000	11.0000		5.0000	
	Loss		3.1359	2.8733		2.8479	3.2288		2.0331	
	Relevant retained		0.7450	0.9950		0.6730	0.9540		1.0000	
	Median Beta MSE		0.0178	0.0183		0.0147	0.0213		0.0045	
	Stdv		0.2428	0.1470		0.2657	0.1690		0.1056	

Table 3.2: Cross Validation and BIC indicate which procedure was used to determine λ_N. Sparsity pattern: The fraction of times the correct sparsity pattern is detected. Median #Var: The median number of variables included. Loss: The MSE when using the estimated parameters on a validation data set drawn from the same distribution as the training set. Relevant retained: The fraction of relevant variables retained in the model. Median Beta MSE: Calculated as explained in the main text. Stdv: Standard deviation of the estimated coefficient of the first variable (which is always in the active set).

reduction.

The LASSO and the Marginal Bridge perform equally well in Experiments G and H (slight advantage for the LASSO) while the LASSO is superior in Experiment I. However, the LASSO also takes a lot longer to compute and the models it chooses are bigger. The following idea which builds on the thoughts of Fan and Lv (2008) could potentially improve the performance of the Marginal Bridge: estimate the Marginal Bridge one or several times more using the residuals from the first (previous) step as left hand side variables. This will lower the priority of those irrelevant variables which seemed relevant only through their high correlation with some of the relevant variables already included.

3.6 Conclusions

This paper introduces the Bridge and Marginal Bridge estimator in a linear panel data setting allowing for random as well as fixed effects. When $p < NT_N$ it is

shown that the Bridge estimator (and Marginal Bridge) has the oracle property. It sets all coefficients that are truly zero to zero and the asymptotic distribution of the estimator of the non zero coefficients is the same as if the sparsity pattern had been known. Monte Carlo experiments underscore this conclusion and are used to investigate the finite sample properties of the procedures. They also reveal that the Schwarz information criterion is more useful than 10-fold cross validation for selecting λ_N. This is encouraging since BIC is also faster than cross validation.

When $p > NT_N$ it is shown that the Marginal Bridge estimator still detects the correct sparsity pattern with probability converging to one. This is true in the random as well as the fixed effects setting under a partial orthogonality assumption on the covariates. However, the Marginal Bridge works well even when the relevant and irrelevant covariates are highly correlated. Furthermore, the Marginal Bridge estimates are extremely fast to compute since only marginal information is used to decide whether a variable is relevant or not. The Marginal Bridge is also shown to perform well in the $p_N < NT_N$ setting. In the $p_N > NT_N$ setting the Marginal Bridge does not always retain all relevant variables. An iterative procedure was proposed to solve this problem. Working out the properties of this procedure is left for future research.

3.7 Appendix

Lemma 1. *Let* $\mathbf{u}$ *be a* $p_N \times 1$ *vector. Then*

$$E\left(\sup_{\|\mathbf{u}\|\leq\delta}\left|\sum_{j=1}^{NT_N}\varepsilon_j\mathbf{x}_j'\mathbf{u}\right|\middle|\mathbf{X}_N\right)\leq\delta\sigma\left(NT_Np_N\right)^{\frac{1}{2}}\left(\frac{1}{NT_Np_N}\sum_{i=1}^{N}\sum_{t=1}^{T_N}\sum_{k=1}^{p_N}x_{itk}^2\right)^{\frac{1}{2}}$$

Proof.

$$\begin{aligned}E\left(\sup_{\|\mathbf{u}\|\leq\delta}\left|\sum_{j=1}^{NT_N}\varepsilon_j\mathbf{x}'{}_j\mathbf{u}\right|^2\middle|\mathbf{X}_N\right)&\leq E\left[\sup_{\|\mathbf{u}\|\leq\delta}\|\mathbf{u}\|^2\left\|\left(\sum_{j=1}^{NT_N}\varepsilon_j\mathbf{x}_j\right)\right\|^2\middle|\mathbf{X}_N\right]\\&\leq\delta^2E\left[\left(\sum_{j=1}^{NT_N}\varepsilon_j\mathbf{x}_j'\right)\left(\sum_{j=1}^{NT_N}\varepsilon_j\mathbf{x}_j\right)\middle|\mathbf{X}_N\right]\end{aligned}$$

where the first inequality follows from the Cauchy-Schwarz inequality. Since

$$\begin{aligned}&E\left(\varepsilon_{it}\mathbf{x}_{it}'\varepsilon_{is}\mathbf{x}_{is}|\mathbf{X}_N\right)=E\left(\mathbf{x}_{it}'\mathbf{x}_{is}\varepsilon_{it}\varepsilon_{is}|\mathbf{X}_{iN}\right)=\mathbf{x}_{it}'\mathbf{x}_{is}E\left(\varepsilon_{it}\varepsilon_{is}|\mathbf{X}_{iN}\right)=0,\ \ t\neq s\\&E\left(\varepsilon_{it}\mathbf{x}_{it}'\varepsilon_{js}\mathbf{x}_{js}|\mathbf{X}_N\right)=E\left(E\left[\varepsilon_{it}\mathbf{x}_{it}'\varepsilon_{js}\mathbf{x}_{js}|\mathbf{X}_N,\varepsilon_{it}\right]|\mathbf{X}_N\right)\\&=E\left(\varepsilon_{it}\mathbf{x}_{it}'\mathbf{x}_{js}E\left[\varepsilon_{js}|\mathbf{X}_N,\varepsilon_{it}\right]|\mathbf{X}_N\right)=E\left(\varepsilon_{it}\mathbf{x}_{it}'\mathbf{x}_{js}E\left[\varepsilon_{js}|\mathbf{X}_{jN}\right]|\mathbf{X}_N\right)=0,\ \ i\neq j\\&E\left(\varepsilon_{it}\mathbf{x}_{it}'\varepsilon_{it}\mathbf{x}_{it}|\mathbf{X}_N\right)=E\left[\mathbf{x}_{it}'\mathbf{x}_{it}\varepsilon_{it}\varepsilon_{it}|\mathbf{X}_{iN}\right]=\mathbf{x}_{it}'\mathbf{x}_{it}E\left[\varepsilon_{it}\varepsilon_{it}|\mathbf{X}_{iN}\right]=\sigma^2\mathbf{x}_{it}'\mathbf{x}_{it}\end{aligned}$$

Hence,

$$\begin{aligned}&E\left(\sup_{\|\mathbf{u}\|\leq\delta}\left|\sum_{j=1}^{NT_N}\varepsilon_j\mathbf{x}_j\mathbf{u}\right|^2\middle|\mathbf{X}_N\right)\leq\delta^2\sigma^2\sum_{j=1}^{NT_N}\mathbf{x}_j'\mathbf{x}_j=\delta^2\sigma^2\sum_{i=1}^{N}\sum_{t=1}^{T_N}\mathbf{x}_{it}'\mathbf{x}_{it}\\&=\delta^2\sigma^2NT_Np_N\frac{1}{NT_Np_N}\sum_{i=1}^{N}\sum_{t=1}^{T_N}\sum_{k=1}^{p_N}x_{itk}^2\end{aligned}$$

and the result follows from the conditional Jensen inequality. □

Lemma 2. *Let* $\{X_n\}_{n\in\mathbb{N}}$ *and* $\{Y_n\}_{n\in\mathbb{N}}$ *be sequences of nonnegative random variables. If there exists an integer* N_0 *and a constant* C *such that for* $n\geq N_0$

$$E\left(\frac{X_n}{Y_n}\right)\leq C$$

then

$$X_n \in O_p(Y_n)$$

Proof. It suffices to show that for any $\varepsilon > 0$ $P\left(\left\{\frac{X_n}{Y_n} > \frac{C}{\varepsilon}\right\}\right) \leq \varepsilon$ for $n \geq N_0$. Assume the opposite is true for some $\varepsilon > 0$ to reach a contradiction. Then,

$$E\left(\frac{X_n}{Y_n}\right) = \int \frac{X_n}{Y_n} dP \geq \int_{\{\frac{X_n}{Y_n} > \frac{C}{\varepsilon}\}} \frac{X_n}{Y_n} dP \geq \int_{\{\frac{X_n}{Y_n} > \frac{C}{\varepsilon}\}} \frac{C}{\varepsilon} dP \geq \frac{C}{\varepsilon} P\left(\left\{\frac{X_n}{Y_n} > \frac{C}{\varepsilon}\right\}\right) > C$$

which is the desired contradiction. □

Proof of Theorem 1. We first show that $\left\|\hat{\beta}_N - \beta_0\right\| \in O_p\left(\left[\frac{p_N + \lambda_N k_N}{NT_N \rho_{1N}}\right]^{\frac{1}{2}}\right)$. Since $\hat{\beta}_N$ minimizes (3.6)

$$\sum_{j=1}^{NT_N} \left(y_j - \mathbf{x}_j' \hat{\beta}_N\right)^2 + \lambda_N \sum_{k=1}^{p_N} |\hat{\beta}_{Nk}|^\gamma \leq \sum_{j=1}^{NT_N} \left(y_j - \mathbf{x}_j' \beta_0\right)^2 + \lambda_N \sum_{k=1}^{p_N} |\beta_{0k}|^\gamma$$

Defining $\eta_N = \lambda_N \sum_{k=1}^{p_N} |\beta_{0k}|^\gamma$ this implies:

$$\begin{aligned}
\eta_N &\geq \sum_{j=1}^{NT_N} \left(y_j - \mathbf{x}_j' \hat{\beta}_N\right)^2 - \sum_{j=1}^{NT_N} \left(y_j - \mathbf{x}_j' \beta_0\right)^2 \\
&= \sum_{j=1}^{NT_N} \left(\left[y_j - \mathbf{x}_j' \hat{\beta}_N\right] - \left[y_j - \mathbf{x}_j' \beta_0\right]\right)\left(\left[y_j - \mathbf{x}_j' \hat{\beta}_N\right] + \left[y_j - \mathbf{x}_j' \beta_0\right]\right) \\
&= \sum_{j=1}^{NT_N} \left[\mathbf{x}_j'(\beta_0 - \hat{\beta}_N)\right]^2 + 2 \sum_{j=1}^{NT_N} \varepsilon_j \mathbf{x}_j' \left(\beta_0 - \hat{\beta}_N\right)
\end{aligned}$$

Now define $\delta_N = (NT_N)^{1/2} \Sigma_N^{1/2} (\hat{\beta}_N - \beta_0)$, $\mathbf{D}_N = (NT_N)^{-1/2} \Sigma_N^{-1/2} \mathbf{X}_N'$ and $\varepsilon_N = (\varepsilon_1, ..., \varepsilon_{NT_N})'$. With these definitions

$$\begin{aligned}
\sum_{j=1}^{NT_N} \left[\mathbf{x}_j'(\beta_0 - \hat{\beta}_N)\right]^2 + 2 \sum_{j=1}^{NT_N} \varepsilon_j \mathbf{x}_j' \left(\beta_0 - \hat{\beta}_N\right) &= \delta_N' \delta_N - 2(\mathbf{D}_N \varepsilon_N)' \delta_N \\
&= ||\delta_N - \mathbf{D}_N \varepsilon_N||^2 - ||\mathbf{D}_N \varepsilon_N||^2
\end{aligned}$$

which implies

$$||\delta_N - \mathbf{D}_N \varepsilon_N||^2 - ||\mathbf{D}_N \varepsilon_N||^2 - \eta_N \leq 0$$

and by the sub additivity of $x \mapsto x^{1/2}$ yields $||\delta_N - \mathbf{D}_N\varepsilon_N|| \leq ||\mathbf{D}_N\varepsilon_N|| + \eta_N^{1/2}$. By sub additivity of the norm $\|\cdot\|$ this implies $||\delta_N|| \leq ||\delta_N - \mathbf{D}_N\varepsilon_N|| + ||\mathbf{D}_N\varepsilon_N|| \leq 2||\mathbf{D}_N\varepsilon_N|| + \eta_N^{1/2}$. Since $(x+y)^2 \leq 2x^2 + 2y^2$ for $x, y \in \mathbf{R}$ by the convexity of $x \mapsto x^2$ one has

$$||\delta_N||^2 \leq 4||\mathbf{D}_N\varepsilon_N||^2 + 2\eta_N$$

Letting $\mathbf{d}_j$ denote the j'th column of $\mathbf{D}_N$ we may write $\mathbf{D}_N\varepsilon_N = \sum_{j=1}^{NT_N} \mathbf{d}_j\varepsilon_j$. Using that $\mathbf{D}_N$ is measurable with respect to $\mathbf{X}_N$, conclude

$$E\left(\mathbf{d}_{it}'\varepsilon_{it}\mathbf{d}_{is}\varepsilon_{is}\right) = E\left(\mathbf{d}_{it}'\mathbf{d}_{is}E\left[\varepsilon_{it}\varepsilon_{is}|\mathbf{X}_N\right]\right) = E\left(\mathbf{d}_{it}'\mathbf{d}_{is}E\left[\varepsilon_{it}\varepsilon_{is}|\mathbf{X}_{iN}\right]\right) = 0,\ s \neq t$$
$$E\left(\mathbf{d}_{it}'\varepsilon_{it}\mathbf{d}_{js}\varepsilon_{js}\right) = E\left(\mathbf{d}_{it}'\varepsilon_{it}\mathbf{d}_{js}E\left[\varepsilon_{js}|\mathbf{X}_N, \varepsilon_{it}\right]\right) = E\left(\mathbf{d}_{it}'\varepsilon_{it}\mathbf{d}_{js}\left[\varepsilon_{js}|\mathbf{X}_{jN}\right]\right) = 0,\ i \neq j$$
$$E\left(\mathbf{d}_{it}'\varepsilon_{it}\mathbf{d}_{it}\varepsilon_{it}\right) = E\left(\mathbf{d}_{it}'\mathbf{d}_{it}E\left[\varepsilon_{it}\varepsilon_{it}|\mathbf{X}_{iN}\right]\right) = E\left(\mathbf{d}_{it}'\mathbf{d}_{it}\right) = \sigma^2 E\left(||\mathbf{d}_{it}||^2\right)$$

Hence,

$$\begin{aligned} E\left(\|\mathbf{D}_N\varepsilon_N\|^2\right) &= E\left(\left\|\sum_{j=1}^{NT_N}\mathbf{d}_j\varepsilon_j\right\|^2\right) = E\left[\left(\sum_{j=1}^{NT_N}\mathbf{d}_j\varepsilon_j\right)'\left(\sum_{j=1}^{NT_N}\mathbf{d}_j\varepsilon_j\right)\right] \\ &= \sigma^2 E\left(\sum_{j=1}^{NT_N}\|\mathbf{d}_j\|^2\right) = \sigma^2 E\left(\mathrm{tr}\left(\mathbf{D}_N'\mathbf{D}_N\right)\right) = \sigma^2 E\left(\mathrm{tr}\left(\mathbf{D}_N\mathbf{D}_N'\right)\right) \\ &= \sigma^2 \mathrm{tr}\left(\mathbf{I}_{p_N}\right) = \sigma^2 p_N \end{aligned}$$

And so, $E\left(||\delta_N||^2\right) \leq 4\sigma^2 p_N + 2\eta_N$. Hence,

$$(NT_N)E\left(\left(\hat{\beta}_N - \beta_0\right)'\Sigma_N\left(\hat{\beta}_N - \beta_0\right)\right) = E(\delta_N'\delta_N) = E(||\delta_N||^2) \leq 4\sigma^2 p_N + 2\eta_N$$

Since the number of non zero coefficients is k_N

$$\eta_N = \lambda_N \sum_{k=1}^{p_N} |\beta_{0j}|^\gamma = \lambda_N \sum_{k=1}^{k_N} |\beta_{0j}|^\gamma \leq \lambda_N k_N b_1^\gamma$$

where the inequality is a consequence of assumption (A5). Since ρ_{1N} is the smallest eigenvalue of Σ_N

$$\rho_{1N}||\hat{\beta}_N-\beta_0||^2=\rho_{1N}\left(\hat{\beta}_N-\beta_0\right)'\left(\hat{\beta}_N-\beta_0\right)\leq\left(\hat{\beta}_N-\beta_0\right)'\Sigma_N\left(\hat{\beta}_N-\beta_0\right)$$

Hence,

$$\begin{aligned}E\left(\rho_{1N}||\hat{\beta}_N-\beta_0||^2\right)&\leq\frac{NT_N}{NT_N}E\left(\left(\hat{\beta}_N-\beta_0\right)'\Sigma_N\left(\hat{\beta}_N-\beta_0\right)\right)\\&\leq\frac{4\sigma^2p_N+2\eta_N}{NT_N}\leq\frac{4\sigma^2p_N+2\lambda_Nk_Nb_1^\gamma}{NT_N}\\&\leq C\frac{p_N+\lambda_Nk_Nb_1^\gamma}{NT_N}\end{aligned}$$

for $C=\max\left(4\sigma^2,2b_1^\gamma\right)$. This implies

$$E\left(\frac{||\hat{\beta}_N-\beta_0||^2}{\left(\frac{p_N+\lambda_Nk_N}{NT_N\rho_{1N}}\right)}\right)\leq C$$

By Lemma 2 this establishes $||\hat{\beta}_N-\beta_0||\in O_p\left(\left[\frac{p_N+\lambda_Nk_N}{NT_N\rho_{1N}}\right]^{\frac{1}{2}}\right)$. Next we show that $||\hat{\beta}_N-\beta_0||\in O_p\left(\rho_{1N}^{-1}\left(p_N/(NT_N)\right)^{1/2}\right)$. Like Huang et al. (2008) we use the idea from the proof of Theorem 3.2.5 in Van der Vaart and Wellner (1996). Let $r_N=\rho_{1N}^{-1}\left(p_N/(NT_N)\right)^{1/2}$. For every N partition the parameter space (excluding β_0) into the disjoint shells $S_{l,N}=\left\{\beta:2^{l-1}<\|\beta-\beta_0\|/r_N\leq 2^l\right\}$ where $l\in\mathbb{Z}$. If $2^M<\left\|\hat{\beta}_N-\beta_0\right\|/r_N$ for a given integer M then $\hat{\beta}_N\in\bigcup_{l>M}S_{l,N}$. For the shell which $\hat{\beta}_N$ belongs to, the infimum of the map $\beta\mapsto L_N(\beta)-L_N(\beta_0)$ is non positive. Hence, for any $\delta>0$ [11]

[11] Note that

$$\begin{aligned}&\left\{\left\|\hat{\beta}_N-\beta_0\right\|/r_N>2^M\right\}\subseteq\bigcup_{l>M}\left\{\inf_{\beta\in S_{l,N}}\left(L_N(\beta)-L_N(\beta_0)\right)\leq 0\right\}\\&\subseteq\left(\bigcup_{l>M}\left\{\inf_{\beta\in S_{l,N}}\left(L_N(\beta)-L_N(\beta_0)\right)\leq 0\right\}\cap\left\{\left\|\hat{\beta}_N-\beta_0\right\|\leq\delta\right\}\right)\cup\left\{\left\|\hat{\beta}_N-\beta_0\right\|>\delta\right\}\\&=\left(\bigcup_{l>M}\left\{\inf_{\beta\in S_{l,N}}\left(L_N(\beta)-L_N(\beta_0)\right)\leq 0\right\}\cap\left\{\left\|\hat{\beta}_N-\beta_0\right\|/r_N\leq\delta/r_N\right\}\right)\cup\left\{\left\|\hat{\beta}_N-\beta_0\right\|>\delta\right\}\\&\subseteq\bigcup_{\substack{l>M\\2^{l-1}<\delta/r_N}}\left\{\inf_{\beta\in S_{l,N}}\left(L_N(\beta)-L_N(\beta_0)\right)\leq 0\right\}\cup\left\{\left\|\hat{\beta}_N-\beta_0\right\|>\delta\right\}\end{aligned}$$

$$P\left(\left\|\hat{\beta}_N-\beta_0\right\|/r_N>2^M\right)$$
$$\leq \sum_{\substack{l>M\\ 2^{l-1}<\delta/r_N}} P\left(\inf_{\beta\in S_{l,N}}\left(L_N(\beta)-L_N(\beta_0)\right)\leq 0\right)+P\left(\left\|\hat{\beta}_N-\beta_0\right\|>\delta\right) \quad (3.13)$$

The last term in (3.13) converges to 0 by the consistency of $\hat{\beta}_N$ shown in the first part of the theorem. The theorem is established by proving that the first term on the right hand side can be made arbitrarily small by choosing M sufficiently large. To this is end let $\beta\in S_{l,N}$ for an arbitrary l summed over, and notice that

$$L_N(\beta)-L_N(\beta_0)=\sum_{j=1}^{NT_N}\left(y_j-\mathbf{x}_j'\beta\right)^2+\lambda_N\sum_{k=1}^{k_N}|\beta_{1k}|^\gamma+\lambda_N\sum_{k=1}^{m_N}|\beta_{2k}|^\gamma$$
$$-\sum_{j=1}^{NT_N}\left(y_j-\mathbf{x}_j'\beta_0\right)^2-\lambda_N\sum_{k=1}^{k_N}|\beta_{01k}|^\gamma$$
$$\geq\sum_{j=1}^{NT_N}\left(y_j-\mathbf{x}_j'\beta\right)^2+\lambda_N\sum_{k=1}^{k_N}|\beta_{1k}|^\gamma-\sum_{j=1}^{NT_N}\left(y_j-\mathbf{x}_j'\beta_0\right)^2-\lambda_N\sum_{k=1}^{k_N}|\beta_{01k}|^\gamma$$
$$=\sum_{j=1}^{NT_N}\left(\mathbf{x}_j'\left[\beta-\beta_0\right]\right)^2-2\sum_{j=1}^{NT_N}\varepsilon_j\mathbf{x}_j'(\beta-\beta_0)+\lambda_N\sum_{k=1}^{k_N}\left(|\beta_{1k}|^\gamma-|\beta_{01k}|^\gamma\right) \quad (3.14)$$

Regarding the first term in (3.14),

$$\sum_{j=1}^{NT_N}\left(\mathbf{x}_j'\left[\beta-\beta_0\right]\right)^2=\left[\beta-\beta_0\right]'\sum_{j=1}^{NT_N}\mathbf{x}_j\mathbf{x}_j'\left[\beta-\beta_0\right]=NT_N\left[\beta-\beta_0\right]'\Sigma_N\left[\beta-\beta_0\right]$$
$$\geq NT_N\|\beta-\beta_0\|^2\rho_{1N}>NT_N2^{2(l-1)}r_N^2\rho_{1N}$$

Regarding the third term in (3.14) we notice that $\beta\in S_{l,N}$ and $2^{l-1}<\delta/r_N$ implies that $\|\beta-\beta_0\|/r_N\leq 2^l<2\delta/r_N$. Hence, it suffices to consider βs satisfying $\|\beta-\beta_0\|<2\delta$. Since $\delta>0$ is arbitrary and the entries of β_{10} are bounded uniformly away from the 0 by b_0 the mean value theorem may be applied to conclude that for some ζ_k between β_{1k} and β_{10k}

and conclude using the subadditivity of P.

$$\left|\lambda_N \sum_{k=1}^{k_N} \left(|\beta_{1k}|^\gamma - |\beta_{01k}|^\gamma\right)\right| = \left|\lambda_N \gamma \sum_{k=1}^{k_N} |\zeta_k|^{\gamma-1} \text{sign}(\zeta_k)\,(\beta_{1k} - \beta_{01k})\right|$$
$$\leq c\lambda_N \sum_{k=1}^{k_N} |\beta_{1k} - \beta_{01k}| \leq c\lambda_N k_N^{1/2} \|\beta - \beta_0\| \leq c\lambda_N k_N^{1/2} 2^l r_N$$

where $c = \gamma(b_0 - 2\delta)^{\gamma-1}$ and the second to last estimate follows from Jensen's inequality. Hence, on $S_{l,N}$, $\lambda_N \sum_{k=1}^{k_N} \left(|\beta_{1k}|^\gamma - |\beta_{01k}|^\gamma\right) \geq -c\lambda_N k_N^{1/2} 2^l r_N$. Therefore, on $S_{l,N}$,

$$L_N(\beta) - L_N(\beta_0) \geq -\left|2\sum_{j=1}^{NT_N} \varepsilon_j \mathbf{x}_j'(\beta - \beta_0)\right| + \rho_{1N} NT_N 2^{2(l-1)} r_N^2 - c\lambda_N k_N^{1/2} 2^l r_N$$

Hence, by the conditional Markov inequality and Lemma 1

$$P\left(\inf_{\beta \in S_{l,N}} \left(L_N(\beta) - L_N(\beta_0)\right) \leq 0 \middle| \mathbf{X}_N\right)$$
$$\leq P\left(\sup_{\beta \in S_{l,N}} \left|2\sum_{j=1}^{NT_N} \varepsilon_j \mathbf{x}_j'(\beta - \beta_0)\right| \geq \rho_{1N} NT_N 2^{2(l-1)} r_N^2 - c\lambda_N k_N^{1/2} 2^l r_N \middle| \mathbf{X}_N\right)$$
$$\leq \frac{E\left(\sup_{\beta \in S_{l,N}} \left|2\sum_{j=1}^{NT_N} \varepsilon_j \mathbf{x}_j'(\beta - \beta_0)\right| \middle| \mathbf{X}_N\right)}{\rho_{1N} NT_N 2^{2(l-1)} r_N^2 - c\lambda_N k_N^{1/2} 2^l r_N}$$
$$\leq \frac{\sigma (NT_N p_N)^{1/2} 2^l r_N \left(\frac{1}{NT_N p_N} \sum_{i=1}^N \sum_{t=1}^{T_N} \sum_{k=1}^{p_N} x_{itk}^2\right)^{1/2}}{\rho_{1N} NT_N 2^{2(l-1)} r_N^2 - c\lambda_N k_N^{1/2} 2^l r_N}$$
$$= \frac{2\sigma \left(\frac{1}{NT_N p_N} \sum_{i=1}^N \sum_{t=1}^{T_N} \sum_{k=1}^{p_N} x_{itk}^2\right)^{1/2}}{2^{l-2} - c\lambda_N \left(k_N/(p_N NT_N)\right)^{1/2}}$$

By assumption (A3) $\lambda_N \left(k_N/(p_N NT_N)\right)^{1/2} \to 0$ and so $2^{l-2} - c\lambda_N \left(k_N/(NT_N)\right)^{1/2} \geq 2^{l-3}$ for N sufficiently large. Hence, by iterated expectations and assumption (A1)

$$P\left(\inf_{\beta\in S_{l,N}}\left(L_N(\beta)-L_N(\beta_0)\right)\leq 0\right)\leq\frac{\sigma\sqrt{K}}{2^{l-4}}$$

Finally, this implies that

$$\sum_{\substack{l>M\\ 2^{l-1}<\delta/r_N}} P\left(\inf_{\beta\in S_{l,N}}\left(L_N(\beta)-L_N(\beta_0)\right)\leq 0\right)\leq\sum_{l>M}\frac{\sigma\sqrt{K}}{2^{l-4}}$$

which is convergent and so the tail can be made arbitrarily small by choosing M sufficiently large. □

Lemma 3. *Suppose* $0<\gamma<1$. *Let* $\hat{\beta}_N=\left(\hat{\beta}_{1N},\,\hat{\beta}_{2N}\right)$. *Then* $\hat{\beta}_{2N}=0$ *with probability converging to* 1 *under assumptions (A1)-(A7).*

Proof. By Theorem 1 $||\hat{\beta}_N-\beta_0||\in O_p(h_N)$ with $h_N=\rho_{1N}^{-1}(p_N/(NT_N))^{1/2}$ so for all $\varepsilon>0$ there exists a constant C such that for N sufficiently large

$$P\left(||\hat{\beta}_N-\beta_0||/h_N>C\right)<\varepsilon\Leftrightarrow P\left(||\hat{\beta}_N-\beta_0||\leq Ch_N\right)\geq 1-\varepsilon$$

Put differently, $\hat{\beta}_N\in\left\{\beta:||\beta-\beta_0||\leq Ch_N\right\}$ with probability converging to 1. Let $\hat{\beta}_{1N}=\beta_{10}+h_N\mathbf{u}_1$ and $\hat{\beta}_{2N}=\beta_{20}+h_N\mathbf{u}_2=h_N\mathbf{u}_2$. Choosing $\hat{\beta}_N$ is then equivalent to choosing $\mathbf{u}_1$ and $\mathbf{u}_2$. For $\mathbf{u}=(\mathbf{u}_1',\mathbf{u}_2')'$ one has $||\mathbf{u}||=||\hat{\beta}_N-\beta_0||/h_N$ which is bounded by C with probability approaching 1. Hence, we may assume $||\mathbf{u}||^2=||\mathbf{u}_1||^2+||\mathbf{u}_2||^2\leq C^2$ and define

$$V_N(\mathbf{u}_1,\mathbf{u}_2)=L_N(\hat{\beta}_{1N},\hat{\beta}_{2N})=L_N(\beta_{10}+h_N\mathbf{u}_1,h_N\mathbf{u}_2)$$

To establish the lemma it now suffices to show that for any $\mathbf{u}$ with $\|\mathbf{u}\|\leq$

C, $V_N(\mathbf{u}_1,\mathbf{u}_2) - V_N(\mathbf{u}_1,\mathbf{0}) > 0$ with probability converging to 1 if $\mathbf{u}_2 \neq 0$. Now,

$$V_N(\mathbf{u}_1,\mathbf{u}_2) - V_N(\mathbf{u}_1,\mathbf{0}) = \sum_{j=1}^{NT_N} \left(y_j - \beta_{01}'\mathbf{w}_j - h_N\mathbf{u}_1'\mathbf{w}_j - h_N\mathbf{u}_2'\mathbf{z}_j\right)^2 + \lambda_N \sum_{k=1}^{k_N} |\beta_{01k} + h_N u_{1k}|^\gamma$$
$$+ \lambda_N \sum_{k=1}^{m_N} |h_N u_{2k}|^\gamma - \sum_{j=1}^{NT_N} \left(y_j - \beta_{01}'\mathbf{w}_j - h_N\mathbf{u}_1'\mathbf{w}_j\right)^2 - \lambda_N \sum_{k=1}^{k_N} |\beta_{01k} + h_N u_{1k}|^\gamma$$
$$= \sum_{j=1}^{NT_N} -h_N(\mathbf{u}_2'\mathbf{z}_j)\left[2(y_j - \beta_{01}'\mathbf{w}_j - h_N\mathbf{u}_1'\mathbf{w}_j) - h_N\mathbf{u}_2'\mathbf{z}_j\right] + \lambda_N \sum_{k=1}^{m_N} |h_N u_{2k}|^\gamma$$
$$= h_N^2 \sum_{j=1}^{NT_N} (\mathbf{z}_j'\mathbf{u}_2)^2 + 2h_N^2 \sum_{j=1}^{NT_N} (\mathbf{w}_j'\mathbf{u}_1)(\mathbf{z}_j'\mathbf{u}_2) - 2h_N \sum_{j=1}^{NT_N} (\mathbf{z}_j'\mathbf{u}_2)\varepsilon_j + \lambda_N h_N^\gamma \sum_{k=1}^{m_N} |u_{2j}|^\gamma$$

Regarding the sum of the first two terms since $2xy \geq -(x^2+y^2)$

$$h_N^2 \sum_{j=1}^{NT_N} (\mathbf{z}_j'\mathbf{u}_2)^2 + 2h_N^2 \sum_{j=1}^{NT_N} (\mathbf{w}_j'\mathbf{u}_1)(\mathbf{z}_j'\mathbf{u}_2) \geq h_N^2 \sum_{j=1}^{NT_N} (\mathbf{z}_j'\mathbf{u}_2)^2 - h_N^2 \sum_{j=1}^{NT_N} \left[(\mathbf{w}_j'\mathbf{u}_1)^2 + (\mathbf{z}_j'\mathbf{u}_2)^2\right]$$
$$= -h_N^2 NT_N \mathbf{u}_1'\Sigma_{1N}\mathbf{u}_1 \geq -\rho_{1N}^{-2} p_N \tau_2 C^2$$

where the last inequality follows from assumption (A2) and the fact that $||\mathbf{u}_1|| \leq C$. Hence,

$$\frac{h_N^2 \sum_{j=1}^{NT_N} (\mathbf{z}_i'\mathbf{u}_2)^2 + 2h_N^2 \sum_{j=1}^{NT_N} (\mathbf{w}_i'\mathbf{u}_1)(\mathbf{z}_i'\mathbf{u}_2)}{p_N \rho_N^{-2}} \geq -\tau_2 C^2$$

Regarding the third term it follows from Jensen's inequality (conditional version)

$$E\left(\left|\sum_{j=1}^{NT_N} (\mathbf{z}_j'\mathbf{u}_2)\varepsilon_j\right| \middle| \mathbf{X}_N\right) \leq \left[E\left(\sum_{j=1}^{NT_N} (\mathbf{z}_j'\mathbf{u}_2)\varepsilon_j\right)^2 \middle| \mathbf{X}_N\right]^{1/2} = \left[\sigma^2 \sum_{j=1}^{NT_N} (\mathbf{z}_j'\mathbf{u}_2)^2\right]^{1/2}$$
$$= \sigma \left[\sum_{j=1}^{NT_N} \mathbf{u}_2'\mathbf{z}_j\mathbf{z}_j'\mathbf{u}_2\right]^{1/2} = \sigma(NT_N)^{1/2}\left(\mathbf{u}_2'\Sigma_{2N}\mathbf{u}_2\right)^{1/2} \leq \sigma(NT_N)^{1/2}\rho_{2N}^{1/2} C$$

where the last inequality used that

$$\rho_{2N} = \max_{\mathbf{u}\in\mathbf{R}^{p_N}} \frac{\mathbf{u}'\Sigma_N\mathbf{u}}{\mathbf{u}'\mathbf{u}} \geq \max_{\mathbf{u}_2\in\mathbf{R}^{m_N}} \frac{(\mathbf{0}',\mathbf{u}_2')\Sigma_N(\mathbf{0}',\mathbf{u}_2')'}{(\mathbf{0}',\mathbf{u}_2')(\mathbf{0}',\mathbf{u}_2')'} = \max_{\mathbf{u}_2\in\mathbf{R}^{m_N}} \frac{\mathbf{u}'\Sigma_{2N}\mathbf{u_2}}{\mathbf{u}_2'\mathbf{u}_2}$$

Hence, since h_N is measurable wrt. $\sigma(\mathbf{X}_N)$,

$$E\left(\left|-2h_N\sum_{j=1}^{NT_N}(\mathbf{z}_j'\mathbf{u}_2)\varepsilon_j\right|\middle|\mathbf{X}_N\right)\leq 2\sigma h_N(NT_N)^{1/2}\rho_{2N}^{1/2}C$$

and so

$$E\left(\frac{\left|-2h_N\sum_{j=1}^{NT_N}(\mathbf{z}_j'\mathbf{u}_2)\varepsilon_j\right|}{h_N(NT_N)^{1/2}\rho_{2N}^{1/2}}\right)\leq 2\sigma C$$

which by Lemma 2 shows that $-2h_N\sum_{j=1}^{NT_N}(\mathbf{z}_j'\mathbf{u}_2)\varepsilon_j$ is $O_p(h_N(NT_N)^{1/2}\rho_{2N}^{1/2})=O_p(\rho_{1N}^{-1}\rho_{2N}^{1/2}p_N^{1/2})$. Hence,

$$\frac{\left|-2h_N\sum_{j=1}^{NT_N}(\mathbf{z}_j'\mathbf{u}_2)\varepsilon_j\right|}{p_N\rho_{1N}^{-2}}\in O_p\left(\frac{\rho_{1N}^{-1}\rho_{2N}^{1/2}p_N^{1/2}}{p_N\rho_{1N}^{-2}}\right)=O_p\left(\frac{\rho_{1N}\rho_{2N}^{1/2}}{p_N^{1/2}}\right)\subseteq O_p(1)$$

by assumption (A7). Regarding the fourth term since $\mathbf{u}_2\neq 0$ we have

$$\lambda_N h_N^{\gamma}/(p_N\rho_{1N}^{-2})=\lambda_N\left[\rho_{1N}^{-1}(p_N/NT_N)^{1/2}\right]^{\gamma}/(p_N\rho_{1N}^{-2})=\lambda_N\rho_{1N}^{2-\gamma}(NT_N)^{-\gamma/2}p_N^{\gamma/2-1}\to\infty$$

by assumption (A4) and so the fourth term diverges to infinity. Since $p_N\rho_{1N}^{-2}\in\Omega_p(1)$ this completes the proof. □

Proof of Theorem 2. The first part has been established in Lemma 3. Since $\hat{\beta}_N$ is consistent it follows from assumption (A5) that for an arbitrary $\varepsilon>0$

$$\begin{aligned}
&P\left(\left\{\min\left\{|\hat{\beta}_{1Nj}|\;|1\leq j\leq k\right\}+\varepsilon<b_0\right\}\right)=P\left(\bigcup_{j=1}^{k}\left\{|\hat{\beta}_{1Nj}|+\varepsilon<b_0\right\}\right)\\
&=P\left(\bigcup_{j=1}^{k}\left\{b_0-|\hat{\beta}_{1Nj}|>\varepsilon\right\}\right)\leq P\left(\bigcup_{j=1}^{k}\left\{|\beta_{10j}|-|\hat{\beta}_{1Nj}|>\varepsilon\right\}\right)\\
&\leq P\left(\bigcup_{j=1}^{k}\left\{|\beta_{10j}-\hat{\beta}_{1Nj}|>\varepsilon\right\}\right)\leq P\left(||\beta_{10}-\hat{\beta}_{1Nj}||>\varepsilon\right)\to 0
\end{aligned}$$

Choosing $\varepsilon=b_0/2$ shows that with probability converging to one $\min\left\{|\hat{\beta}_{1Nj}|\;|1\leq j\leq k\right\}\geq b_0/2$ and so $\hat{\beta}_{1N}$ is bounded away from 0. Hence

L_N is differentiable at $\hat{\beta}_{1N}$ with probability converging to one. And so $\hat{\beta}_{1N}$ satisfies

$$\frac{\partial}{\partial \beta_1} L_N(\hat{\beta}_{1N}, \hat{\beta}_{2N}) = 0$$

That is,

$$-2\sum_{j=1}^{NT_N}\left(y_j - \mathbf{w}_j'\hat{\beta}_{1N} - \mathbf{z}_j'\hat{\beta}_{2N}\right)\mathbf{w}_j + \lambda_N\gamma\psi_N = 0$$

with probability converging to 1 where ψ_N is a $k\times 1$ vector with l'th entry given by $\psi_{Nl} = |\hat{\beta}_{1Nl}|^{\gamma-1}\text{sign}(\hat{\beta}_{1Nl})$. This can be rewritten as

$$-2\sum_{j=1}^{NT_N}\left(\varepsilon_j - \mathbf{w}_j'(\hat{\beta}_{1N} - \beta_{10}) - \mathbf{z}_j'\hat{\beta}_{2N}\right)\mathbf{w}_j + \lambda_N\gamma\psi_N = 0 \Leftrightarrow$$

$$-2\sum_{j=1}^{NT_N}\varepsilon_j\mathbf{w}_j + 2\sum_{j=1}^{NT_N}\mathbf{w}_j\mathbf{w}_j'(\hat{\beta}_{1N} - \beta_{10}) + 2\sum_{j=1}^{NT_N}\mathbf{z}_j'\hat{\beta}_{2N}\mathbf{w}_j + \lambda_N\gamma\psi_N = 0 \Leftrightarrow$$

$$\Sigma_{1N}(\hat{\beta}_{1N} - \beta_{10}) = \frac{1}{NT_N}\sum_{j=1}^{NT_N}\varepsilon_j\mathbf{w}_j - \frac{\lambda_N\gamma\psi_N}{2NT_N} - \frac{1}{NT_N}\sum_{j=1}^{NT_N}\mathbf{z}_j'\hat{\beta}_{2N}\mathbf{w}_j$$

Hence, for any $k\times 1$ vector α

$$(NT_N)^{1/2}\alpha'(\hat{\beta}_{1N} - \beta_{10}) = (NT_N)^{-1/2}\sum_{j=1}^{NT_N}\alpha'\Sigma_{1N}^{-1}\varepsilon_j\mathbf{w}_j$$

$$-(1/2)\gamma(NT_N)^{-1/2}\lambda_N\alpha'\Sigma_{1N}^{-1}\psi_N - (NT_N)^{-1/2}\sum_{j=1}^{NT_N}\alpha'\Sigma_{1N}^{-1}\mathbf{z}_j'\hat{\beta}_{2N}\mathbf{w}_j$$

Since $P(\hat{\beta}_{2N} = 0) \to 1$ the last term equals 0 with probability converging to 1. From the Cauchy-Schwarz inequality in $\mathbf{R}^k$ it follows that

$$|\alpha'\Sigma_{1N}^{-1}\psi_N| \le ||\alpha'\Sigma_{1N}^{-1}||||\psi_N||$$

Since

$$||\alpha'\Sigma_{1N}^{-1}||^2 = \alpha'\Sigma_{1N}^{-1}\Sigma_{1N}^{-1}\alpha = \alpha'\Sigma_{1N}^{-2}\alpha \le \tau_{1N}^{-2}||\alpha||^2 \le \tau_1^{-2}||\alpha||^2$$

by assumption (A2), we get with probability converging to one

$$\begin{aligned}|(1/2)(NT_N)^{-1/2}\gamma\lambda_N\alpha'\Sigma_{1N}^{-1}\psi_N| &\le (1/2)(NT_N)^{-1/2}\gamma\lambda_N\tau_1^{-1}||\alpha||||\psi_N|| \\ &\le (1/2)(NT_N)^{-1/2}\gamma\lambda_N\tau_1^{-1}||\alpha||k^{1/2}(b_0/2)^{(\gamma-1)} \\ &= (1/2)\gamma\tau_1^{-1}||\alpha||(b_0/2)^{(\gamma-1)}T_N^{-1/2}N^{-1/2}\lambda_N k^{1/2} \to 0\end{aligned}$$

by assumption (A3). Hence,

$$(NT_N)^{1/2}\alpha'(\hat{\beta}_{1N}-\beta_{10}) \in (NT_N)^{-1/2}\sum_{j=1}^{NT_N}\alpha'\Sigma_{1N}^{-1}\varepsilon_j\mathbf{w}_j + o_p(1)$$

Since $s_N^{-1} = 1/\sqrt{\sigma^2\alpha'\Sigma_{1N}^{-1}\alpha} \leq \tau_2^{1/2}/(\sigma||\alpha||)$ by assumption (A2) it is also true that

$$(NT_N)^{1/2}s_N^{-1}\alpha'(\hat{\beta}_{1N}-\beta_{10}) \in (NT_N)^{-1/2}s_N^{-1}\sum_{j=1}^{NT_N}\alpha'\Sigma_{1N}^{-1}\varepsilon_j\mathbf{w}_j + o_p(1) \quad (3.15)$$

Now, defining $\mathbf{W}_{i,N} = (\mathbf{w}_{i1},...,\mathbf{w}_{iT_N})'$ and $\varepsilon_{i,N} = (\varepsilon_{i1},...,\varepsilon_{iNT_N})'$ one first notices that[12]

$$\frac{1}{NT_N}\sum_{j=1}^{NT_N}\mathbf{w}_j\mathbf{w}_j' \to \lim_{N\to\infty}\frac{1}{NT_N}\sum_{j=1}^{NT_N}E\left(\mathbf{w}_j\mathbf{w}_j'\right) = \lim_{N\to\infty}E\left(\frac{1}{T_N}\mathbf{W}_{1N}'\mathbf{W}_{1N}\right) \quad (3.16)$$

where the first limit is in probability. To see why (3.16) is true let Z_j be a fixed entry in $\mathbf{w}_j\mathbf{w}_j'$, $j = 1,...,NT_N$. Letting $\eta > 0$ be arbitrary and using the Markov inequality

$$\begin{aligned}
&P\left(\left|\frac{1}{NT_N}\sum_{j=1}^{NT_N}Z_j - \frac{1}{NT_N}\sum_{j=1}^{NT_N}E\left(Z_j\right)\right| > \eta\right) \leq \frac{E\left(\frac{1}{NT_N}\sum_{j=1}^{NT_N}\left[Z_j - E\left(Z_j\right)\right]\right)^2}{\eta^2} \\
&= \frac{E\left(\sum_{i=1}^{N}\sum_{t=1}^{T_N}\left[Z_{it} - E\left(Z_{it}\right)\right]\right)^2}{(NT_N\eta)^2} = \frac{\sum_{i=1}^{N}E\left(\sum_{t=1}^{T_N}\left[Z_{it} - E\left(Z_{it}\right)\right]\right)^2}{(NT_N\eta)^2} \\
&= \frac{N\left(\sum_{t=1}^{T_N}Var(Z_{1t}) + 2\sum_{t=1}^{T_N}\sum_{s>t}^{T_N}Cov\left(Z_{1t},\, Z_{1s}\right)\right)}{(NT_N\eta)^2} \\
&\leq \frac{N\left(T_N\max_{1\leq t\leq T_N}Var\left(Z_{1t}\right) + 2T_N(T_N-1)/2\max_{1\leq t\leq T_N}Var\left(Z_{1t}\right)\right)}{(NT_N\eta)^2} \\
&= \frac{\max_{1\leq t\leq T_N}Var\left(Z_{1t}\right)}{N\eta^2} \to 0
\end{aligned}$$

[12] All limits are taken elementwise in the matrices.

Hence,

$$(NT_N)^{-1/2}s_N^{-1}\sum_{j=1}^{NT_N}\alpha'\Sigma_{1N}^{-1}\varepsilon_j\mathbf{w}_j = \frac{(NT_N)^{-1/2}\sum_{j=1}^{NT_N}\alpha'\left(\frac{1}{NT_N}\sum_{j=1}^{NT_N}\mathbf{w}_j\mathbf{w}_j'\right)^{-1}\varepsilon_j\mathbf{w}_j}{\sqrt{\sigma^2\alpha'\left(\frac{1}{NT_N}\sum_{j=1}^{NT_N}\mathbf{w}_j\mathbf{w}_j'\right)^{-1}\alpha}}$$

$$\in \frac{N^{-1/2}}{\sqrt{\sigma^2\alpha'\left(E\left[\frac{1}{T_N}\mathbf{W}_{1N}'\mathbf{W}_{1N}\right]\right)^{-1}\alpha}}\sum_{i=1}^{N}\alpha'\left(E\left[\frac{1}{T_N}\mathbf{W}_{1N}'\mathbf{W}_{1N}\right]\right)^{-1}T_N^{-1/2}\mathbf{W}_{iN}'\varepsilon_{iN}+o_p(1)$$

Now,

$$E\left(\sum_{i=1}^{N}\alpha'\left(E\left[\frac{1}{T_N}\mathbf{W}_{1N}'\mathbf{W}_{1N}\right]\right)^{-1}T_N^{-1/2}\mathbf{W}_{iN}'\varepsilon_{iN}\right)=0$$

by iterated expectations and

$$\begin{aligned}
r_N^2 :&= E\left[\left(\sum_{i=1}^{N}\alpha'\left(E\left[\frac{1}{T_N}\mathbf{W}_{1N}'\mathbf{W}_{1N}\right]\right)^{-1}T_N^{-1/2}\mathbf{W}_{iN}'\varepsilon_{iN}\right)^2\right]\\
&=\sum_{i=1}^{N}E\left(\alpha'\left(E\left[\frac{1}{T_N}\mathbf{W}_{1N}'\mathbf{W}_{1N}\right]\right)^{-1}T_N^{-1/2}\mathbf{W}_{iN}'\varepsilon_{iN}\varepsilon_{iN}'\mathbf{W}_{iN}T_N^{-1/2}\left(E\left[\frac{1}{T_N}\mathbf{W}_{1N}'\mathbf{W}_{1N}\right]\right)^{-1}\alpha\right)\\
&=\sigma^2N\alpha'\left(E\left[\frac{1}{T_N}\mathbf{W}_{1N}'\mathbf{W}_{1N}\right]\right)^{-1}\alpha
\end{aligned}$$

Finally, let $U_{iN}=\alpha'\left(E\left[\frac{1}{T_N}\mathbf{W}_{1N}'\mathbf{W}_{1N}\right]\right)^{-1}T_N^{-1/2}\mathbf{W}_{iN}'\varepsilon_{iN}$. Since $E(U_{iN}^2)=\sigma^2\alpha'\left(E\left[\frac{1}{T_N}\mathbf{W}_{1N}'\mathbf{W}_{1N}\right]\right)^{-1}\alpha$ is convergent it is bounded. Furthermore,

$$\alpha'E\left(\frac{1}{T_N}\mathbf{W}_{1N}'\mathbf{W}_{1N}\right)\alpha=E\left(\alpha'\frac{1}{NT_N}\mathbf{W}_N'\mathbf{W}_N\alpha\right)\in\left[\alpha'\alpha\tau_1,\alpha'\alpha\tau_2\right]$$

and so the eigenvalues of $E\left[\frac{1}{T_N}\mathbf{W}_{1N}'\mathbf{W}_{1N}\right]$ are also contained in $[\tau_1,\tau_2]$. This implies $r_N^2=\sigma^2N\alpha'\left(E\left[\frac{1}{T_N}\mathbf{W}_{1N}'\mathbf{W}_{1N}\right]\right)^{-1}\alpha\geq\sigma^2N\alpha'\alpha/\tau_2$. Hence, the Lindeberg condition is satisfied since for all $\delta>0$

$$\lim_{N\to\infty}\left\{r_N^{-2}\sum_{i=1}^{N}\int_{\left\{|U_{i,N}|>\delta r_N\right\}}U_{iN}^2dP\right\}\leq\lim_{N\to\infty}\frac{\tau_2}{\alpha'\alpha\sigma^2}\int_{\left\{|U_{1N}|>\delta\sqrt{\frac{\alpha'\alpha\sigma^2}{\tau_2}}\sqrt{N}\right\}}U_{1N}^2dP=0$$

since for any $\rho>0$ (let $\delta\sqrt{\frac{\alpha'\alpha\sigma^2}{\tau_2}}=K$)

$$\lim_{N\to\infty}P\left(U_{1N}^2\mathbf{1}_{\left\{|U_{1N}|>K\sqrt{N}\right\}}>\rho\right)\leq\lim_{N\to\infty}P\left(\left\{|U_{1N}|>K\sqrt{N}\right\}\right)\leq\lim_{N\to\infty}\frac{E\left(U_{1N}^2\right)}{K^2N}=0$$

and $\left\{U_{1N}^2\right\}_{N=1}^{\infty}$ is uniformly integrable[13]. Hence,

$$\frac{N^{-1/2}}{\sqrt{\sigma^2\alpha'\left(E\left[\frac{1}{T_N}\mathbf{W}_{1N}'\mathbf{W}_{1N}\right]\right)^{-1}\alpha}}\sum_{i=1}^{N}\alpha'\left(E\left[\frac{1}{T_N}\mathbf{W}_{1N}'\mathbf{W}_{1N}\right]\right)^{-1}T_N^{-1/2}\mathbf{W}_{iN}'\varepsilon_{iN}\xrightarrow{d}N(0,1)$$

And so by (3.15),

$$(NT_N)^{1/2}s_N^{-1}\alpha'(\hat{\beta}_{1N}-\beta_{10})\xrightarrow{d}N(0,1)$$

or equivalently,

$$(NT_N)^{1/2}(\hat{\beta}_{1N}-\beta_{10})\xrightarrow{d}N\left(0,\sigma^2\left(\lim_{N\to\infty}E\left[\frac{1}{T_N}\mathbf{W}_{1N}'\mathbf{W}_{1N}\right]\right)^{-1}\right)\qquad(3.17)$$

□

Proof of Theorem 3. (i) If $T_N=T$ for a fixed T, $U_{1N}=U_1$ for all N where U_1 is defined in the obvious way, does not depend on N and belongs to L^2. Hence,

$$\lim_{K\to\infty}\sup_{1\leq N<\infty}\int_{\left\{|U_{1N}|>K\right\}}U_{1N}^2dP=\lim_{K\to\infty}\int_{\left\{|U_1|>K\right\}}U_1^2dP=0$$

by Lebesgue's Dominated Convergence Theorem.

[13]The uniform integrability of $\left\{U_{1N}^2\right\}_{N=1}^{\infty}$ implies that $\left\{U_{1N}^2\mathbf{1}_{\left\{|U_{1N}|>K\sqrt{N}\right\}}\right\}_{N=1}^{\infty}$ is uniformly integrable which is what is needed to utilize the well known fact that convergence in measure plus uniform integrability is equivalent to convergence in L^1.

(ii) By the Cauchy-Schwarz inequality

$$|U_{1N}| \leq \left\| \alpha' \left(E\left[\frac{1}{T_N} \mathbf{W}'_{1N} \mathbf{W}_{1N} \right] \right)^{-1} \right\| \left\| T_N^{-1/2} \mathbf{W}'_{1N} \varepsilon_{1N} \right\|$$

Since $\left\| \alpha' \left(E\left[\frac{1}{T_N} \mathbf{W}'_{1N} \mathbf{W}_{1N} \right] \right)^{-1} \right\|$ is convergent it is bounded by a constant C. Hence,

$$U_{1N}^2 \leq C^2 \left\| T_N^{-1/2} \mathbf{W}'_{1N} \varepsilon_{1N} \right\|^2 = C^2 \sum_{i=1}^{k} \frac{1}{T_N} \sum_{j=1}^{T_N} \left(\mathbf{W}_{1N}^{ji} \varepsilon_{1N}^{j} \right)^2$$

where $\mathbf{W}_{1N}^{ji}$ is jth row in the ith column of $\mathbf{W}_{1N}$. Since the rows of $\mathbf{W}_{1N}$ are identically distributed and $\mathbf{W}_{1N}^{ji} \perp\!\!\!\perp \varepsilon_{1N}^{j}$ one has (calculate the characteristic functions and notice they are identical) $\mathbf{W}_{1N}^{ji} \varepsilon_{1N}^{j} \sim \mathbf{W}_{1N}^{1i} \varepsilon_{1N}^{1}$ where $\mathbf{W}_{1N}^{1i} \varepsilon_{1N}^{1} \sim Z_i$ for some $Z_i \in L^2$. Hence, $\left\{ \left\{ \left(\mathbf{W}_{1N}^{ji} \varepsilon_{1N}^{j} \right)^2 \right\}_{j=1}^{T_N} \right\}_{N=1}^{\infty}$ is uniformly integrable for all $1 \leq i \leq k$ since

$$\begin{aligned}
&\lim_{K\to\infty} \sup_{1\leq N<\infty} \sup_{1\leq j\leq T_N} \int_{\left\{ \left(\mathbf{W}_{1N}^{ji} \varepsilon_{1N}^{j} \right)^2 > K \right\}} \left(\mathbf{W}_{1N}^{ji} \varepsilon_{1N}^{j} \right)^2 dP \\
&= \lim_{K\to\infty} \sup_{1\leq N<\infty} \sup_{1\leq j\leq T_N} \int_{\left\{ x^2 > K \right\}} x^2 dP_{\mathbf{W}_{1N}^{ji} \varepsilon_{1N}^{j}}(x) \\
&= \lim_{K\to\infty} \sup_{1\leq N<\infty} \sup_{1\leq j\leq T_N} \int_{\left\{ x^2 > K \right\}} x^2 dP_{Z_i}(x) \\
&= \lim_{K\to\infty} \int_{\left\{ x^2 > K \right\}} x^2 dP_{Z_i}(x) = 0
\end{aligned}$$

by Lebesgue's Dominated Convergence Theorem. By Hoffmann-Jørgensen (1994) (page 338)[14] this implies that $\left\{ \frac{1}{T_N} \sum_{j=1}^{T_N} \left(\mathbf{W}_{1N}^{ji} \varepsilon_{1N}^{j} \right)^2 \right\}_{N=1}^{\infty}$ is uniformly integrable which in turn implies that $\left\{ \sum_{i=1}^{k} \frac{1}{T_N} \sum_{j=1}^{T_N} \left(\mathbf{W}_{1N}^{ji} \varepsilon_{1N}^{j} \right)^2 \right\}_{N=1}^{\infty}$ is uniformly integrable by Hoffmann-Jørgensen (1994) (page 337)[15]. Since

[14]The partial averages of a uniformly integrable sequence are themselves uniformly integrable.

[15]Finite sums of uniformly integrable sequences are themselves uniformly integrable.

$\left\{C^2\sum_{i=1}^{k}\frac{1}{T_N}\sum_{j=1}^{T_N}\left(\mathbf{W}_{1N}^{ji}\varepsilon_{1N}^{j}\right)^2\right\}_{N=1}^{\infty}$ dominates $\left\{U_{1N}^2\right\}$ this yields the desired result.

(iii) If $\mathbf{W}_{1N}$ and ε_{1N} are uniformly bounded U_{1N} has moments of any order and so $\left\{U_{1N}^2\right\}_{N=1}^{\infty}$ is uniformly integrable. □

Proof of Corollary 1. For fixed T_N (3.17) reads

$$N^{1/2}(\hat{\beta}_{1N}-\beta_{10})\xrightarrow{d}N\left(0,\sigma^2\left(E\left[\mathbf{W}_1'\mathbf{W}_1\right]\right)^{-1}\right)$$

where absence of subscript N indicates that the matrices no longer depend on T_N. In the fixed effects setting $\mathbf{W}_1=\left(\mathbf{D}\mathbf{D}'\right)^{-1/2}\mathbf{D}\tilde{\mathbf{W}}_1$ and so

$$\begin{aligned}E\left(\mathbf{W}_1'\mathbf{W}_1\right)&=E\left(\tilde{\mathbf{W}}_1'\mathbf{D}'\left(\mathbf{D}\mathbf{D}'\right)^{-1/2}\left(\mathbf{D}\mathbf{D}'\right)^{-1/2}\mathbf{D}\tilde{\mathbf{W}}_1\right)\\&=E\left(\tilde{\mathbf{W}}_1'\mathbf{D}'\left(\mathbf{D}\mathbf{D}'\right)^{-1}\mathbf{D}\tilde{\mathbf{W}}_1\right)=E\left(\ddot{\tilde{\mathbf{W}}}_1'\ddot{\tilde{\mathbf{W}}}_1\right)\end{aligned}$$

where the last inequality used that $\mathbf{D}'\left(\mathbf{D}\mathbf{D}'\right)^{-1}\mathbf{D}$ is symmetric and idempotent and that premultiplication of it corresponds to columnwise demeaning.

The proof of part (ii) is similar using $\mathbf{W}_1=\Omega^{-1/2}\tilde{\mathbf{W}}_1$. □

Next we turn to the properties of the Marginal Bridge estimator.

Lemma 4. *For q even and any $w_N>0$,*

$$P\left(w_N>\max_{1\le k\le m}\left|\sum_{j=1}^{NT}x_{jk}\varepsilon_j\right|\right)\ge 1-\frac{Km(NT_N)^{q/2}}{w_N^q}$$

Proof. By the Markov inequality,

$$\begin{aligned}P\left(w_N>\max_{1\le k\le m}\left|\sum_{j=1}^{NT}x_{jk}\varepsilon_j\right|\right)&=1-P\left(\max_{1\le k\le m}\left|\sum_{j=1}^{NT_N}x_{jk}\varepsilon_j\right|\ge w_N\right)\\&\ge 1-\frac{E\left(\max_{1\le k\le m}\left|\sum_{j=1}^{NT_N}x_{jk}\varepsilon_j\right|^q\right)}{w_N^q}\end{aligned}$$

Since for any sequence of random variables $\{Z_k\}_{k=1}^m\subseteq L^q$

$$E\left(\max_{1\le k\le m}Z_k^q\right)\le E\left(\sum_{k=1}^{m}Z_k^q\right)\le m\max_{1\le k\le m}E\left(Z_k^q\right)$$

it suffices to show that $E\left(\sum_{j=1}^{NT_N} x_{jk}\varepsilon_j\right)^q \leq K(NT_N)^{q/2}$ for all $k \in \{1,...,m\}$. By the multinomial theorem

$$E\left(\sum_{j=1}^{NT_N} x_{jk}\varepsilon_j\right)^q = E\left(\sum_{|\alpha|=q} \binom{q}{\alpha} \prod_{j=1}^{NT_N} (x_{jk}\varepsilon_j)^{\alpha_j}\right)$$

where we have used the multi-index convention $\alpha = (\alpha_1,...,\alpha_{NT})$ and $|\alpha| = q$ means $\sum_{j=1}^{NT_N} \alpha_j = q$. The sum is taken over all such vectors α whose entries add up to q. Since $(\mathbf{X}_{iN}, \varepsilon_{iN})$ is i.i.d. and $\mathbf{X}_{iN} \perp\!\!\!\perp \varepsilon_{iN}$ it follows that $(\mathbf{X}_{1N}, \varepsilon_{1N}, ..., \mathbf{X}_{NN}, \varepsilon_{NN})$ are independent (calculate the characteristic function and observe that it factorizes). This implies $\mathbf{X}_N \perp\!\!\!\perp \varepsilon_N$. Finally, it is seen that $\varepsilon_{11}, ..., \varepsilon_{NT_N}$ are independent and (again calculate the characteristic function and observe that it factorizes by using $(\mathbf{X}_{iN}, \varepsilon_{iN})$ is i.i.d and the that uncorrelatedness implies independence for gaussian variables.). From these observations it follows that all summands including an $\alpha_j = 1$ for some $j = 1,...,NT_N$ equal 0 since $E(\varepsilon_j) = 0$ for all $j = 1,...,NT_N$. It remains to be shown that no other term is of greater order than $O\left((NT_N)^{q/2}\right)$. So assume $\min_{1\leq j\leq NT_N} \alpha_j \geq 2$ [16]. Consider the set A of all vectors α with s entries different from 0 and larger than or equal to 2 with non-zero values $a_1, ..., a_s$ and $a_1 + ... + a_s = q$.

[16]Here we follow the convention of setting $t^0 = 1$ even when $t = 0$ and so we only have to consider the case $\min_{1\leq j\leq NT_N} \alpha_j \geq 2$.

$$
\begin{aligned}
&E\left(\sum_{\alpha\in A}\prod_{j=1}^{NT_N}(x_{jk}\varepsilon_j)^{\alpha_j}\right)\\
&=E\left(\sum_{j_1=1}^{NT_N}\sum_{j_2\neq j_1}^{NT_N}\ldots\sum_{j_s\notin\{j_1,\ldots,j_{s-1}\}}^{NT_N}(x_{j_1k}\varepsilon_{j_1})^{a_1}(x_{j_2k}\varepsilon_{j_2})^{a_2}\cdot\ldots\cdot(x_{j_sk}\varepsilon_{j_s})^{a_s}\right)\\
&=KE\left(\sum_{j_1=1}^{NT_N}\sum_{j_2\neq j_1}^{NT_N}\ldots\sum_{j_s\notin\{j_1,\ldots,j_{s-1}\}}^{NT_N}(x_{j_1k})^{a_1}(x_{j_2k})^{a_2}\cdot\ldots\cdot(x_{j_sk})^{a_s}\right)\\
&\leq K(NT_N)^{q/2}E\left(\sum_{j_1=1}^{NT_N}\sum_{j_2\neq j_1}^{NT_N}\ldots\sum_{j_s\notin\{j_1,\ldots,j_{s-1}\}}^{NT_N}\left[\frac{x_{j_1k}^2}{NT_N}\right]^{a_1/2}\left[\frac{x_{j_2k}^2}{NT_N}\right]^{a_2/2}\cdot\ldots\cdot\left[\frac{x_{j_sk}^2}{NT_N}\right]^{a_s/2}\right)\\
&=K(NT_N)^{q/2}E\left(\sum_{j_1=1}^{NT_N}\left[\frac{x_{j_1k}^2}{NT_N}\right]^{a_1/2}\sum_{j_2\neq j_1}^{NT_N}\left[\frac{x_{j_2k}^2}{NT_N}\right]^{a_2/2}\cdot\ldots\cdot\sum_{j_s\notin\{j_1,\ldots,j_{s-1}\}}^{NT_N}\left[\frac{x_{j_sk}^2}{NT_N}\right]^{a_s/2}\right)\\
&\leq K(NT_N)^{q/2}
\end{aligned}
$$

where the second equality used $\mathbf{X}_N \perp\!\!\!\perp \varepsilon_N$ and $K := E(\varepsilon_1^{a_1})\cdot\ldots\cdot E(\varepsilon_1^{a_s})$. The first estimate follows from the nonnegativity of all terms on the right hand side. The last estimate uses that for any subset S of $\{1,\ldots,NT_N\}$, $\sum_{j\in S}\left[x_{jk}^2/(NT_N)\right]^{a_i/2} \leq \sum_{j=1}^{NT_N}\left[x_{jk}^2/(NT_N)\right]^{a_i/2} \leq \sum_{j=1}^{NT_N}\left[x_{jk}^2/(NT_N)\right] = 1$ since $\left[x_{jk}^2/(NT_N)\right] \leq 1$ and so $\left[x_{jk}^2/(NT_N)\right]^{a_i/2} \leq \left[x_{jk}^2/(NT_N)\right]$ for $a_i \geq 2$. Since s was arbitrary and there are only finitely many terms of the above kind this establishes the lemma.

□

Proof of Theorem 4. Let $\varepsilon_N = (\varepsilon_1,\ldots,\varepsilon_{NT_N})'$ and recall $\xi_{Nk} =$

$1/(NT_N)\sum_{j=1}^{NT_N}\mathbf{w}_j'\beta_{10}x_{jk}$. Then,

$$
\begin{aligned}
U_N(\beta) &= \sum_{k=1}^{p_N}\sum_{j=1}^{NT_N}\left(y_j - x_{jk}\beta_k\right)^2 + \lambda_N\sum_{k=1}^{p_N}|\beta_k|^\gamma \\
&= \sum_{k=1}^{p_N}\left[\sum_{j=1}^{NT_N} y_j^2 + NT_N\beta_k^2 - 2\sum_{j=1}^{NT_N} y_j x_{jk}\beta_k\right] + \lambda_N\sum_{k=1}^{p_N}|\beta_k|^\gamma \\
&= \sum_{k=1}^{p_N}\left[\sum_{j=1}^{NT_N}\varepsilon_j^2 + \sum_{j=1}^{NT_N}(\mathbf{w}_j'\beta_{01})^2 + 2\sum_{j=1}^{NT_N}(\mathbf{w}_j'\beta_{01})\varepsilon_j + NT_N\beta_k^2 - 2\sum_{j=1}^{NT_N}(\varepsilon_j + \mathbf{w}_j'\beta_{01})x_{jk}\beta_k\right] \\
&+ \lambda_N\sum_{k=1}^{p_N}|\beta_k|^\gamma \\
&= \sum_{k=1}^{p_N}\left[\sum_{j=1}^{NT_N}\varepsilon_j^2 + \sum_{j=1}^{NT_N}(\mathbf{w}_j'\beta_{01})^2 + 2\sum_{j=1}^{NT_N}(\mathbf{w}_j'\beta_{01})\varepsilon_j + NT_N\beta_k^2 - 2(\varepsilon_N'\mathbf{x}_k + NT_N\xi_{Nk})\beta_k\right] \\
&+ \lambda_N\sum_{k=1}^{p_N}|\beta_k|^\gamma
\end{aligned}
$$

So minimizing $U_N(\beta)$ is equivalent to minimizing $\sum_{k=1}^{p_N}\left[NT_N\beta_k^2 - 2(\varepsilon_N'\mathbf{x}_k + NT_N\xi_{Nk})\beta_k + \lambda_N|\beta_k|^\gamma\right]$. Since $0<\gamma<1$ it follows from Lemma A of Knight and Fu (2000) that $\beta_k = 0$ if and only if

$$\lambda_N/(NT_N) > |\varepsilon_N'\mathbf{x}_k/(NT_N) + \xi_{Nk}|^{2-\gamma}c_\gamma$$

where $c_\gamma = \left(\frac{2}{2-\gamma}\right)\left(\frac{2(2-\gamma)}{2-\gamma}\right)^{1-\gamma}$. Defining $w_N = c_\gamma^{-1/(2-\gamma)}(\lambda_N/(NT_N)^{\gamma/2})^{1/(2-\gamma)}$ the above inequality is equivalent to

$$w_N > (NT_N)^{-1/2}|\varepsilon_N'\mathbf{x}_k + (NT_N)\xi_{Nk}|$$

So to prove the theorem it is enough to show

$$P\left(w_N > (NT_N)^{-1/2}\max_{k\in J_N}\left|\varepsilon_N'\mathbf{x}_k + (NT_N)\xi_{Nk}\right|\right) \to 1 \tag{3.18}$$

and

$$P\left(w_N > (NT_N)^{-1/2}\min_{k\in K_N}\left|\varepsilon_N'\mathbf{x}_k + (NT_N)\xi_{Nk}\right|\right) \to 0 \tag{3.19}$$

We first prove (3.18). On $A_N = \left\{ \frac{\left|\sum_{j=1}^{NT_N} x_{jk}x_{jl}\right|}{(NT_N)^{1/2}} \le c_0,\ k \in K_N, l \in J_N \right\}$ and under assumption (B6)

$$\max_{l\in J_N}(NT_N)^{1/2}\left|\xi_{Nl}\right| = \max_{l\in J_N}\frac{1}{(NT_N)^{1/2}}\left|\sum_{j=1}^{NT_N}\sum_{k=1}^{k_N} x_{jk}\beta_{10k}x_{jl}\right|$$
$$= \max_{l\in J_N}\frac{1}{(NT_N)^{1/2}}\left|\sum_{k=1}^{k_N}\beta_{10k}\sum_{j=1}^{NT_N} x_{jk}x_{jl}\right| \le \max_{l\in J_N}\sum_{k=1}^{k_N}|\beta_{10k}|\left|\frac{1}{(NT_N)^{1/2}}\sum_{j=1}^{NT_N} x_{jk}x_{jl}\right|$$
$$\le b_1c_0k_N$$

Hence,

$$P\left(w_N > (NT_N)^{-1/2}\max_{k\in J_N}|\varepsilon_N'\mathbf{x}_k + (NT_N)\xi_{Nk}|\right)$$
$$\ge P\left(w_N > (NT_N)^{-1/2}\max_{k\in J_N}|\varepsilon_N'\mathbf{x}_k| + (NT_N)^{1/2}\max_{k\in J_N}|\xi_{Nk}|,\ A_N\right)$$
$$\ge P\left(w_N > (NT_N)^{-1/2}\max_{k\in J_N}|\varepsilon_N'\mathbf{x}_k| + b_1c_0k_N,\ A_N\right)$$
$$\ge P\left(w_N > (NT_N)^{-1/2}\max_{k\in J_N}|\varepsilon_N'\mathbf{x}_k| + b_1c_0k_N\right) + P(A_N) - 1$$

where the last estimate follows from the inclusion-exclusion principle. By Lemma 4 for every even q,

$$P\left((NT_N)^{1/2}(w_N - b_1c_0k_N) > \max_{k\in J_N}|\varepsilon_N'\mathbf{x}_k|\right) \ge 1 - \frac{Km_N(NT_N)^{q/2}}{\left((NT_N)^{1/2}[w_N - b_1c_0k_N]\right)^q}$$
$$= 1 - \frac{Km_N/w_N^q}{\left(1 - b_1c_0k_N/w_N\right)^q}$$

Furthermore, by assumption (B4) we may choose q sufficiently large such that

$$m_N/w_N^q \in O(1)\frac{m_N}{\left(\lambda_N(NT_N)^{-\gamma/2}\right)^{q/(2-\gamma)}} \to 0$$

and by assumption (B3)

$$b_1c_0k_N/w_N \in O(1)\frac{k_N}{\left(\lambda_N(NT_N)^{-\gamma/2}\right)^{1/(2-\gamma)}} \to 0$$

Finally, $P(A_N)$ can be made arbitrarily close to 1 by assumption (B5) which establishes (3.18). Next we verify (3.19).

$$\begin{aligned}
&P\left(w_N > (NT_N)^{-1/2} \min_{k\in K_N} |\varepsilon_N'\mathbf{x}_k + (NT_N)\xi_{Nk}|\right) \\
&= P\left(\bigcup_{k\in K_N} \left\{w_N > |(NT_N)^{-1/2}\varepsilon_N'\mathbf{x}_k + (NT_N)^{1/2}\xi_{Nk}|\right\}\right) \\
&\leq \sum_{k\in K_N} P\left(\left\{w_N > |(NT_N)^{-1/2}\varepsilon_N'\mathbf{x}_k + (NT_N)^{1/2}\xi_{Nk}|\right\}\right) \qquad (3.20)
\end{aligned}$$

Since $\min_{k\in K_N} |\xi_{Nk}| \geq \xi_0 > 0$ by assumption (B1) we may write,

$$\begin{aligned}
&P\left(\left\{w_N > |(NT_N)^{-1/2}\varepsilon_N'\mathbf{x}_k + (NT_N)^{1/2}\xi_{Nk}|\right\}\right) \\
&\leq P\left(\left\{w_N > (NT_N)^{1/2}|\xi_{Nk}| - (NT_N)^{-1/2}|\varepsilon_N'\mathbf{x}_k|\right\}\right) \\
&\leq P\left(\left\{w_N > (NT_N)^{1/2}\xi_0 - (NT_N)^{-1/2}|\varepsilon_N'\mathbf{x}_k|\right\}\right) \\
&= P\left(\left\{(NT_N)^{-1/2}|\varepsilon_N'\mathbf{x}_k| > (NT_N)^{1/2}\xi_0 - w_N\right\}\right) \\
&= 1 - P\left(\left\{(NT_N)^{1/2}\xi_0 - w_N \geq (NT_N)^{-1/2}|\varepsilon_N'\mathbf{x}_k|\right\}\right) \\
&\leq 1 - P\left(\left\{(NT_N)^{1/2}\xi_0 - w_N > (NT_N)^{-1/2}|\varepsilon_N'\mathbf{x}_k|\right\}\right)
\end{aligned}$$

By Lemma 4,

$$P\left(\left\{(NT_N)^{1/2}\left((NT_N)^{1/2}\xi_0 - w_N\right) > |\varepsilon_N'\mathbf{x}_k|\right\}\right) \geq 1 - \frac{K(NT_N)^{q/2}}{\left((NT_N)^{1/2}[(NT_N)^{1/2}\xi_0 - w_N]\right)^q} \qquad (3.21)$$

$$= 1 - \frac{K}{((NT_N)^{1/2}\xi_0 - w_N)^q} \qquad (3.22)$$

And so for all $k \in K_N$

$$P\left(\left\{w_N > |(NT_N)^{-1/2}\varepsilon_N'\mathbf{x}_k + (NT_N)^{1/2}\xi_{Nk}|\right\}\right) \leq \frac{K}{((NT_N)^{1/2}\xi_0 - w_N)^q}$$

Inserting this into (3.20) yields

$$P\left(w_N > (NT_N)^{-1/2} \min_{k \in k_N} |\varepsilon_N' \mathbf{x}_k + (NT_N)\xi_{Nk}|\right) \leq \frac{Kk_N}{((NT_N)^{1/2}\xi_0 - w_N)^q} = \frac{Kk_N/(NT_N)^{q/2}}{\left(\xi_0 - w_N/(NT_N)^{1/2}\right)^q}$$

By assumption (B2),

$$\frac{w_N}{(NT_N)^{1/2}} \in O(1)\left(\frac{\lambda_N (NT_N)^{-\gamma/2}}{(NT_N)^{(2-\gamma)/2}}\right)^{1/(2-\gamma)} = O(1)\left(\lambda_N/(NT_N)\right)^{1/(2-\gamma)} \subseteq o(1)$$

Furthermore, for any $q \geq 1$, $k_N/(NT_N)^{q/2} \to 0$ by (B3). Hence,

$$P\left(w_N > (NT_N)^{-1/2} \min_{k \in K_N} |\varepsilon_N' \mathbf{x}_k + (NT_N)\xi_{Nk}|\right) \to 0.$$

□

3.8 Bibliography

Arellano, M. (2003). *Panel data econometrics*. Oxford University Press, Oxford.

Candes, E. and T. Tao (2007). The Dantzig selector: statistical estimation when p is much larger than n. *Annals of Statistics 35*, 2313–2351.

Fan, J. and J. Lv (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70*, 849–911.

Friedman, J., T. Hastie, H. Höfling, and R. Tibshirani (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics 1*(2), 302–332.

Hoffmann-Jørgensen, J. (1994). *Probability with a View Toward Statistics, Volume 1*. Chapman and Hall Probability Series, New-York: Chapman & Hall.

Huang, J., J. Horowitz, and S. Ma (2008). Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Annals of Statistics 36*, 587–613.

Hunter, D. and R. Li (2005). Variable selection using MM algorithms. *Annals of Statistics 33*, 1617.

Knight, K. and W. Fu (2000). Asymptotics for lasso-type estimators. *Annals of Statistics 28*, 1356–1378.

Meinshausen, N. and B. Yu (2009). Lasso-type recovery of sparse representations for high-dimensional data. *Annals of Statistics 37*, 246–270.

Stoyanov, J. (1997). *Counterexamples in probability* (2 ed.). Chichester-New York.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological) 58*, 267–288.

Van der Vaart, A. and J. Wellner (1996). *Weak convergence and empirical processes*. Springer Verlag, New York.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association 101*, 1418–1429.

SCHOOL OF ECONOMICS AND MANAGEMENT

AARHUS UNIVERSITY – BARTHOLINS ALLÉ 10 – BUILDING 1322
DK-8000 AARHUS C – TEL. +45 8942 1111 - www.econ.au.dk

PhD Theses:

1999-4	Philipp J.H. Schröder, Aspects of Transition in Central and Eastern Europe.
1999-5	Robert Rene Dogonowski, Aspects of Classical and Contemporary European Fiscal Policy Issues.
1999-6	Peter Raahauge, Dynamic Programming in Computational Economics.
1999-7	Torben Dall Schmidt, Social Insurance, Incentives and Economic Integration.
1999	Jørgen Vig Pedersen, An Asset-Based Explanation of Strategic Advantage.
1999	Bjarke Jensen, Five Essays on Contingent Claim Valuation.
1999	Ken Lamdahl Bechmann, Five Essays on Convertible Bonds and Capital Structure Theory.
1999	Birgitte Holt Andersen, Structural Analysis of the Earth Observation Industry.
2000-1	Jakob Roland Munch, Economic Integration and Industrial Location in Unionized Countries.
2000-2	Christian Møller Dahl, Essays on Nonlinear Econometric Time Series Modelling.
2000-3	Mette C. Deding, Aspects of Income Distributions in a Labour Market Perspective.
2000-4	Michael Jansson, Testing the Null Hypothesis of Cointegration.
2000-5	Svend Jespersen, Aspects of Economic Growth and the Distribution of Wealth.
2001-1	Michael Svarer, Application of Search Models.
2001-2	Morten Berg Jensen, Financial Models for Stocks, Interest Rates, and Options: Theory and Estimation.
2001-3	Niels C. Beier, Propagation of Nominal Shocks in Open Economies.
2001-4	Mette Verner, Causes and Consequences of Interrruptions in the Labour Market.
2001-5	Tobias Nybo Rasmussen, Dynamic Computable General Equilibrium Models: Essays on Environmental Regulation and Economic Growth.

2001-6 Søren Vester Sørensen, Three Essays on the Propagation of Monetary Shocks in Open Economies.

2001-7 Rasmus Højbjerg Jacobsen, Essays on Endogenous Policies under Labor Union Influence and their Implications.

2001-8 Peter Ejler Storgaard, Price Rigidity in Closed and Open Economies: Causes and Effects.

2001 Charlotte Strunk-Hansen, Studies in Financial Econometrics.

2002-1 Mette Rose Skaksen, Multinational Enterprises: Interactions with the Labor Market.

2002-2 Nikolaj Malchow-Møller, Dynamic Behaviour and Agricultural Households in Nicaragua.

2002-3 Boriss Siliverstovs, Multicointegration, Nonlinearity, and Forecasting.

2002-4 Søren Tang Sørensen, Aspects of Sequential Auctions and Industrial Agglomeration.

2002-5 Peter Myhre Lildholdt, Essays on Seasonality, Long Memory, and Volatility.

2002-6 Sean Hove, Three Essays on Mobility and Income Distribution Dynamics.

2002 Hanne Kargaard Thomsen, The Learning organization from a management point of view - Theoretical perspectives and empirical findings in four Danish service organizations.

2002 Johannes Liebach Lüneborg, Technology Acquisition, Structure, and Performance in The Nordic Banking Industry.

2003-1 Carter Bloch, Aspects of Economic Policy in Emerging Markets.

2003-2 Morten Ørregaard Nielsen, Multivariate Fractional Integration and Cointegration.

2003 Michael Knie-Andersen, Customer Relationship Management in the Financial Sector.

2004-1 Lars Stentoft, Least Squares Monte-Carlo and GARCH Methods for American Options.

2004-2 Brian Krogh Graversen, Employment Effects of Active Labour Market Programmes: Do the Programmes Help Welfare Benefit Recipients to Find Jobs?

2004-3 Dmitri Koulikov, Long Memory Models for Volatility and High Frequency Financial Data Econometrics.

2004-4 René Kirkegaard, Essays on Auction Theory.

2004-5 Christian Kjær, Essays on Bargaining and the Formation of Coalitions.

2005-1 Julia Chiriaeva, Credibility of Fixed Exchange Rate Arrangements.

2005-2 Morten Spange, Fiscal Stabilization Policies and Labour Market Rigidities.

2005-3 Bjarne Brendstrup, Essays on the Empirical Analysis of Auctions.

2005-4 Lars Skipper, Essays on Estimation of Causal Relationships in the Danish Labour Market.

2005-5 Ott Toomet, Marginalisation and Discouragement: Regional Aspects and the Impact of Benefits.

2005-6 Marianne Simonsen, Essays on Motherhood and Female Labour Supply.

2005 Hesham Morten Gabr, Strategic Groups: The Ghosts of Yesterday when it comes to Understanding Firm Performance within Industries?

2005 Malene Shin-Jensen, Essays on Term Structure Models, Interest Rate Derivatives and Credit Risk.

2006-1 Peter Sandholt Jensen, Essays on Growth Empirics and Economic Development.

2006-2 Allan Sørensen, Economic Integration, Ageing and Labour Market Outcomes

2006-3 Philipp Festerling, Essays on Competition Policy

2006-4 Carina Sponholtz, Essays on Empirical Corporate Finance

2006-5 Claus Thrane-Jensen, Capital Forms and the Entrepreneur – A contingency approach on new venture creation

2006-6 Thomas Busch, Econometric Modeling of Volatility and Price Behavior in Asset and Derivative Markets

2007-1 Jesper Bagger, Essays on Earnings Dynamics and Job Mobility

2007-2 Niels Stender, Essays on Marketing Engineering

2007-3 Mads Peter Pilkjær Harmsen, Three Essays in Behavioral and Experimental Economics

2007-4 Juanna Schrøter Joensen, Determinants and Consequences of Human Capital Investments

2007-5 Peter Tind Larsen, Essays on Capital Structure and Credit Risk

2008-1	Toke Lilhauge Hjortshøj, Essays on Empirical Corporate Finance – Managerial Incentives, Information Disclosure, and Bond Covenants
2008-2	Jie Zhu, Essays on Econometric Analysis of Price and Volatility Behavior in Asset Markets
2008-3	David Glavind Skovmand, Libor Market Models - Theory and Applications
2008-4	Martin Seneca, Aspects of Household Heterogeneity in New Keynesian Economics
2008-5	Agne Lauzadyte, Active Labour Market Policies and Labour Market Transitions in Denmark: an Analysis of Event History Data
2009-1	Christian Dahl Winther, Strategic timing of product introduction under heterogeneous demand
2009-2	Martin Møller Andreasen, DSGE Models and Term Structure Models with Macroeconomic Variables
2009-3	Frank Steen Nielsen, On the estimation of fractionally integrated processes
2009-4	Maria Knoth Humlum, Essays on Human Capital Accumulation and Educational Choices
2009-5	Yu Wang, Economic Analysis of Open Source Software
2009-6	Benjamin W. Blunck, Creating and Appropriating Value from Mergers and Acquisitions – A Merger Wave Perspective
2009-7	Rune Mølgaard, Essays on Dynamic Asset Allocation and Electricity Derivatives
2009-8	Jens Iversen, Financial Integration and Real Exchange Rates: Essays in International Macroeconomics
2009-9	Torben Sørensen, Essays on Human Capital and the Labour Market
2009-10	Jonas Staghøj, Essays on Unemployment and Active Labour Market Policies
2009-11	Rune Majlund Vejlin, Essays in Labor Economics
2009-12	Nisar Ahmad, Three Essays on Empirical Labour Economics
2010-1	Jesper Rosenberg Hansen, Essays on Application of Management Theory in Public Organizations - Changes due to New Public Management
2010-2	Torben Beedholm Rasmussen, Essays on Dynamic Interest Rate Models and Tests for Jumps in Asset Prices

2010-3 Anders Ryom Villadsen, Sources of Organizational Similarity and Divergence: Essays on Decision Making and Change in Danish Municipalities Using Organizational Institutionalism

2010-4 Eske Stig Hansen, Essays in Electricity Market Modeling

2010-5 Marías Halldór Gestsson, Essays on Macroeconomics and Economic Policy

2010-6 Niels Skipper, Essays on the Demand for Prescription Drugs

2010-7 Thomas Quistgaard Pedersen, Return predictability and dynamic asset allocation

2011-1 Malene Kallestrup Lamb, Health, Retirement and Mortality

2011-2 Lene Kromann, Essays on Productivity and Labour Supply

2011-3 Firew Bekele Woldeyes, Essays on Government Budgets in Developing Countries

2011-4 Anders Bredahl Kock, Forecasting and Oracle Efficient Econometrics