PARAMETRIC AND NON-PARAMETRIC SYSTEM MODELLING

Henrik Aalborg Nielsen

Department of Mathematical Modelling
Technical University of Denmark
Ph.D. Thesis No. 70
Kgs. Lyngby 1999

IMM

ISSN 0909-3192

This document was prepared with LaTeX and printed by Jespersen Offset.

Preface

This thesis was prepared at the Department of Mathematical Modelling, the Technical University of Denmark in partial fulfillment of the requirements for acquiring the Ph.D. degree in engineering.

The thesis deals with different aspects of mathematical modelling of systems using data and partial knowledge about the structure of the systems. The main focus is on extensions of non-parametric methods, but also stochastic differential equations and neural networks are considered.

The thesis consists of a summary report and a collection of ten research papers written during the period 1996–1999, and elsewhere published.

Lyngby, December 1999

Henrik Nielsen

Henrik Aalborg Nielsen

Acknowledgements

In carrying out the work described in this thesis I have received important assistance from many people. First of all I want to address my gratitude to my supervisors Prof. Henrik Madsen from my own department and Prof. Jan Holst from Mathematical Statistics at Lund University for their help and guidance and general willingness to enter into discussions. Here at the department Henrik has over a little more than a decade build up a well-working group within the field of applied statistics and with many contacts to the industry.

Thanks also to my colleagues at the department for their invaluable cooperation, help, and discussions. Especially, I would like to thank Research Assistant Prof. Torben Skov Nielsen with whom I have shared office since I started at the department. Besides many other subjects, Torben and I have discussed the methods in this thesis extensively. Torben is also of invaluable help when I challenge my skills within the field of C programming.

Finally, I would like to express my sincere thanks to the Energy Research Program of the Danish Ministry of Energy (1323/93-0020, 1753/95-0001, and 1323/98-0025) who supported a number of projects financially and to the external partners with whom I have worked on these projects.

Papers included in the thesis

[A] Henrik Aalborg Nielsen, Torben Skov Nielsen, and Henrik Madsen. Conditional parametric ARX-models. *Journal of Time Series Analysis*, 1998. Submitted.

[B] Henrik Aalborg Nielsen. LFLM version 1.0, an S-PLUS / R library for locally weighted fitting of linear models. Technical Report 22, Department of Mathematical Modelling, Technical University of Denmark, Lyngby, Denmark, 1997.

[C] Henrik Aalborg Nielsen, Torben Skov Nielsen, Alfred Karsten Joensen, Henrik Madsen, and Jan Holst. Tracking time-varying coefficient-functions. *Int. J. of Adaptive Control and Signal Processing*, 1999. Preliminary accepted for publication.

[D] Alfred Karsten Joensen, Gregor Giebel, Lars Landberg, Henrik Madsen, and Henrik Aalborg Nielsen. Model output statistics applied to wind power prediction. In *Wind Energy for the Next Millenium*, European Wind Energy Conference, pages 1177–1180, Nice, France, March 1999.

[E] Alfred Karsten Joensen, Henrik Madsen, Henrik Aalborg Nielsen and Torben Skov Nielsen. Tracking time-varying parameters with local regression. *Automatica*, 1999. To appear.

[F] Payman Sadegh, Henrik Aalborg Nielsen, and Henrik Madsen. A semi-parametric approach for decomposition of absorption spectra in the presence of unknown components. Technical Report 17, Department of Mathematical Modelling, Technical University of Denmark, Lyngby, Denmark, 1999.

[G] Henrik Aalborg Nielsen and Henrik Madsen. A generalization of some classical time series tools. *Computational Statistics and Data Analysis*, 1999. Submitted.

[H] Henrik Aalborg Nielsen and Henrik Madsen. Wind power prediction using ARX models and neural networks. In M. H. Hamza, editor, *Proceedings of the Fifteenth IASTED International Conference on Modelling, Identification and Control*, pages 310–313, February 1996.

[I] Lars Henrik Hansen, Judith L. Jacobsen, Henrik Aalborg Nielsen, and Torben Skov Nielsen. Approximating building components using stochastic differential equations. In J. J. Bloem, editor, *System Identification Competition*, pages 149–162. Joint Research Centre, European Commission, 1996. EUR 16359 EN.

[J] Jakob Bak, Henrik Aalborg Nielsen, and Henrik Madsen. Goodness of fit of stochastic differential equations. In Peter Linde & Anders Holm, editors, *21st Symposium of applied statistics*, Copenhagen Business School, Copenhagen, January 1999.

Summary

The present thesis consists of ten research papers published in the period 1996-1999 together with a summary report. The thesis deals with different aspects of mathematical modelling of systems using data and, if possible, partial knowledge about the systems.

In the first part of the thesis the focus is on combinations of parametric and non-parametric methods of regression. This combination can be in terms of additive models where e.g. one or more non-parametric term is added to a linear regression model. It can also be in terms of conditional parametric models where the coefficients of a linear model are estimated as functions of some explanatory variable(s). Also, software for handling the estimation is presented. The software runs under S-PLUS and R and contains also a number of tools useful when doing model diagnostics or interpreting the results.

Adaptive estimation is also considered. It is shown that adaptive estimation in conditional parametric models can be performed by combining the well known methods of local polynomial regression and recursive least squares with exponential forgetting. The approach used for estimation in conditional parametric models also highlights how recursive least squares with exponential forgetting can be generalized and improved by approximating the time-varying parameters with polynomials locally in time.

In one of the papers well known tools for structural identification of linear time series are generalized to the non-linear time series case. For this purpose non-parametric methods together with additive models are suggested. Also, a new approach specifically designed to detect nonlinearities is introduced. Confidence intervals are constructed by use of bootstrapping.

As a link between non-parametric and parametric methods a paper dealing with neural networks is included. In this paper, neural networks are used for predicting the electricity production of a wind farm. The results are compared with results obtained using an adaptively estimated ARX-model. Finally, two papers on stochastic differential equations are included. In the first paper, among other aspects, the properties of a method for parameter estimation in stochastic differential equations is considered within the field of heat dynamics of buildings. In the second paper a lack-of-fit test for stochastic differential equations is presented. The test can be applied to both linear and non-linear stochastic differential equations.

Some applications are presented in the papers. In the summary report references are made to a number of other applications.

Resumé

Nærværende afhandling består af ti artikler publiceret i perioden 1996-1999 samt et sammendrag og en perspektivering heraf. I afhandlingen behandles aspekter af matematisk modellering af systemer vha. data og, såfremt det er muligt, delvis viden om disse systemer.

Den første del af artiklerne fokuserer på kombinationer af parametriske og ikke-parametriske regressionsmetoder. Sådanne kombinationer kan være additive, hvor f.eks. et eller flere ikke-parametriske led adderes til en lineær regressionsmodel. En anden mulighed er betinget parametriske modeller, hvor koefficienterne i en lineær model estimeres som funktioner af en eller flere forklarende variable. Endvidere præsenteres et EDB-program til håndtering af og estimation i sådanne modeller. Programmet er en udvidelse til S-PLUS og R. Programmet inkluderer også en række værktøjer, der er nyttige i forbindelse med diagnostik og fortolkning af resultater.

Endvidere behandles adaptiv estimation. Det vises, at der ved at kombinere adaptiv estimation i lineære modeller med lokal polynomiel regression, som begge er velkendte metoder, fås en metode, der kan håndtere adaptiv estimation i betinget parametriske modeller. Den anvendte metode til estimation i betinget parametriske modeller tydeliggør også, hvorledes den rekursive mindste kvadraters metode med eksponentiel glemsel kan generaliseres og forbedres ved at approksimere de tidsvarierende parametre med polynomier, der er lokale i tid.

Ikke-parametriske metoder bruges sammen med additive modeller til at generalisere velkendte metoder til strukturel identifikation af lineære tidsrækker. Således opnås nye metoder, der kan håndtere ikke-lineære tids-

rækker. Endvidere introduceres et nyt værktøj specifikt designet til at detektere ikke-lineariteter. Konfidensintervaller konstrueres vha. bootstrapping.

Som et forbindelsesled mellem ikke-parametriske og parametriske metoder er der inkluderet en artikel vedr. neurale netværk. Her bruges neurale netværk til at forudsige elproduktionen i en vindmøllepark, og der sammenlignes med prædiktioner på basis af en adaptivt estimeret ARX-model. Til slut er to artikler vedr. stokastiske differentialligninger inkluderet. Den første artikel vedrører varmedynamik for bygninger. Her undersøges bl.a. egenskaberne for en metode til estimation af parametrene i stokastiske differentialligninger. I den anden artikel præsenteres et test for lack-of-fit af stokastiske differentialligninger. Metoden kan benyttes i forbindelse med både lineære og ikke-lineære systemer.

I den første del af afhandlingen refereres der til en række anvendelser. Desuden præsenteres andre anvendelser i artiklerne.

Contents

Chapter 1

Introduction

This thesis consists of ten research papers. The approach to modelling applied in the papers is to use observations from the system under consideration. However, some aspects of the model will always be based on knowledge about the system obtained from other sources. For some systems this knowledge can be rather specific. For instance the heat dynamics of a wall can be approximated by a set of ordinary differential equations and if noise is present it is appropriate to formulate the problem in terms of stochastic differential equations. Observations from the system under consideration are then used to obtain estimates of the parameters in the model whereby a final model is obtained.

Other systems may not allow such detailed information to be provided. This is typically the case when the system under consideration consists of a number of units for which detailed information is not available or difficult / expensive to obtain. For instance it is clear that the energy needed for heating and hot tap water in a town depends on the outdoor temperature, on the time of day, and on a number of other quantities, see e.g. (Nielsen & Madsen 1999). It may be possible to use relations known from e.g. the theory of heat transfer to obtain more details. However, to a large extend the structure of the model will have to be based on data, partly because deductive methods are rarely applicable for modelling human behaviour. Finally, for many systems the characteristics will change slowly over time. This may be due to wear and tear, extensions, or other reasons. Consequently, adaptive methods of estimation are often necessary, especially in on-line applications.

Experimental design is an important aspect when aiming at obtaining models from data. Although some of the papers included in this thesis address aspects related to experimental design, it is not considered very explicitly. This is mainly because it is often impossible or very expensive to perform experiments. An obvious example is the modelling of the heat consumption in a town. However, in one of the applications mentioned in Chapter 3 experimental design is considered in the context of power consumption, see also (Nielsen & Madsen 1997). When it is not possible to perform an experiment a model derived from theoretical considerations might have to be condensed in some sense in order to retain identifiability.

In practice the model obtained may not be the primary interest, but the model is used to generate predictions of the response, e.g. the heat consumption in a town. These predictions will often be based on uncertain information about the explanatory variables, e.g. the future climate. For a simple setup corresponding to this case Jonsson (1994) has shown that the unbiased estimates should not be used. Furthermore in (Nielsen & Madsen 1999) simulations are used to verify that this is also true for a far more complicated setup. Intuitively, this is quite plausible in that an explanatory variable which is only known with a relatively large uncertainty should not be trusted too much, i.e. in a linear model the corresponding parameter estimate should be shrinked towards zero. The consequence of this is that the appropriateness of the estimation method is highly related to the planed application of the model. This is somewhat contrary to normal statistical practice in which it is traditionally assumed that estimates must be unbiased, or marginally biased if this can reduce the variance of the estimates. An other consequence is that parameters obtained from theoretical considerations (assuming this is possible) are also not always optimal. For control applications these aspects are considered in detail by Gevers (1996), see also (Hjalmarsson, Gevers & De Bruyne 1996).

1.1 Overview of the thesis

A short overview of the papers are presented in Chapter 2 and in Chapter 3 an overview of some actual applications of the methods are presented. Chapter 4 concludes on the thesis.

After this introductory part the papers follow. The first papers, paper A–G, address methods applicable to systems for which detailed information is not available. The methods considered relate to various aspects of non-parametric and semi-parametric methods of regression. In general these methods aim at estimating the dependence of a response on some explanatory variables without specifying a parametric form of this dependence, see e.g. (Hastie & Tibshirani 1990) for an overview of such methods. In this thesis models are considered which allow knowledge about the structure of the system to be build into the model without specifying a complete parametric form.

Then a paper in which neural networks are applied is included. When the output unit is linear, neural networks can approximate any smooth function by increasing the number of neurons (Ripley 1996). In this sense neural networks are not different from non-parametric methods of regression.

Finally, two papers on stochastic differential equations follow. One of these papers is an example of a system where many details are known as outlined above. The other paper presents a new approach for testing the goodness of fit of stochastic differential equations.

1.2 Bibliographic notes

Early work on local polynomial regression includes (Stone 1977, Cleveland 1979, Cleveland 1981), although, as described by Cleveland & Loader (1996), it dates back to the 19'th century. A comprehensive overview of local polynomial regression can be found in (Cleveland & Devlin 1988, Cleveland, Devlin & Grosse 1988) and in (Hastie & Loader 1993) an account for the benefit of local polynomial regression over kernel regression is given. Hastie & Tibshirani (1990, Chapters 2 and 3) gives an excellent overview of smoothing techniques in general and also considers selection of smoothing parameters, degrees of freedom, Bayesian approaches, and more. More details on smoothing splines can be found in (Eubank 1988, Wahba 1990). Regression splines (Wold 1974) is a simple alternative method which can be applied simply by constructing a basis and use this in a linear regression model. However, Hastie & Tibshirani (1990, pp. 251-254) argues that regression splines may produce misleading results. Pseudosplines (Hastie 1996) can be seen as a bridge between regression splines and smoothing splines. A forthcoming book by Ruppert, Wand, and Carroll describes this subject. Wavelet shrinkage (Donoho & Johnstone 1995) is a smoothing technique very suitable for situations where the smoothness of the underlying function changes over the range in which it is to be estimated. For local polynomial regression alternative approaches exists, see (Cleveland & Loader 1996, Section 9) and (Loader 1999). Härdle, Lütkepohl & Chen (1997) gives a review of non-parametric methods in time series analysis.

For guidance on selection of smoothing parameters, cross-validation and related methods have traditionally been used, see (Hastie & Tibshirani 1990). Often leave-one-out cross-validation is used, but there is some evidence that this is not optimal (Breiman & Spector 1992, Shao 1993) and instead K-fold cross-validation could be used. Also for selection of smoothing parameters the plug-in approach (Hall, Sheather, Jones & Marron 1991) is sometimes used, see also (Ruppert, Sheather & Wand 1995) and (Fan & Gijbels 1996). The applicability of the approach is strongly disagreed with by some people, mainly because the approach aim at estimating the bias and use this information to select the bandwidth. Cleveland & Loader (1996, Section 10.3) argues that information on bias should go directly into the estimate of the unknown function, see also (Loader 1995, Section 6). Specificly for time series Hart (1996) considers the subject of selection of smoothing parameters.

Generalized additive models were introduced by Hastie & Tibshirani (1986) and an overview can be found in (Hastie & Tibshirani 1990). The Seasonal Trend Loess (STL) procedure for decomposing time series (Cleveland, Cleveland, McRae & Terpenning 1990) is actually an application of the additive model. Along these lines Hastie & Tibshirani (1993) introduced the varying-coefficients model.

There is a vast literature on neural networks and on stochastic differential equations. This will not be considered in detail. A statistical oriented book on neural networks is (Ripley 1996). Venables & Ripley (1997) can be consulted for a brief introduction, see also (Ripley 1995). Numerical methods for solution of stochastic differential equations are described by Kloeden & Platen (1992). The approach used for estimation of embedded parameters in stochastic differential equations in this thesis is described in (Madsen & Melgaard 1991, Melgaard & Madsen 1993, Melgaard 1994).

Chapter 2

Overview of included papers

Paper A deals with a class of models which we call *conditional parametric models*. These models are linear regression models in which the coefficients are replaced with smooth functions of one or more additional explanatory variables. When these additional variables are constant the model reduces to a linear regression model, hereof the name. Another obvious name for these models often used is *varying-coefficients models*, but Hastie & Tibshirani (1993) use this term for models in which the arguments of the coefficient-functions are not necessarily the same. Estimation of the coefficient-functions can be accomplished by a method very similar to local polynomial regression (Cleveland & Devlin 1988). For time series settings lagged values of the response are easily included. In this case the model is called a conditional parametric ARX-model. The method is applied to model the relation between the conditions at the plant (supply temperature and flow) in a district heating system and the temperature at a specific location in the network.

The function `loess` in S-PLUS (Statistical Sciences 1995) can be used for estimation in some simple conditional parametric models. Paper B describes a software implementation for estimation in conditional parametric models. The software contains no constraints on the size of the problem. The software runs under S-PLUS (Becker, Chambers &

Wilks 1988, Statistical Sciences 1993) and R (Ihaka & Gentleman 1996) and also methods for e.g. prediction and graphics are supplied. The paper also include some details about degrees of freedom and about non-constant error variance (heteroscedasticity).

Adaptive and recursive estimation is considered in the Papers C, D and E. In Paper C a novel method for adaptive estimation in conditional parametric models is developed. The method combines the weighted least squares estimation method suggested for conditional parametric models with the well known method of recursive least squares with exponential forgetting (Ljung & Söderström 1983) used for adaptive estimation in linear models. The method is formulated recursively and simulations are used to illustrate the behaviour of the method. In paper D the method is applied for wind power prediction and compared to other methods, which are outperformed in this case. Paper E deals with adaptive and recursive estimation in linear ARX-models. A method often applied in this context is recursive least squares with exponential forgetting. This method can be viewed as a conditional parametric model (varying-coefficients model in the paper) estimated by locally in time approximating the coefficients by constants. This observation leads to a generalization in which the coefficients are approximated by polynomials locally in time. By use of exponential forgetting the method can be formulated recursively. When the underlying parameters change smoothly over time, examples have shown the method to be superior to the existing recursive estimation method with exponential forgetting.

For some applications it is desirable to include a non-parametric term in a linear regression model. One such application is the decomposition of absorption spectra where the concentrations of certain quantities traditionally are estimated by projecting an observed mixture spectrum on to the linear space generated by a number of reference spectra. However, when the mixture contains additional quantities the estimated concentrations may be biased. The bias can be reduced by simultaneously estimating the spectrum corresponding to the additional quantities. This application is described in Paper F.

New tools for structural identification of non-linear time series models are presented in Paper G. The foundation of the tools is the observation that squared estimates of autocorrelation and partial autocorrelation can be obtained by calculating coefficients of determination (R-squared) from residual sums of squares obtained from fits of linear models. The flexibility of the non-parametric methods and the fact that values of R-squared can be obtained on the basis of other than linear models are then used to generalize the sample autocorrelation function and the sample partial autocorrelation function. Another tool specificly designed to detect departures from linearity is also introduced. In the paper the linear models underlying the traditional tools are replaced with local polynomial regression models (Cleveland & Devlin 1988) and additive models (Hastie & Tibshirani 1990, Chapter 4), but other models could serve as a basis. Confidence intervals are constructed by the use of bootstrapping (Efron & Tibshirani 1993). The methods are illustrated on both simulated and real data. It is demonstrated that they can be used to detect lag dependences and departures from linearity which can not be detected using the sample autocorrelation function or the sample partial autocorrelation function. In principle the generalizations are not restricted to the time series settings.

The remaining papers deal with various aspects within dynamic systems modelling. Paper H compares the use of neural networks and ARX-models for the prediction of the wind power production in a wind farm placed near the west coast of Denmark.

Hereafter papers on stochastic differential equations are included. Paper I presents an application related to the modelling of the heat dynamics of buildings. In the paper quite explicit knowledge about the structure of the system is assumed and the estimates obtained have a physical interpretation. This approach is often called grey box modelling, see e.g. (Bohlin & Graebe 1995). Building components are of the distributed parameter type, i.e. they should in principle be modelled by partial differential equations. In the paper approximations using ordinary (but stochastic) differential equations are used.

Goodness of fit of stochastic differential equations are considered in Paper J. The approach can be applied to a wide range of models, provided that an adequate method of simulation exists. This includes non-linear stochastic differential equations where both the drift and the diffusion are state dependent. In the paper the method is formulated for one-dimensional stochastic differential equation but it can be used in one dimension at a time for multivariate models. The method may be seen as an application of the bootstrap technique (Efron & Tibshirani 1993) in that it is based on a predetermined number of simulations of inter-observational trajectories. For univariate models, and when the model is adequate, the rank of the actual observation as compared to the end-points of the simulated trajectories will follow a multinomial distribution with equal cell frequencies. From this observation a standard χ^2-test is derived.

Chapter 3

Applications

Conditional parametric models as described in Papers A and B and semi-parametric models as used in Paper F are used extensively in (Nielsen & Madsen 1999), where prediction models and methods for the heat consumption in a large district heating network are developed. In the report cross-validation is used for guidance when selecting smoothing parameters. The methods are developed for on-line applications, and it is assumed that meteorological forecasts will be available on-line. In the report simple stationary relations known from the theory of heat transfer (Incropera & DeWitt 1985) are used to arrive at an initial model structure. A feature of this initial model is that the heat convection coefficient on the outside of the walls of the building is expected to vary with the wind speed. However, because the system consists of numerous buildings, more details are difficult to obtain. Instead conditional parametric models are used, whereby the wind speed is allowed to control some coefficients which depend on the heat convection coefficient. Fortunately, it turns out that the coefficient-functions are well approximated by straight lines whereby the actual prediction methods can be build using well known methods such as recursive least squares with exponential forgetting (Ljung & Söderström 1983, Ljung 1987). Periodic B-spline bases are used as an alternative to Fourier expansions to model the diurnal variation of the heat consumption. Conditional parametric models and also adaptive estimation in such models (cf. Paper C) are used in (Nielsen, Madsen & Nielsen 1999) for the development of on-line wind power predictors based on meteorological forecasts.

Models of the electricity consumption in the Eastern part of Denmark are considered in (Nielsen & Madsen 1997). Here conditional parametric models are used to explore a possibly non-linear relation, which in turn is used to build non-linear regression models. A similar model is considered in (Nielsen, Andersen & Madsen 1998), where the annual variation is modelled using B-spline bases. The model is evaluated using cross-validation. The model is implemented as a set of SAS macros and used at ELKRAFT A.m.b.A., Ballerup, Denmark. The SAS macros are documented in (Nielsen & Madsen 1998).

Non-parametric and related methods are also used extensively in (Nielsen & Madsen 1997) where the effect of a power conservation campaign is addressed. Originally, two trial substations with corresponding control substations were selected by use of cluster analysis as described in (Nielsen 1996); an English translation is included in Appendix D of (Nielsen & Madsen 1997). The number of substations selected were based on economic constraints. Extensive information on ways to reduce power consumption were then offered to each household of the trial substations (Sørensen 1997). Hereafter hourly measurements of power consumption for the four substations obtained before, during, and after the actual trial were analyzed. The methods used include a modification of the time series decomposition method Seasonal Trend Loess (STL) (Cleveland et al. 1990), which allows the diurnal variation to vary between different types of days. As noted by (Hastie & Tibshirani 1990, Section 8.5) STL is actually a special kind of an additive model estimated using backfitting. In (Nielsen & Madsen 1997) it is also shown that the four series in the full trial can be decomposed simultaneously into main an interaction effects of the trend and seasonal components as in analysis of variance. This is achieved by combining conditional parametric models and additive models. Furthermore information criteria are used for the selection of bandwidths, and it is shown that the residuals should be corrected for temporal correlation before applying the information criteria. An inappropriate behaviour of the Bayes Information Criterion (BIC) is observed and in the discussion in (Nielsen & Madsen 1997) it is argued that one of the basic assumptions used by (Schwarz 1978) in the derivation of BIC is violated for smoothing parameter selection.

Chapter 4

Conclusion and Discussion

The main part of the papers included in this thesis are concerned with various aspects of non-parametric and semi-parametric methods. One paper presents an application of neural networks, which as mentioned in the introduction, can be seen as a non-parametric method. Finally, two papers on stochastic differential equations are included.

Stochastic differential equations and traditional non-parametric methods can be seen as two quite contradictory approaches to modelling. When stochastic differential equations are used rather detailed knowledge about the system is required. As opposed to this non-parametric methods “let the data speak”. However, when there are many explanatory variables prior knowledge is required to arrive at a sensible model, also in the non-parametric case. In this sense conditional parametric models (cf. Paper A) provides a simple link between linear parametric models and non-parametric models. The approach is attractive since it shares many properties with local polynomial regression (cf. Paper B) and furthermore it is based on weighted least squares for which adequate numerical solutions exist, see e.g. (Miller 1992) and Paper B. To make the approach readily applicable software for estimation in such models and for handling of estimation results has been developed (cf. Paper B). To enhance flexibility the software is embedded in S-PLUS and R. As shown in Paper C the close relation to the least squares solution for linear models allows for adaptive and recursive estimation, see also Paper D.

Furthermore, the close relation with least squares, allows for a generalization and improvement of existing methods for adaptive and recursive estimation in linear models (cf. Paper E).

For local regression Cleveland & Loader (1996, Section 10.1) gives a simple and convincing account for the bias and variance of the estimated regression function. For the variance of the estimated functions in conditional parametric models the approximative results of Hastie & Tibshirani (1993) can be used. Bootstrapping might be more appropriate since it rely on fewer approximations / assumptions and it is applicable even when autoregressive terms are included in the model. For conditional parametric models an understanding of the sources of bias on the estimated functions will be more demanding than for local regression and further research is needed. One analysis should be based on the assumption that the structure of the true system is identical to that of the conditional parametric model. Hereafter the effect of approximating the functions locally by low-order polynomials can be investigated using Taylor series as it is done by Cleveland & Loader (1996). If instead the expected response of the true system is a completely general function of the explanatory variables it might be possible, by comparing the two Taylor series, to obtain further understanding of the origin of bias in this case. Also, the effect on bias of including autoregressive terms should be investigated. For conditional parametric models with autoregressive terms it is likely that bias can be introduced if the model error is not white noise. However, compared to the case of linear autoregressive models the principle called "whitening by window" (Hart 1996) might actually reduce bias.

Additive models provide an alternative approach by which structure can be build into the model. These models are rather flexible in that they can be used additively to combine linear models with smoothers. The linear models can also be combined with conditional parametric models since the latter share properties with local polynomial regression. Estimation in additive models is normally performed using backfitting (Hastie & Tibshirani 1990), but recently Fan, Härdle & Mammen (1998) has suggested an alternative approach. Backfitting is an iterative procedure, but when there is only one smoother involved a closed form solution exist (Hastie & Tibshirani 1990, p. 118). This solution is used in Paper F for decomposition of absorption spectra. In the paper it is shown why the traditional method for decomposition of absorption spectra may produce biased results in case of unknown components. The paper also presents a novel approach which amounts to averaging the weighted least squares criterion of conditional parametric models over the observations and requiring the estimates in the linear part of the model to be constant across observations. For the example presented in Paper F the novel approach perform better than the closed form solution to the backfitting iterations. Further research is needed to clarify under which circumstances this holds.

Smoothers and additive models are used together with bootstrapping (Efron & Tibshirani 1993) in Paper G to generalize well known tools for structural identification of time series. Hereby structural identification of non-linear time series is allowed for. It is attractive that results obtained by the new tools can be presented in a way similar to results of the classical tools. Furthermore, when the models used in the new tools are replaced by linear models, the new tools reduce to the classical tools. In Paper G additive models are fitted using backfitting. A closed-form solution can be provided if the additive models are approximated using pseudosplines (Hastie 1996) and this may also simplify the calculation of confidence intervals. Further research is needed to clarify these aspects.

Application of the non- and semi-parametric methods used in this thesis is relatively simple. However, one or more smoothing parameters must be selected. In some of the applications mentioned in Chapter 3 cross-validation or information criteria are used for guidance in selecting the smoothing parameter. It must be noted that based on these criteria the optimal smoothing parameter is not well defined and therefore in practice some judgment is called upon, see also (Hastie & Tibshirani 1990, Section 3.4). Another approach is plug-in methods, see e.g. (Fan & Gijbels 1996). These methods are not considered in this thesis, but in Section 1.2 some references to the literature are mentioned. As it is evident from eg. (Breiman & Spector 1992, Shao 1993) model selection and hence also selection of a smoothing parameter is a difficult theoretical and practical problem. When using local polynomial approximations the selection of a bandwidth can also be based on prior knowledge about the smoothness of the functions to be estimated. This information can in some cases be obtained from plots of the data. For this purpose first and second order polynomials seems to be most useful in that questions like "for how large a bandwidth would a first / second order polynomial fit the data?" is relatively easy to answer. Such questions are difficult to answer for kernel smoothers (local constants), for smoothing splines, and also for higher order polynomials.

Contrary to conditional parametric models, estimation in neural networks (cf. Paper H) seems to be complicated due to the inherent nonlinear structure. For this reason it is often appropriate to try a number of different starting values. Similar problems can arise when estimating parameters in stochastic differential equations (cf. Paper I). However, if the structure of the model considered is based on prior physical knowledge, as e.g. the theory of heat transfer, then often appropriate starting values can be obtained by considering the actual system. Therefore often, in practice, the application of stochastic differential equations is not as problematic as it might seem from a theoretical point of view.

Bibliography

Becker, R. A., Chambers, J. M. & Wilks, A. R. (1988), *The New S Language: A Programming Environment for Data Analysis and Graphics*, Wadsworth, Belmont, CA.

Bohlin, T. & Graebe, S. F. (1995), 'Issues in nonlinear stochastic grey box identification', *International Journal of Adaptive Control and Signal Processing* **9**, 465–490.

Breiman, L. & Spector, P. (1992), 'Submodel selection and evaluation in regression. The X-random case', *International Statistical Review* **60**, 291–319.

Cleveland, W. S. (1979), 'Robust locally weighted regression and smoothing scatterplots', *Journal of the American Statistical Association* **74**, 829–836.

Cleveland, R. B., Cleveland, W. S., McRae, J. E. & Terpenning, I. (1990), 'STL: A seasonal-trend decomposition procedure based on loess', *Journal of Official Statistics* **6**, 3–33. (C/R: pp. 33-73; Corr. to Comment: pp. 343-348).

Cleveland, W. S. (1981), 'LOWESS: A program for smoothing scatterplots by robust locally weighted regression', *The American Statistician* **35**, 54.

Cleveland, W. S. & Devlin, S. J. (1988), 'Locally weighted regression: An approach to regression analysis by local fitting', *Journal of the American Statistical Association* **83**, 596–610.

Cleveland, W. S., Devlin, S. J. & Grosse, E. (1988), 'Regression by local fitting: Methods, properties, and computational algorithms', *Journal of Econometrics* **37**, 87–114.

Cleveland, W. S. & Loader, C. (1996), Smoothing by local regression: Principles and methods, *in* W. Härdle & M. G. Schimek, eds, 'Statistical Theory and Computational Aspects of Smoothing. Proceedings of the COMPSTAT '94 Satellite Meeting', Physica-Verlag, Heidelberg, pp. 10–49. (Discussion: pp. 80-127).

Donoho, D. L. & Johnstone, I. M. (1995), 'Adapting to unknown smoothness via wavelet shrinkage', *Journal of the American Statistical Association* **90**, 1200–1224.

Efron, B. & Tibshirani, R. J. (1993), *An Introduction to the Bootstrap*, Chapman & Hall, London/New York.

Eubank, R. L. (1988), *Spline Smoothing and Nonparametric Regression*, Marcel Dekker, New York.

Fan, J. & Gijbels, I. (1996), *Local Polynomial Modelling and Its Applications*, Chapman & Hall, London/New York.

Fan, J., Härdle, W. & Mammen, E. (1998), 'Direct estimation of low-dimensional components in additive models', *The Annals of Statistics* **26**(3), 943–971.

Gevers, M. (1996), 'Identification for control', *Annual Reviews in Control* **20**, 95–196.

Hall, P., Sheather, S. J., Jones, M. C. & Marron, J. S. (1991), 'On optimal data-based bandwidth selection in kernel density estimation', *Biometrika* **78**, 263–269.

Härdle, W., Lütkepohl, H. & Chen, R. (1997), 'A review of nonparametric time series analysis', *International Statistical Review* **65**, 49–72.

Hart, J. D. (1996), 'Some automated methods of smoothing time-dependent data', *Journal of Nonparametric Statistics* **6**, 115–142.

Hastie, T. (1996), 'Pseudosplines', *Journal of the Royal Statistical Society, Series B, Methodological* **58**, 379–396.

Hastie, T. J. & Tibshirani, R. J. (1990), *Generalized Additive Models*, Chapman & Hall, London/New York.

Hastie, T. & Loader, C. (1993), 'Local regression: Automatic kernel carpentry', *Statistical Science* **8**, 120–129. (Discussion: pp. 129-143).

Hastie, T. & Tibshirani, R. (1986), 'Generalized additive models', *Statistical Science* **1**, 297–310. (C/R: pp. 310-318).

Hastie, T. & Tibshirani, R. (1993), 'Varying-coefficient models', *Journal of the Royal Statistical Society, Series B, Methodological* **55**, 757–796.

Hjalmarsson, H., Gevers, M. & De Bruyne, F. (1996), 'For model-based control design, closed-loop identification gives better performance', *Automatica* **32**(12), 1659–1673.

Ihaka, R. & Gentleman, R. (1996), 'R: A language for data analysis and graphics', *Journal of Computational and Graphical Statistics* **5**, 299–314.

Incropera, F. P. & DeWitt, D. P. (1985), *Fundamentals of Heat and Mass Transfer*, Second edn, John Wiley & Sons.

Jonsson, B. (1994), 'Prediction with a linear regression model and errors in a regressor', *International Journal of Forecasting* **10**, 549–555.

Kloeden, P. E. & Platen, E. (1992), *Numerical Solution of Stochastic Differential Equations*, Springer-Verlag, Berlin/New York.

Ljung, L. (1987), *System Identification: Theory for the User*, Prentice-Hall, Englewood Cliffs, NJ.

Ljung, L. & Söderström, T. (1983), *Theory and Practice of Recursive Identification*, MIT Press, Cambridge, MA.

Loader, C. (1995), Old faithful erupts: Bandwidth selection reviewed, Technical report, AT&T Bell Laboratories. http://cm.bell-labs.com/cm/ms/departments/sia/doc/95.9.ps.

Loader, C. (1999), *Local regression and likelihood*, Springer, New York.

Madsen, H. & Melgaard, H. (1991), The mathematical and numerical methods used in CTLSM – a program for ML-estimation in stochastic, continuous time dynamical models, Technical Report 7, Institute of Mathematical Statistics and Operations Research, Technical University of Denmark, Lyngby, Denmark.

Melgaard, H. (1994), Identification of physical models, PhD thesis, Department of Mathematical Modelling, Technical University of Denmark, Lyngby, Denmark.

Melgaard, H. & Madsen, H. (1993), CTLSM continuous time linear stochastic modelling, Technical Report 1, Institute of Mathematical Statistics and Operations Research, Technical University of Denmark, Lyngby, Denmark.

Miller, A. J. (1992), '[Algorithm AS 274] Least squares routines to supplement those of Gentleman', *Applied Statistics* **41**, 458–478. (Correction: 94V43 pp. 678).

Nielsen, H. A. & Madsen, H. (1997), *Development of methods for evaluation of electricity savings and load levelling Measures, Part 3: Experimental Assessment of the Effect of a Power Conservation Campaign*, Department of Mathematical Modelling, Technical University of Denmark, Lyngby, Denmark in collaboration with NESA A/S, Hellerup, Denmark. EFP95/1753/95-0001.

Nielsen, H. A. (1996), 'Note on grouping substations prior to execution of a trial (in Danish: Notat om gruppering af transformerudføringer med henblik på udførelse af forsøg)', Department of Mathematical Modelling, Technical University of Denmark, Lyngby, Denmark.

Nielsen, H. A., Andersen, K. K. & Madsen, H. (1998), Empirically determined model of the power consumption in East Denmark (in Danish: Empirisk bestemt model for elforbruget i Østdanmark), Technical Report 18, Department of Mathematical Modelling, Technical University of Denmark, Lyngby, Denmark.

Nielsen, H. A. & Madsen, H. (1997), *Development of methods for evaluation of electricity savings and load levelling Measures, Part 1: Aggregated Power Consumption Models for the Eastern Part of Denmark*, Department of Mathematical Modelling, Technical University of Denmark, Lyngby, Denmark in collaboration with NESA A/S, Hellerup, Denmark. EFP95/1753/95-0001.

Nielsen, H. A. & Madsen, H. (1998), SAS-macros for estimation and prediction in a model of the power consumption (in Danish: SAS-makroer til estimation og prædiktion i elforbrugsmodel), Technical Report 19, Department of Mathematical Modelling, Technical University of Denmark, Lyngby, Denmark.

Nielsen, H. A. & Madsen, H. (1999), *Forecasting the Heat Consumption in District Heating Systems using Meteorological Forecasts*, Department of Mathematical Modelling, Technical University of Denmark, Lyngby, Denmark. EFP98/1323/98-0025, draft.

Nielsen, T. S., Madsen, H. & Nielsen, H. A. (1999), *Using Meteorological Forecasts in On-line Predictions of Wind Power*, Department of Mathematical Modelling, Technical University of Denmark, Lyngby, Denmark.

Ripley, B. D. (1995), Statistical ideas for selecting network architectures, *in* B. Kappen & S. Gielen, eds, 'Neural Networks: Artificial Intelligence and Industrial Applications', Springer, pp. 183–190. http://www.stats.ox.ac.uk/pub/neural/papers/SNN.ps.Z.

Ripley, B. D. (1996), *Pattern Recognition and Neural Networks*, Cambridge University Press, Cambridge, UK.

Ruppert, D., Sheather, S. J. & Wand, M. P. (1995), 'An effective bandwidth selector for local least squares regression', *Journal of the American Statistical Association* **90**, 1257–1270. (Correction: 96V91 pp. 1380).

Schwarz, G. (1978), 'Estimating the dimension of a model', *The Annals of Statistics* **6**, 461–464.

Shao, J. (1993), 'Linear model selection by cross-validation', *Journal of the American Statistical Association* **88**, 486–494.

Sørensen, M. S. (1997), 'Development of methods for evaluation of electricity saving and load levelling measures, part 2: The planning and implementation of a power conservation campaign', NESA A/S, Hellerup, Denmark.

Statistical Sciences (1993), *S-PLUS User's Manual, Version 3.2*, StatSci, a division of MathSoft, Inc., Seattle.

Statistical Sciences (1995), *S-PLUS Guide to Statistical & Mathematical Analysis, Version 3.3*, StatSci, a division of MathSoft, Inc., Seattle.

Stone, C. J. (1977), 'Consistent nonparametric regression', *The Annals of Statistics* **5**, 595–620. (C/R: pp. 620-645).

Venables, W. N. & Ripley, B. D. (1997), *Modern Applied Statistics With S-Plus*, Second edn, Springer-Verlag, Berlin/New York.

Wahba, G. (1990), *Spline Models for Observational Data*, Society for Industrial and Applied Mathematics, Philadelphia.

Wold, S. (1974), 'Spline functions in data analysis', *Technometrics* **16**, 1–11.

Papers

A

Paper A

Conditional parametric ARX-models

Submitted to *Journal of Time Series Analysis*. A previous version is published as

> Henrik Aalborg Nielsen, Torben Skov Nielsen, and Henrik Madsen. ARX-models with parameter variations estimated by local fitting. In Yoshikazu Sawaragi and Setsuo Sagara, editors, *11th IFAC Symposium on System Identification*, volume 2, pages 475–480, 1997.

Conditional parametric ARX-models

Henrik Aalborg Nielsen[1], Torben Skov Nielsen[1], and Henrik Madsen[1]

Abstract

In this paper conditional parametric ARX-models are suggested and studied by simulation. These non-linear models are traditional ARX-models in which the parameters are replaced by smooth functions. The estimation method is based on the ideas of locally weighted regression. It is demonstrated that kernel estimates (local constants) are in general inferior to local quadratic estimates. For the considered application, modelling of temperatures in a district heating system, the input sequences are correlated. Simulations indicate that correlation to this extend results in unreliable kernel estimates, whereas the local quadratic estimates are quite reliable.

Keywords: Non-linear models; non-parametric methods; kernel estimates; local polynomial regression; ARX-models; time series.

1 Introduction

Linear models in which the parameters are replaced by smooth functions are denoted varying-coefficients models. Estimation in these models has been developed for the regression framework, see e.g. (Hastie & Tibshirani 1993). In this paper a special class of the models in which all coefficients are controlled by the same argument are considered, these are also denoted conditional parametric models, see e.g. (Anderson, Fang & Olkin 1994). This class of models is applied for autoregressive processes with external input and the resulting models will be denoted conditional parametric ARX-models. These models are similar to smooth threshold autoregressive models, see e.g. (Tong 1990), but more general since a transition is related to each coefficient and a non-parametric form is assumed for these transitions. The method of estimation is closely related

[1]Department of Mathematical Modelling, Technical University of Denmark, DK-2800 Lyngby, Denmark

to locally weighted regression (Stone 1977, Cleveland 1979, Cleveland & Devlin 1988, Cleveland, Devlin & Grosse 1988). Because of the autoregressive property the fitted values will not be linear combinations of the observations and results concerning bias and variance obtained for the regression framework (Cleveland & Devlin 1988, Hastie & Loader 1993) can not be used. For this reason the method is studied by simulation.

Our interest in these models originates from the modelling of temperatures in a district heating system. In such a system the energy needed for heating and hot tap-water in the individual households is supplied from a central heating utility. The energy is distributed as hot water through a system of pipelines covering the area supplied. In the system an increased energy demand is first meet by increasing the flow rate in the system and, when the maximum flow rate is reached, by increasing the supply temperature at the utility. The energy demand in a district heating system typically exhibits a strong diurnal variation with the peak load occuring during the morning hours. A similar pattern can be found in the observed flow rates, although this is also influenced by variations in the supply temperature. Consequently, the time delay for an increase in the supply temperature to be observed in a household inlet also has a diurnal variation.

Models of the relationship between supply temperature and inlet temperature are of high interest from a control point of view. Previous studies have lead to a library of ARX-models with different time delays and with a diurnal variation in the model parameters. Methods for on-line estimating of the varying time delay as well as a controller which takes full advantage of this model structure have previously been published (Søgaard & Madsen 1991, Palsson, Madsen & Søgard 1994, Madsen, Nielsen & Søgaard 1996). A more direct approach is to use one ARX-model but with parameters replaced by smooth functions of the flow rate. This approach is addressed in this paper. This further has the advantage, that the need for on-line estimation of the time delay is eliminated.

In Section 2 the conditional parametric model and the estimation methods are outlined. The performance of the estimators are studied by simulation in Section 3. An application to real data from a district heating system is described in Section 4. Some of this material has previously been presented at a conference, see (Nielsen, Nielsen & Madsen 1997).

2 Model and estimation

A conditional parametric model is a linear regression model with the parameters replaced by smooth functions. The name of the model comes from observing that if the argument of the functions are fixed then the model is an ordinary linear model.

Models of this type are briefly described in the literature (Hastie & Tibshirani 1993, Anderson et al. 1994). Below a more general description of some aspects of the method is presented. The method of estimation is closely related to locally weighted regression (Stone 1977, Cleveland & Devlin 1988, Cleveland et al. 1988, Hastie & Loader 1993).

2.1 The regression case

Assume that observations of the response y_t and the explanatory variables $\boldsymbol{x}_t$ and $\boldsymbol{z}_t$ exist for observation numbers $t = 1, \ldots, n$. An intercept is included in the model by putting the first element of $\boldsymbol{z}_t$ equal to one. The conditional parametric model for this setup is

$$y_t = \boldsymbol{z}_t^T \boldsymbol{\theta}(\boldsymbol{x}_t) + e_t \quad (t = 1, \ldots, n), \tag{1}$$

where e_t is i.i.d. $N(0, \sigma^2)$ and $\boldsymbol{\theta}(\cdot)$ is a vector of functions, with values in $\mathcal{R}$, to be estimated. The functions $\boldsymbol{\theta}(\cdot)$ is only estimated for distinct values of their argument $\boldsymbol{x}$. In this paper the approach taken is to estimate $\boldsymbol{\theta}(\cdot)$ at points sufficiently close for linear interpolation. Below $\boldsymbol{x}$ denotes a single such point within the space spanned by the observations $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$.

The estimation of $\boldsymbol{\theta}(\boldsymbol{x})$ is accomplished by calculating the weighted least squares estimate of the parameter vector, i.e. close to $\boldsymbol{x}$ the function $\boldsymbol{\theta}(\cdot)$ is approximated by a constant vector. The weight on observation t is related to the distance from $\boldsymbol{x}$ to $\boldsymbol{x}_t$, such that

$$w_t(\boldsymbol{x}) = W(||\boldsymbol{x}_t - \boldsymbol{x}||/h(\boldsymbol{x})), \tag{2}$$

where $W : \mathcal{R}_0 \to \mathcal{R}_0$ is a nowhere increasing function. In this paper the tricube function

$$W(u) = \begin{cases} (1-u^3)^3, & u \in [0;1) \\ 0, & u \in [1;\infty) \end{cases} \tag{3}$$

is used. $||\boldsymbol{x}_t - \boldsymbol{x}||$ is the Euclidean distance between $\boldsymbol{x}_t$ and $\boldsymbol{x}$. The scalar $h(\boldsymbol{x}) > 0$ is called the bandwidth. If $h(\boldsymbol{x})$ is constant for all values of $\boldsymbol{x}$ it is denoted a fixed bandwidth. If $h(\boldsymbol{x})$ is chosen so that a certain fraction (α) of the observations is within the bandwidth it is denoted a nearest neighbour bandwidth. If $\boldsymbol{x}$ has dimension of two, or larger, scaling of the individual elements of $\boldsymbol{x}_t$ before applying the method should be considered, see e.g. (Cleveland & Devlin 1988). A rotation of the coordinate system, in which $\boldsymbol{x}_t$ is measured, could also be relevant. Note that if $\boldsymbol{z}_t = 1$ for all t the method of estimation reduces to determining the scalar $\hat{\theta}(\boldsymbol{x})$ so that $\sum_{t=1}^{n} w_t(\boldsymbol{x})(y_t - \hat{\theta}(\boldsymbol{x}))^2$ is minimized, i.e. the method reduces to kernel estimation, see also (Härdle 1990, p. 30) or (Hastie & Loader 1993). For this reason the described method of estimation of $\boldsymbol{\theta}(\boldsymbol{x})$ in (1) is called kernel or local constant estimation.

If the bandwidth $h(\boldsymbol{x})$ is sufficiently small the approximation of $\boldsymbol{\theta}(\cdot)$ as a constant vector near $\boldsymbol{x}$ is good. The consequence is, however, that a relatively low number of observations is used to estimate $\boldsymbol{\theta}(\boldsymbol{x})$, resulting in a noisy estimate, or bias if the bandwidth is increased. See also the comments on kernel estimates in (Anderson et al. 1994) or (Hastie & Loader 1993).

It is, however, well known that locally to $\boldsymbol{x}$ the elements of $\boldsymbol{\theta}(\cdot)$ may be approximated by polynomials, and in many cases these will provide better approximations for larger bandwidths than those corresponding to local constants. Local polynomial approximations are easily included in the method described. Let $\theta_j(\cdot)$ be the j'th element of $\boldsymbol{\theta}(\cdot)$ and let $\boldsymbol{p}_d(\boldsymbol{x})$ be a column vector of terms in a d-order polynomial evaluated at $\boldsymbol{x}$. If for instance $\boldsymbol{x} = [x_1 \ x_2]^T$ then $\boldsymbol{p}_2(\boldsymbol{x}) = [1 \ x_1 \ x_2 \ x_1^2 \ x_1x_2 \ x_2^2]^T$. Furthermore, let $\boldsymbol{z}_t = [z_{1t} \ldots z_{pt}]^T$. With

$$\boldsymbol{u}_t^T = \left[z_{1t}\boldsymbol{p}_{d(1)}^T(\boldsymbol{x}_t) \ldots z_{jt}\boldsymbol{p}_{d(j)}^T(\boldsymbol{x}_t) \ldots z_{pt}\boldsymbol{p}_{d(p)}^T(\boldsymbol{x}_t)\right] \tag{4}$$

and

$$\hat{\boldsymbol{\phi}}^T(\boldsymbol{x}) = [\hat{\boldsymbol{\phi}}_1^T(\boldsymbol{x}) \ldots \hat{\boldsymbol{\phi}}_j^T(\boldsymbol{x}) \ldots \hat{\boldsymbol{\phi}}_p^T(\boldsymbol{x})], \tag{5}$$

where $\hat{\boldsymbol{\phi}}_j(\boldsymbol{x})$ is a column vector of local constant estimates at $\boldsymbol{x}$ corresponding to $z_{jt}\boldsymbol{p}_{d(j)}(\boldsymbol{x}_t)$. The estimation is handled as described above, but fitting the linear model

$$y_t = \boldsymbol{u}_t^T \boldsymbol{\phi}_x + e_t \quad (t = 1, \ldots, N), \tag{6}$$

locally to $\boldsymbol{x}$. Hereafter the elements of $\hat{\boldsymbol{\theta}}(\boldsymbol{x})$ are calculated as

$$\hat{\theta}_j(\boldsymbol{x}) = \boldsymbol{p}_{d(j)}^T(\boldsymbol{x})\, \hat{\boldsymbol{\phi}}_j(\boldsymbol{x}) \quad (j = 1, \ldots p). \tag{7}$$

When $\boldsymbol{z}_j = 1$ for all j this method is identical to the method by Cleveland & Devlin (1988), with the exception that Cleveland & Devlin center the elements of $\boldsymbol{x}_i$ used in $\boldsymbol{p}_d(\boldsymbol{x}_t)$ around $\boldsymbol{x}$ and consequently $\boldsymbol{p}_d(\boldsymbol{x}_t)$ must be recalculated for each value of $\boldsymbol{x}$ considered.

Let $\boldsymbol{U}$ be a matrix with rows $\boldsymbol{u}_t^T$, let $\boldsymbol{W}(\boldsymbol{x}_t)$ be a diagonal matrix containing the weights on the observations when $\boldsymbol{x} = \boldsymbol{x}_t$, and let $\boldsymbol{y}$ be a column vector containing the observations. Using this notation the fitted value for observation t can be written $\hat{y}_t = \boldsymbol{u}_t^T[\boldsymbol{U}^T\boldsymbol{W}(\boldsymbol{x}_t)\boldsymbol{U}]^{-1}\boldsymbol{U}^T\boldsymbol{W}(\boldsymbol{x}_t)\boldsymbol{y}$. Since no element of $\boldsymbol{y}$ is used in $\boldsymbol{u}_t^T$ it then follows that the fitted values are linear combinations of the observations. This property is shared with, e.g., locally weighted regression and forms the basis of discussions regarding bias and variance, see e.g. (Cleveland & Devlin 1988, Hastie & Loader 1993).

The weighted least squares problem is solved by the algorithm described in (Miller 1992). The algorithm was originally written in Fortran but here a port to C by A. Shah is used (`pub/C-numanal/as274_fc.tar.z` from `usc.edu`, using anonymous ftp).

2.2 ARX-models

It is well known that a linear ARX-model can be written in the form (1), where t is the time index, $\boldsymbol{\theta}(\cdot)$ is a constant parameter vector and $\boldsymbol{z}_t$ contains input and lagged values of y_t. For this reason the method described in Section 2.1 is easily extended to models of the form:

$$y_t = \sum_{i \in L_y} a_i(x_{t-m}) y_{t-i} + \sum_{i \in L_u} b_i(x_{t-m}) u_{t-i} + e_t, \tag{8}$$

where t is the time index, y_t is the response, x_t and u_t are inputs, $\{e_t\}$ is i.i.d. $N(0, \sigma^2)$, L_y and L_u are sets of positive integers defining the autoregressive and input lags in the model, and m is a positive integer. Finally, $a_i(\cdot)$ and $b_i(\cdot)$ are unknown but smooth functions which are to be estimated. Extensions to multivariate x_t and u_t are strait forward. Models belonging to the class defined by (8), including the multivariate extensions just mentioned, will be denoted conditional parametric ARX-models.

For this class of models the matrix $\boldsymbol{U}$, which were introduced in the previous section, will depend on the observations of the response and hence the fitted values will not be linear combinations of the observations.

3 Simulation study

To study the estimation method described in Section 2.1 for models of the class (8) simulations using the model

$$y_t = a_1(x_{t-1})y_{t-1} + \sum_{i \in \{2,4\}} b_i(x_{t-1})u_{t-i} + e_t, \tag{9}$$

are performed. In the simulations e_t is normally distributed white noise with zero mean and variance σ_e^2. The functions used are $a_1(x) = 0.5 + 0.5/(1 + \mathrm{e}^{4x-0.5})$, $b_2(x) = 0.045\sqrt{\Phi_{(0.09,0.015)}(x)}$, and $b_4(x) = 0.025\sqrt{\Phi_{(0.05,0.01)}(x)}$, where $\Phi_{(\mu,\sigma)}(\cdot)$ is the Gaussian probability distribution function with mean μ and standard deviation σ. Note that for all values of x used in the simulations the values of $b_2(x)$ and $b_4(x)$ are positive, and the pole ($a_1(x)$) is between 0.5 and 1. However, in the simulations the pole takes values between 0.75 and 0.79.

In the district heating case it is not possible to experiment with the system and hence the response corresponding to uncorrelated inputs can not be obtained. To study the performance of the estimation method in a relevant setting measurements of half-hourly averages of supply temperature (u_t) and flow (x_t) from a running district heating plant is used. In order to investigate certain aspects of the method simulations with white noise inputs are also presented. The distributions of the white noise inputs are chosen so that the range corresponds to the range of observations from the district heating plant.

In the simulations local constant (kernel) estimates are compared to local quadratic estimates. Local quadratic estimates are chosen rather than local linear estimates since the first type is expected to have less bias near peaks and to have similar performance in other regions.

In Section 3.1 simulated white noise input with a uniform distribution are used and the performance of local constant (kernel) estimates are compared with local quadratic estimates. Simulated white noise input with a normal distribution are used in Section 3.2 and local quadratic estimates are investigated for cases where data is sparse in some regions. Simulations with correlated input are addressed in Section 3.3 and in this section local constant and local quadratic estimates are compared when a lag not included in the simulation is included in the estimation. Histograms of the sequences of x_t are shown in Figure 1.

The length of the simulated series is approximately 40% of the length of the real data series used in Section 4. To obtain roughly the same variance of the estimates in the simulation and in the application, the simulation variance is chosen to be 40% of the variance obtained in Section 4. With a robust estimate of the variance obtained from the central 95% of the residuals, the simulation variance of e_t in (9) is 0.27^2.

The function values are estimated at 50 equidistantly spaced points. Results for nearest neighbour bandwidths with $\alpha = 0.1, 0.2, \ldots, 0.6$ are presented. In the figures the true values are indicated by a dotted line.

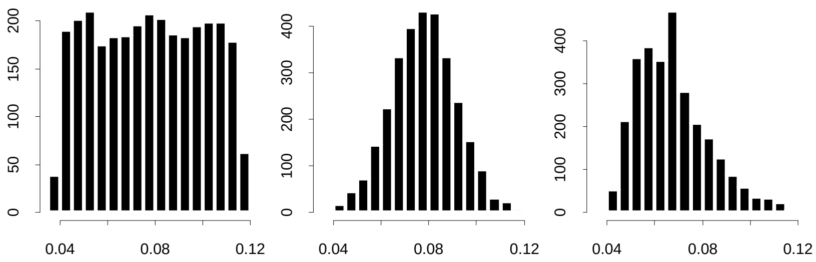


Figure 1: Histograms of x_t used in the simulations. From left to right; uniformly distributed white noise, normally distributed white noise, and measurements of flow (m^3/s) from a running plant.

3.1 Uniformly distributed input

Minimum and maximum observed values of supply temperature and low pass filtered flow (c.f. Section 3.3) are used as limits of the uniform distribution from which the white noise input sequences, of length 3000, are generated, i.e. x_t is i.i.d. $U(0.0390, 0.116)$ and u_t is i.i.d. $U(78.2, 94.2)$.

In Figure 2 local constant (kernel) and local quadratic estimates of the functions $a_1(\cdot)$, $b_2(\cdot)$, and $b_4(\cdot)$ in (9) are shown. It is clearly seen that the local quadratic is superior to the local constant approximation. This is especially true near the border of the interval spanned by the observations and in areas with large curvature. Also note that the local quadratic approximation is quite insensitive to the choice of bandwidth.

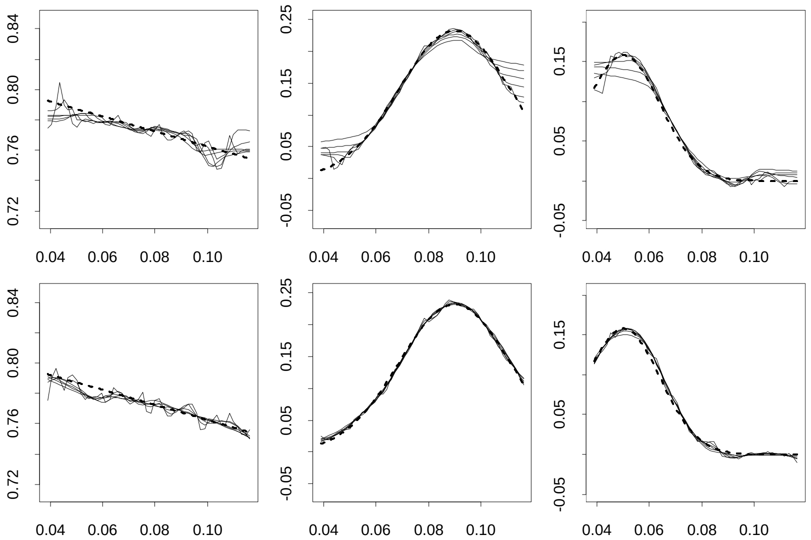


Figure 2: Uniformly distributed white noise input. Local constant (top row) and local quadratic (bottom row). From left to right; $\hat{a}_1(x)$, $\hat{b}_2(x)$, and $\hat{b}_4(x)$.

3.2 Normally distributed input

To investigate the performance of the local quadratic estimates in cases where the data are sparse near the border of the interval spanned by the observations a simulation corresponding to Section 3.1 is performed with normally distributed white noise input.

The mean and variance of the normal distributions are chosen so that the lower and upper limits of the uniform distributions of Section 3.1 is exceeded with probability 0.25%, i.e. x_t is i.i.d. $N(0.0776, 0.0138^2)$ and u_t is i.i.d. $N(86.2, 2.86^2)$, 15 values of each of the simulated input sequences are expected to exceed the limits. Values exceeding the limits are truncated.

In Figure 3 local quadratic estimates of the functions $a_1(\cdot)$, $b_2(\cdot)$, and $b_4(\cdot)$ in (9) are shown. When compared with the distribution of x_t in Figure 1 it is seen that the estimates are relatively far from the true values in regions with sparse data. To further investigate this aspect Figure 4 indicate the pointwise distribution of the estimates, calculated for $\alpha = 0.3$ at 20 equidistantly spaced points.

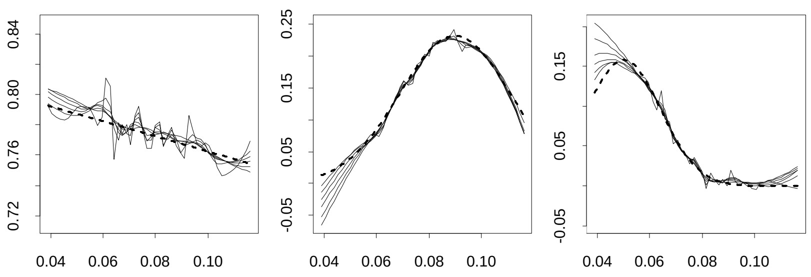


Figure 3: Normally distributed white noise input and local quadratic estimates. From left to right; $\hat{a}_1(x)$, $\hat{b}_2(x)$, and $\hat{b}_4(x)$.

3.3 Measurements from a running plant as input

In order to address the influence of correlated input sequences on the performance of the estimation method, measurements of 2873 half-hourly averages of flow (x_t, m^3/s) and supply temperature (u_t, °C) from a

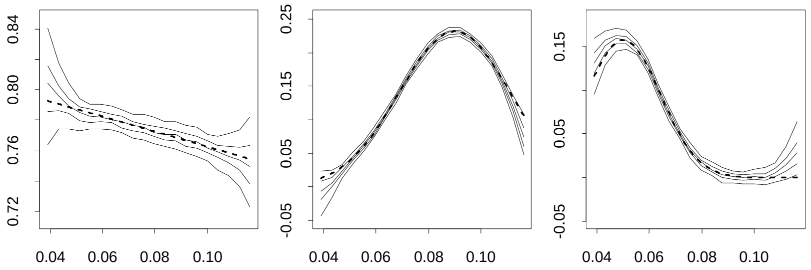


Figure 4: Quantiles (5%, 25%, 50%, 75%, and 95%) of one hundred simulations ($\alpha = 0.3$) with the same input sequences as in Figure 3. From left to right; $\hat{a}_1(x)$, $\hat{b}_2(x)$, and $\hat{b}_4(x)$.

running district heating plant are used. To mimic the real situation (c.f. Section 4.2) the flow is filtered with exponential smoothening using a forgetting factor of 0.8.

Figure 5 shows the results obtained with local quadratic estimates when the structure of the true system is assumed known. Compared with the previous plots it is clear that the correlation of the input sequences deteriorates the estimates, even in regions where data is dense. The main characteristics are, however, still identifiable.

In Figures 6 and 7 the results when including $b_3(x_{t-1})u_{t-3}$ in the model used for estimation are shown. In this case the local constant estimates perform very poor. The local quadratic estimates perform well, except for regions were data is sparse (compare with the histogram in Figure 1).

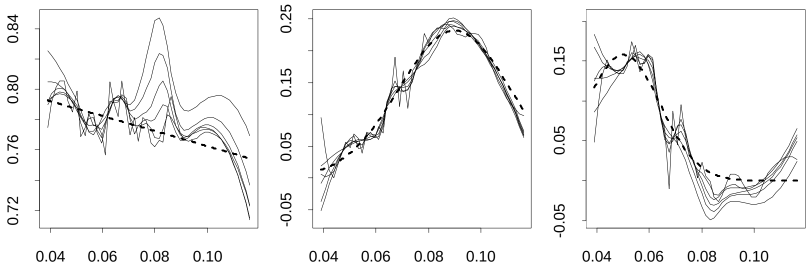


Figure 5: Measurements of flow (m^3/s) from a running plant as input and local quadratic estimates. From left to right; $\hat{a}_1(x)$, $\hat{b}_2(x)$, and $\hat{b}_4(x)$.

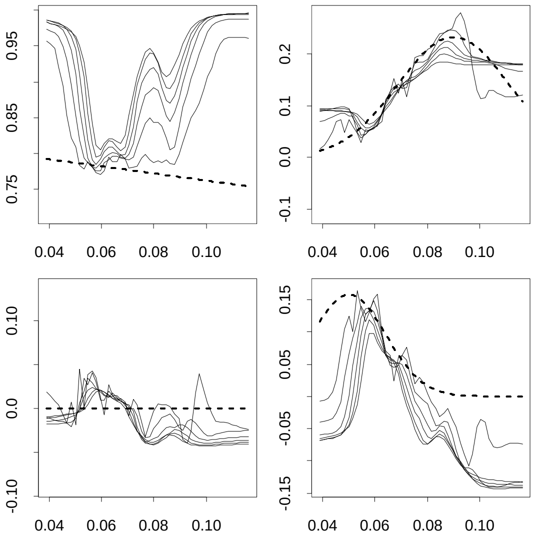


Figure 6: Measurements of flow (m^3/s) from a running plant as input and local constant estimates. From left to right; $\hat{a}_1(x)$, $\hat{b}_2(x)$, $\hat{b}_3(x)$, and $\hat{b}_4(x)$.

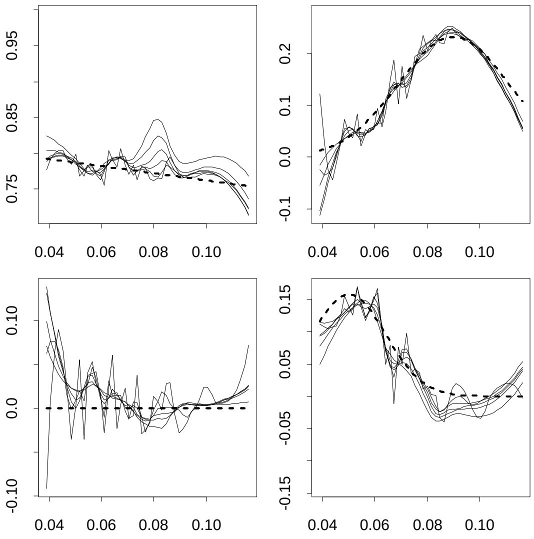


Figure 7: Measurements of flow from a running plant as input and local quadratic estimates. From left to right; $\hat{a}_1(x)$, $\hat{b}_2(x)$, $\hat{b}_3(x)$, and $\hat{b}_4(x)$.

The performance in border regions can be improved if prior knowledge about the individual functions is available. If the estimates corresponding to Figure 7 are recalculated with the local approximations $a_1(x)$ - strait line, $b_2(x)$ and $b_4(x)$ - quadratic, $b_3(x)$ - constant. This removes much of the fluctuation of $\hat{b}_3(x)$ and also increases the performance of the remaining estimates.

4 Application to a real system

In this section the method is applied to data obtained from the district heating plant "Høje Taastrup Fjernvarme" near Copenhagen in Denmark. For the periods considered the energy were supplied from one plant only.

4.1 Data

The data covers the periods from 1 April until 31 May and 1 September until 17 December, 1996. Data consists of five minute samples of supply temperature and flow at the plant together with the network temperature at a major consumer, consisting of 84 households. In 1996 that consumer used 1.2% of the produced energy.

The measurements of the flow and temperatures are occasionally erroneous. In order to find these measurements the data were investigated by visual methods; five minutes samples were excluded for 999 time values. Based on the observations half-hour averages were calculated, and these were excluded if any of the five minute samples were excluded or missing. In total 724 half-hour averages were excluded, yielding a total of 7388 half-hour values which are assumed valid.

4.2 Models

The network temperature is modelled by models of the structure (8), where one time step corresponds to 30 minutes, y_t represents the network temperature, u_t represents the supply temperature, and x_t represents a filtered value of the flow.

Consider a simple district heating network with only one consumer and one plant. Disregarding diffusion, the time delay $\tau(t)$ of a water particle leaving the plant at time t is related to the distance d between plant and consumer by the relation $d = \int_t^{t+\tau(t)} v(s)ds$, where $v(s)$ is the velocity of the particle at time s. From this simple model it is seen that the flow should be filtered by averaging over past values, but with varying horizon. In practice this is implemented by assuming a known volume for the district heating pipe between the plant and the major consumer. Hereafter the filtered values of the flow is found as the average of the five minute samples filling the pipe, see (Nielsen et al. 1997) for a brief description of how the volume is estimated.

Below results corresponding to two different model structures are presented, (i) the conditional parametric Finite Impulse Response (FIR) model $y_t = \sum_{i=0}^{30} b_i(x_t)u_{t-i} + e_t$, and (ii) the conditional parametric ARX-model $y_t = a(x_t)y_{t-1} + \sum_{i=3}^{15} b_i(x_t)u_{t-i} + e_t$.

4.3 Results

For both the FIR-model and the ARX-model described in the previous section local quadratic estimates and nearest neighbour bandwidths are used. The coefficient functions are estimated at 50 equally spaced points ranging from the 2% to the 98% quantile of the filtered flow.

The coefficient functions of the FIR-model are estimated for α equal to $0.1, 0.2, \ldots, 0.9$. In Figure 8 the impulse response as a function of the flow is displayed for $\alpha = 0.4$. Equivalent plots for the remaining bandwidths revealed that for $\alpha \leq 0.2$ the fits are too noisy, whereas in all cases sufficiently smoothness is obtained for $\alpha = 0.5$. Only minor differences in the fits are observed for $\alpha \in \{0.3, 0.4, 0.5\}$.

In Figure 9 a contour plot corresponding to Figure 8 is shown. From the plot the varying time delay of the system is revealed, it seems to vary from three lags when the flow is large to approximately ten lags when the flow is near its minimum. The peak at lag zero for the lower flows is

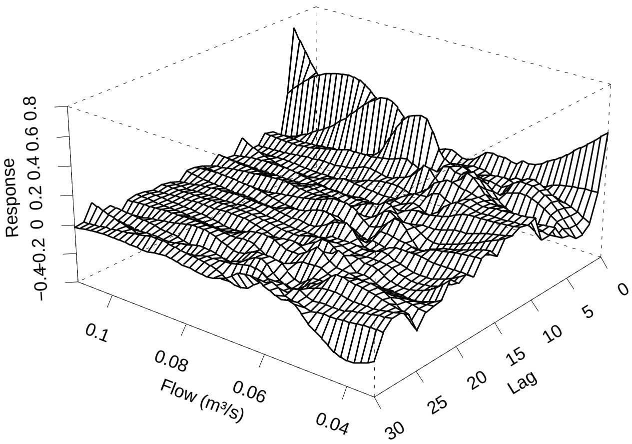


Figure 8: Impulse response function of the FIR-model.

clearly an artifact. It is probably due to the correlation structure of the flow. The sample autocorrelation function shows a local peak at lag 24, and therefore the peak at lag zero can be compensated by higher lags.

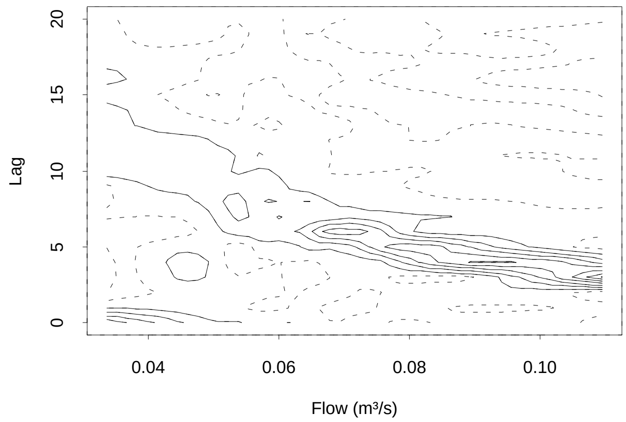


Figure 9: Contour plot of the impulse response function of the FIR-model ($\alpha = 0.4$). The contour lines is plotted from -0.1 to 0.7 in steps of 0.1. Lines corresponding to non-positive values are dotted.

The residuals of the FIR-model show a diurnal variation. The sample inverse autocorrelation function of the residuals indicates that a AR(1)-model can account for most of the correlation, this corresponds to the order of the auto regression used in (Madsen et al. 1996) which considers the same district heating system.

Based on the results obtained for the FIR-model the ARX-model described in Section 4.2 is purposed. The coefficient functions of the model are estimated for α equal to 0.3, 0.4, and 0.5. Very similar results are obtained for the three different bandwidths. For $\alpha = 0.4$ the impulse response as a function of the flow is displayed in Figure 10. The varying time delay is clearly revealed. In Figure 11 the stationary gain of the two models and the pole of the ARX-model are shown. From the values of the stationary gain it is seen that the temperature loss changes from 6% when the flow is large to 12-15% when the flow is small.

The residuals of the ARX-model show a weak diurnal variation, only very weak autocorrelation (0.07 in lag two), and no dependence of supply temperature and flow. If the coefficients of the ARX-model are fitted as global constants the central 95% of the residuals spans 2.1 $°C$ opposed to 1.7 $°C$ for the conditional parametric ARX-model.

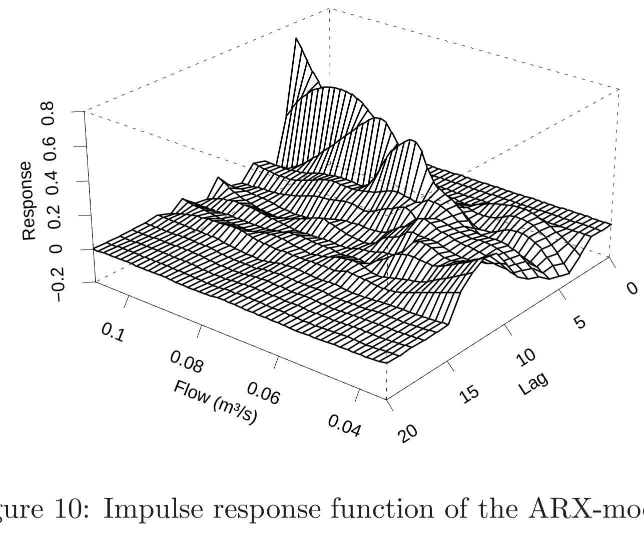


Figure 10: Impulse response function of the ARX-model.

5 Conclusions

In this paper conditional parametric ARX-models are suggested and studied. These non-linear models are obtained as traditional ARX-models in which the parameters are replaced by smooth functions. Simulations show that kernel estimates (local constants) in general are inferior

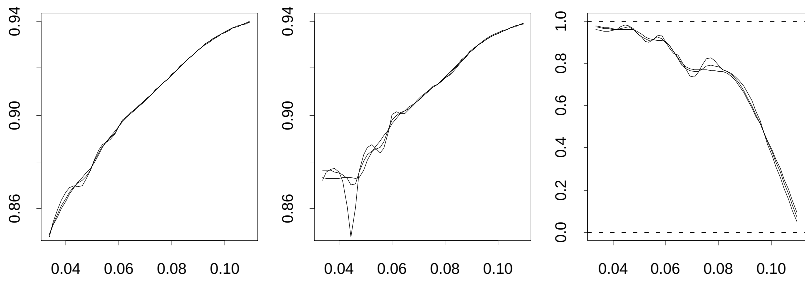


Figure 11: Stationary gain of the FIR-model (left) and ARX-model (middle) and the pole of the ARX-model (right) all plotted against the flow and for α equal to 0.3, 0.4, and 0.5.

to local quadratic estimates. In the case of correlated input sequences the kernel estimates are rather unreliable, but local quadratic estimates are in general quite reliable.

If the observations are sparse in border regions of the variable(s) defining the weight, the local quadratic estimate may have increased bias and/or variance in these regions. Consequently it is important to investigate the distribution of the variable(s) defining the weight when interpreting the estimates.

When applied to real data from a district heating system in which case numerous coefficient functions is necessary the method seem to perform well in that the results obtained are quite plausible. Furthermore, the residuals behave appropriately.

6 Discussion

The poor performance of the kernel estimates is probably related to the inability of the approximation to fit the data locally. For the estimation of a regression function ($\boldsymbol{z}_t = 1$) this has been described by Hastie & Loader (1993).

In this paper the selection of the bandwidth is not considered. If prior knowledge of the curvature of the coefficient-functions are available this may be used to select a bandwidth such that a polynomial of a certain order is a reasonable local approximation. A more informal method is to select a bandwidth for which the estimated functions attains a certain degree of smoothness. In the context of time series it seems appropriate to use forward validation (Hjorth 1994) for guidance on bandwidth selection. As for cross validation, this approach will probably yield a flat optimum (Hastie & Tibshirani 1990, p. 42), and so, some judgment is called upon.

The interpretation of the function value estimates would be greatly facilitated by results concerning the statistical properties of the estimates. It is believed that it is possible to derive such properties based on the results obtained for locally weighted regression, see e.g. (Cleveland & Devlin 1988). Without the theoretical work it is still possible to address the precision of the function estimates by for instance bootstrapping.

The conditional parametric ARX-models may be too flexible for some applications. Especially it is of interest to be able to fit some of the functions as global constants. This could probably be accomplished by applying the methods described in (Hastie & Tibshirani 1993).

For temperature control applications in district heating systems it is expected that the identified conditional parametric ARX-model is superior to more traditional adaptive controllers as considered for instance in (Madsen et al. 1996), because the non-linearity is directly modelled as opposed to the successive linearizations used in traditional adaptive estimation techniques.

References

Anderson, T. W., Fang, K. T. & Olkin, I., eds (1994), *Multivariate Analysis and Its Applications*, Institute of Mathematical Statistics, Hayward, chapter Coplots, Nonparametric Regression, and conditionally Parametric Fits, pp. 21–36.

Cleveland, W. S. (1979), 'Robust locally weighted regression and smoothing scatterplots', *Journal of the American Statistical Association* **74**, 829–836.

Cleveland, W. S. & Devlin, S. J. (1988), 'Locally weighted regression: An approach to regression analysis by local fitting', *Journal of the American Statistical Association* **83**, 596–610.

Cleveland, W. S., Devlin, S. J. & Grosse, E. (1988), 'Regression by local fitting: Methods, properties, and computational algorithms', *Journal of Econometrics* **37**, 87–114.

Härdle, W. (1990), *Applied Nonparametric Regression*, Cambridge University Press, Cambridge, UK.

Hastie, T. & Loader, C. (1993), 'Local regression: Automatic kernel carpentry', *Statistical Science* **8**, 120–129. (Discussion: p129-143).

Hastie, T. & Tibshirani, R. (1990), *Generalized Additive Models*, Chapman & Hall, London/New York.

Hastie, T. & Tibshirani, R. (1993), 'Varying-coefficient models', *Journal of the Royal Statistical Society, Series B, Methodological* **55**, 757–796.

Hjorth, J. S. U. (1994), *Computer Intensive Statistical Methods: Validation Model Selection and Bootstrap*, Chapman & Hall, London/New York.

Madsen, H., Nielsen, T. S. & Søgaard, H. T. (1996), *Control of Supply Temperature in District Heating Systems*, Department of Mathematical Modelling, Technical University of Denmark, Lyngby, Denmark.

Miller, A. J. (1992), '[Algorithm AS 274] Least squares routines to supplement those of Gentleman', *Applied Statistics* **41**, 458–478. (Correction: 94V43 p678).

Nielsen, H. A., Nielsen, T. S. & Madsen, H. (1997), ARX-models with parameter variations estimated by local fitting, *in* Y. Sawaragi & S. Sagara, eds, '11th IFAC Symbosium on System Identification', Vol. 2, pp. 475–480.

Palsson, O., Madsen, H. & Søgard, H. (1994), 'Generalized predictive control for non-stationary systems', *Automatica* **30**, 1991–1997.

Søgaard, H. & Madsen, H. (1991), On-line estimation of time-varying delays in district heating system, *in* 'Proc. of the 1991 European Simulation Multiconference', pp. 619–624.

Stone, C. J. (1977), 'Consistent nonparametric regression', *The Annals of Statistics* **5**, 595–620. (C/R: p620-645).

Tong, H. (1990), *Non-linear Time Series. A Dynamical System Approach*, Oxford University Press, Oxford.

B

Paper B

LFLM Version 1.0 – An S-PLUS / R library for locally weighted fitting of linear models

Originally published as

> Henrik Aalborg Nielsen. LFLM version 1.0, an S-PLUS / R library for locally weighted fitting of linear models. Technical Report 22, Department of Mathematical Modelling, Technical University of Denmark, DK-2800 Lyngby, Denmark, 1997.

LFLM Version 1.0

An S-PLUS / R library for locally weighted fitting of linear models

Henrik Aalborg Nielsen[1]

Abstract

The conditional parametric model is an extension of the well known linear regression model, obtained by replacing the parameters by smooth functions. Estimation in such models may be accomplished by fitting a, possibly larger, linear model locally to some explanatory variable(s). In this report the conditional parametric model is described together with a method of estimation. An S-PLUS / R implementation is described and an example given. Since the user interface is similar to other S-PLUS / R functions for regression the software is easy to use. Furthermore, to increase speed, the core of the program is written in the ANSI-C programming language. The software allows for experimentation with variable bandwidth selection procedures and evaluation structures.

1 Introduction

Conditional parametric models are considered in (Chambers & Hastie 1991), (Hastie & Tibshirani 1993), and (Anderson, Fang & Olkin 1994). These models may be viewed as linear regression models in which the parameters are replaced by smooth functions of other explanatory variables. The purpose of the analysis is to estimate and make inference about these functions, without assuming a parametric form. The models considered in (Chambers & Hastie 1991) and (Anderson et al. 1994) are rather restricted, since the originating linear model need to be a strait line or a linear (hyper) surface. Conditional parametric models are a subset of the varying-coefficient models described in (Hastie & Tibshirani 1993), but the general description requires a more complicated estimation procedure.

[1]Department of Mathematical Modelling, Technical University of Denmark, DK-2800 Lyngby, Denmark

In Section 2 a general description of the model is given, together with a method for estimation. In Section 3 an S-PLUS / R library which is able to estimate the smooth functions is described. An example on how to use the software is shown in Section 4. In Section 5 details on how to obtain the source code are given. Finally, in Section 6 we conclude on the paper.

2 Theory

2.1 Model

A conditional parametric model is a model of the form

$$Y_i = \boldsymbol{z}_i^T \boldsymbol{\theta}(\boldsymbol{x}_i) + e_i; \quad i = 1, \ldots, N, \tag{1}$$

where the response Y_i is a stochastic variable, $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$ are explanatory variables, e_i is i.i.d. $N(0, \sigma^2)$, $\boldsymbol{\theta}(\cdot)$ is a vector of unknown but smooth functions with values in $\mathcal{R}$, and $i = 1, \ldots, N$ are observation numbers. When $\boldsymbol{x}_i$ is constant across observations the model reduces to a linear model, hereof the name.

2.2 Local constant estimates

Estimation in (1) aims at estimating the functions $\boldsymbol{\theta}(\cdot)$ within the space spanned by the observations of $\boldsymbol{x}_i; i = 1, \ldots, N$. The functions is only estimated for distinct values of the argument $\boldsymbol{x}$. Below $\boldsymbol{x}$ denotes one single of these points and $\hat{\boldsymbol{\theta}}(\boldsymbol{x})$ denotes the estimates of the coefficient functions, when the coefficient functions are evaluated at $\boldsymbol{x}$.

One solution to the estimation problem is to replace $\boldsymbol{\theta}(\boldsymbol{x}_i)$ in (1) with a constant vector $\boldsymbol{\theta}(\boldsymbol{x})$ and fit the resulting model locally to $\boldsymbol{x}$, using weighted least squares. Below two similar methods of allocating weights to the observations are described, for both methods the weight function $W : \mathcal{R}_0 \rightarrow \mathcal{R}_0$ is a nowhere increasing function. The weight functions available in LFLM are listed in Table 1 on page 57.

In the case of a spherical kernel the weight on observation i is determined by the Euclidean distance $||\boldsymbol{x}_i - \boldsymbol{x}||$ between $\boldsymbol{x}_i$ and $\boldsymbol{x}$, i.e.

$$w_i(\boldsymbol{x}) = W\left(\frac{||\boldsymbol{x}_i - \boldsymbol{x}||}{h(\boldsymbol{x})}\right). \tag{2}$$

A product kernel is characterized by distances being calculated for one dimension at a time, i.e.

$$w_i(\boldsymbol{x}) = \prod_j W\left(\frac{|x_i(j) - x(j)|}{h(\boldsymbol{x})}\right), \tag{3}$$

where the multiplication is over the dimensions of $\boldsymbol{x}$. The scalar $h(\boldsymbol{x}) > 0$ is called the bandwidth. If $h(\boldsymbol{x})$ is constant for all values of $\boldsymbol{x}$ it is denoted a fixed bandwidth. If $h(\boldsymbol{x})$ is chosen so that a certain fraction (α) of the observations fulfill $||\boldsymbol{x}_i - \boldsymbol{x}|| \leq h(x)$ it is denoted a nearest neighbour bandwidth. If $\boldsymbol{x}$ has dimension of two or larger, scaling of the individual elements of $\boldsymbol{x}_i$ before applying the method should be considered, see e.g. (Cleveland & Devlin 1988). Rotating the coordinate system in which $\boldsymbol{x}_i$ is measured may also be relevant.

Note that if $\boldsymbol{z}_i = 1$ for all i the method of estimation reduces to determining the scalar $\hat{\theta}(\boldsymbol{x})$ such that $\sum_{i=1}^{n} w_i(\boldsymbol{x})(y_i - \hat{\theta}(\boldsymbol{x}))^2$ is minimized, i.e. the method reduces to kernel estimation (Härdle 1990, p. 30). For this reason the described method of estimation of $\boldsymbol{\theta}(\boldsymbol{x})$ in (1) is called kernel or local constant estimation.

2.3 Local polynomial estimates

If the bandwidth $h(\boldsymbol{x})$ is sufficiently small the approximation of $\boldsymbol{\theta}(\cdot)$ as a constant vector near $\boldsymbol{x}$ is good. This implies that a relatively low number of observations is used to estimate $\boldsymbol{\theta}(\boldsymbol{x})$, resulting in a noisy estimate or large bias if the bandwidth is increased. See also the comments on kernel estimates in (Anderson et al. 1994)

It is, however, well known that locally to $\boldsymbol{x}$ the elements of $\boldsymbol{\theta}(\cdot)$ may be approximated by polynomials, and in many cases these will be good approximations for larger bandwidths than those corresponding to local constants. Local polynomial approximations are easily included in the method described. Let $\theta_j(\cdot)$ be the j'th element of $\boldsymbol{\theta}(\cdot)$ and let $\boldsymbol{P}_d(\boldsymbol{x})$ be a column vector of terms in a d-order polynomial evaluated at $\boldsymbol{x}$, if for instance $\boldsymbol{x} = [x_1\ x_2]^T$ then $\boldsymbol{P}_2(\boldsymbol{x}) = [1\ x_1\ x_2\ x_1^2\ x_1 x_2\ x_2^2]^T$. Furthermore, let $\boldsymbol{z}_i = [z_{1i} \dots z_{pi}]^T$. With

$$\boldsymbol{u}_i^T = \left[z_{1i} \boldsymbol{P}_{d(1)}^T(\boldsymbol{x}_i) \dots z_{ji} \boldsymbol{P}_{d(j)}^T(\boldsymbol{x}_i) \dots z_{pi} \boldsymbol{P}_{d(p)}^T(\boldsymbol{x}_i) \right] \tag{4}$$

and

$$\hat{\boldsymbol{\phi}}^T(\boldsymbol{x}) = [\hat{\boldsymbol{\phi}}_1^T(\boldsymbol{x}) \dots \hat{\boldsymbol{\phi}}_j^T(\boldsymbol{x}) \dots \hat{\boldsymbol{\phi}}_p^T(\boldsymbol{x})], \tag{5}$$

where $\hat{\boldsymbol{\phi}}_j(\boldsymbol{x})$ is a column vector of local constant estimates at $\boldsymbol{x}$ corresponding to $z_{ji} \boldsymbol{P}_{d(j)}(\boldsymbol{x}_i)$, estimation is handled as described in Section 2.2, but fitting the linear model

$$Y_i = \boldsymbol{u}_i^T \boldsymbol{\phi}(\boldsymbol{x}) + e_i; \quad i = 1, \dots, N, \tag{6}$$

locally to $\boldsymbol{x}$. Hereafter the elements of $\boldsymbol{\theta}(\boldsymbol{x})$ is estimated by

$$\hat{\theta}_j(\boldsymbol{x}) = \boldsymbol{P}_{d(j)}^T(\boldsymbol{x})\, \hat{\boldsymbol{\phi}}_j(\boldsymbol{x}); \quad j = 1, \dots p. \tag{7}$$

When $\boldsymbol{z}_j = 1$ for all j this method is identical to the method by Cleveland & Devlin (1988), with the exception that they center the elements of $\boldsymbol{x}_i$ used in $\boldsymbol{P}_d(\boldsymbol{x}_i)$ around $\boldsymbol{x}$ and so $\boldsymbol{P}_d(\boldsymbol{x}_i)$ must be recalculated for each value of $\boldsymbol{x}$ considered.

Example: If the first element of $\boldsymbol{\theta}(\cdot)$ is approximated locally by a 2nd order polynomial and if $\boldsymbol{x}_i = [x_{1i}\ x_{2i}]^T$ then z_{1i} in (1) is replaced by the elements $z_{1i},\ z_{1i}x_{1i},\ z_{1i}x_{2i},\ z_{1i}x_{1i}^2, z_{1i}x_{1i}x_{2i},$ and $z_{1i}x_{2i}^2$. If the corresponding parameters are denoted $\phi_{1,0},\ \phi_{1,1},\ \phi_{1,2},\ \phi_{1,11},\ \phi_{1,12},$ and $\phi_{1,22}$ the estimate of $\theta_1(\boldsymbol{x})$ is $\hat{\phi}_{1,0} + \hat{\phi}_{1,1}x_1 + \hat{\phi}_{1,2}x_2 + \hat{\phi}_{1,11}x_1^2 + \hat{\phi}_{1,12}x_1x_2 + \hat{\phi}_{1,22}x_2^2$.
$\square$

2.4 Heteroscedasticity

For estimation purposes only observations in an neighbour-hood of $\boldsymbol{x}$ is used. Consequently, it is not required that e_i has constant variance. It is sufficient that the variance is approximately constant within a neighbourhood of $\boldsymbol{x}_i$, i.e. we may write e_i is i.i.d. $N(0, \sigma^2(\boldsymbol{x}_i))$.

It is possible to extend this to local polynomial estimation of $\sigma^2(\boldsymbol{x}_i)$. However, it is required that the local approximations used are strictly positive and it would be obvious to approximate $\log(\sigma^2(\boldsymbol{x}_i))$. Note that this may pose problems if the true variance is very small. Furthermore, it is necessary to use a local likelihood method (Hastie & Tibshirani 1990). These methods are not included in LFLM.

2.5 The smoother matrix

As for linear regression the fitted values $\hat{y}_i,\ i = 1, \ldots, N$ is linear combinations of the observations $y_i,\ i = 1, \ldots, N$. If the observations are arranged in a column vector $\boldsymbol{Y}$ this may be expressed as

$$\hat{\boldsymbol{Y}} = \boldsymbol{L}\boldsymbol{Y}, \tag{8}$$

where the $n \times n$ matrix $\boldsymbol{L}$ is called the smoother matrix and is independent of $\boldsymbol{Y}$. For linear regression the matrix is often called the hat-matrix (Jørgensen 1993).

The property (8) is shared with other smoothers (Hastie & Tibshirani 1990) and hence the model, variance, and error degrees of freedom can be calculated

$$\begin{aligned} df_{mod} &= \operatorname{tr}(\boldsymbol{L}) && (9)\\ df_{var} &= \operatorname{tr}(\boldsymbol{L}\boldsymbol{L}^T) && (10)\\ df_{err} &= N - \operatorname{tr}(2\boldsymbol{L} - \boldsymbol{L}\boldsymbol{L}^T) && (11) \end{aligned}$$

For linear regression df_{mod}, df_{var}, and $N - df_{err}$ are identical due to the special properties of $\boldsymbol{L}$ in this case.

To see that (8) is true let $\boldsymbol{U}$ be the design matrix corresponding to local constant estimates, i.e. row i is $\boldsymbol{u}_i$ from (4). Let $\boldsymbol{W}_i$ be a diagonal matrix, where element (k, k) is the weight on observation k when $\boldsymbol{x}_i$ is used as fitting point, i.e. $\boldsymbol{x} = \boldsymbol{x}_i$. The local constant (intermediate) estimates can then be expressed as

$$\hat{\boldsymbol{\phi}}(\boldsymbol{x}_i) = \left[\boldsymbol{U}^T\boldsymbol{W}_i\boldsymbol{U}\right]^{-1}\boldsymbol{U}^T\boldsymbol{W}_i\boldsymbol{Y}. \tag{12}$$

Hence the fitted value at observation i equals

$$\hat{y}_i = \boldsymbol{u}_i^T\hat{\boldsymbol{\phi}}(\boldsymbol{x}_i) = \boldsymbol{u}_i^T\left[\boldsymbol{U}^T\boldsymbol{W}_i\boldsymbol{U}\right]^{-1}\boldsymbol{U}^T\boldsymbol{W}_i\boldsymbol{Y}. \tag{13}$$

Consequently, the vector of fitted values can be written as (8). If the weights are not chosen based on the values of $\boldsymbol{Y}$ or $\hat{\boldsymbol{Y}}$ then $\boldsymbol{L}$ only depend on $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$, $i = 1, \ldots, N$.

3 Software

The weighted least squares problem, described in Section 2, is solved by the algorithm described in (Miller 1992). The algorithm was originally written in Fortran but here a port to C by A. Shah is used. This was obtained as `pub/C-numanal/as274_fc.tar.z` by anonymous ftp from `usc.edu`.

The local constant estimation is implemented as a function written in ANSI-C. This is also true for most of the calculations related to the smoother matrix. The remaining part of the program is written in S-PLUS / R.

The S-PLUS / R front end constitutes a user friendly interface which is described in this section. Some experience with S-PLUS or R is assumed, including the notion of classes and methods (Statistical Sciences 1993). An example on how to use the software is given in Section 4.

Below a fixed width font indicates computer terms, e.g. files, S-PLUS / R objects, and arguments to S-PLUS / R functions. If the string is ended with "`()`" this indicates that reference to a S-PLUS / R function is made.

3.1 The main function

The main function of LFLM is `lflm()`. The arguments of this function are described in this section. Below the term "fitting points" is used to denote the points in which the estimates are to be calculated ($\boldsymbol{x}$ in Section 2.2). The handling of weight functions and bandwidth are inspired by the LOCFIT program by Clive Loader, Lucent Technologies, Bell Laboratories, see:
`http://cm.bell-labs.com/stat/project/locfit/index.html`

3.1.1 Required arguments:

The arguments listed below must be supplied to `lflm()`.

- `formula:` Model specification, see Section 3.1.3.
- `alpha:` The bandwidth to use, by default this is interpreted as a fixed bandwidth. If a nearest neighbour bandwidth is used `alpha` specifies the fraction of observations to be covered for each fitting point. In this case `alpha` may have length two, the second element is then taken as a lower bound on the bandwidth found by the nearest neighbour method. To allow experimentation with variable-bandwidth selection procedures, it is also possible to specify an individual bandwidth for each fitting point. In this case the length of `alpha` must equal the number of fitting points, see also the description of argument `bt` in Section 3.1.2.
- `data:` Data frame containing the data.

3.1.2 Optional arguments:

Default values are supplied for the following arguments. These are by no means guaranteed to be appropriate, especially the bandwidth type `bt`, the degree of the local polynomial approximations `degree`, and the number / placement of fitting points `n.points` and `x.points` should be considered.

- `degree:` Degree of the local polynomial approximations. If the length is one this is used for all elements of $\boldsymbol{\theta}(\cdot)$. If the length is different from one it must equal the number of functions to estimate and the order must correspond to the global part of `formula`. Default: `2`.
- `CP:` If `TRUE` crossproduct terms are included in the local polynomial approximations. As for `degree` the length must either be one or equal the number of functions to estimate. Default: `TRUE`.

- `scale`: Vector, the length of which correspond to the dimension of $\boldsymbol{x}$. For each dimension of the fitting points and the corresponding observations the values are divided by the corresponding element of `scale` before distances, and consequently weights, are calculated. Default: No scaling.

- `bt`: Type of bandwidth; fixed (`"fix"`), nearest neighbour (`"nn"`), or user specified (`"user"`), i.e. an individual bandwidth for each fitting point. If user specified bandwidths are used the argument `x.points`, described below, must be supplied. Default: `"fix"`.

- `kt`: Type of kernel; spherical (`"sph"`) or product (`"prod"`), see Section 2.2. Default: `"sph"`.

- `kern`: Type of kernel or weight function; box (`"box"`), triangle (`"tangl"`), tricube (`"tcub"`), or Gaussian (`"gauss"`), see Table 1. Default: `"tcub"`.

- `n.points`: Number of fitting points per dimension of the data corresponding to the local formula, c.f. Sections 3.1.3 and 3.3. This argument is disregarded if `x.points` is specified. Default: `10`.

- `x.points`: Data frame containing the fitting points. The names and order must correspond to the local formula. This argument allows experimentation with evaluation structures. Default: Constructed using `n.points`.

- `na.action`: Function specifying the action to take when missing data (`NA`) is found in `data`, often `na.omit` is desirable. Default: `na.fail`.

- `weight`: Vector, which length is the number of observations with no missing values, specifying the weight on the observations. Default: unweighted.

- `calc.df`: Should the degrees of freedom of the model be calculated? In version 1.0 this is possible only when the estimates are calculated for all observations. Default: `FALSE`.

- `save.smoother.matrix`: Should the smoother matrix be calculated and saved? In version 1.0 this is possible only when the estimates are calculated for all observations. Default: `FALSE`.

Name	kern argument	Weight function
Box	"box"	$W(u) = \begin{cases} 1, & u \in [0;1) \\ 0, & u \in [1;\infty) \end{cases}$
Triangle	"tangl"	$W(u) = \begin{cases} 1-u, & u \in [0;1) \\ 0, & u \in [1;\infty) \end{cases}$
Tribube	"tcub"	$W(u) = \begin{cases} (1-u^3)^3, & u \in [0;1) \\ 0, & u \in [1;\infty) \end{cases}$
Gauss	"gauss"	$W(u) = \exp(-u^2/2)$

Table 1: Weight functions available in LFLM.

- `circ`: Should the distances be calculated between points on the unit circle? This argument may only be `TRUE` for local constant estimates and for one-dimensional fitting points. Default: `FALSE`.

- `dump`: Should `lflm()` dump if an error from the C-code is trapped. Default: `TRUE`.

3.1.3 Model specification:

Model specification (`formula`) follows the usual S-PLUS / R formula language

```
<response> ~ (<global formula>) | (<local formula>)
```

where the global formula is a S-PLUS / R model specification corresponding to $\boldsymbol{z}$ in (1), and the local formula corresponds to $\boldsymbol{x}$ in (1). The elements in the local formula must be separated by asterisks.

Example: If `x1`, `x2`, `z1`, `z2`, and `y` are numeric vectors (*not factors*) then

```
y ~ (z1 + z2) | (x1 * x2),
```

specifies the model $y_i = \theta_0(x_{1i}, x_{2i}) + \theta_1(x_{1i}, x_{2i})z_{1i} + \theta_0(x_{1i}, x_{2i})z_{2i} + e_i$. As usual the intercept term can be dropped from the model by replacing `z1 + z2` with `-1 + z1 + z2`.

Note: The function does not handle factor variables and functions of numeric variables included in the global formula correctly. Factor variables must be replaced by a number of coding variables before using the program. Similar steps must be used to include functions of numeric variables.

3.1.4 Output:

After successful completion `lflm()` returns a list of class `lflm` with the following components:

- `call`: An image of the call that produced the list.
- `data.name`: Under S-PLUS: The name of the data frame used for estimation. Under R: the string `"unknown"`; assign it manually after the call to `lflm()`.
- `run.time`: The date and time at which `lflm()` were called, as returned by `date()`.
- `formula`, `degree`, `CP`, `circ`, `kt`, `kern`, `bt`, `alpha`, and `scale`: Copies from the call to `lflm()`.
- `x.points`: Data frame in which the rows are the fitting points used.
- `est`: Data frame in which the rows correspond to the rows in `x.points` and the columns correspond to the function estimates.
- `df`: If requested; the degrees of freedom of the model, otherwise: `NA`.
- `S`: If requested; the smoother matrix, otherwise: `NA`.
- `bandwidth`: Vector containing the actual bandwidth used for each fitting point, i.e. the positions correspond to the rows of `x.points`.

- `rank.defic`: Vector containing non-negative integers indicating the rank deficiency for each fitting point, as reported by the WLS algorithm. A warning will be issued by `lflm()` if these are not all zero.

- `loc.nparam`: The number of local constants used, i.e. the number of parameters estimated by the WLS algorithm.

- `err.func`: If positive one of the C-functions in the WLS algorithm has returned an error condition. The value can be used to locate the function in the C code. By default a positive value will cause `lflm()` to stop without returning a result.

- `err.val`: If `err.func` is positive; the error value returned by the function, see (Miller 1992).

3.1.5 R notes

As mentioned above the name of the data frame used are not returned when the software runs under R. To be able to calculate the fitted values or the residuals, see Section 3.2, the correct name must be assigned to the list returned by `lflm()`. For instance if the result is saved in the list `fit` and the data frame containing the data are called `mydata`, the command `fit$data.name <- "mydata"` must be issued.

Furthermore, when using R (version 0.50), a call to `lflm()` will cause the warning `"some row names are duplicated; argument ignored"`. This can safely be disregarded.

3.2 Methods

Methods, corresponding to objects of class `lflm`, are supplied for the functions `coef()`, `fitted()`, `lines()`, `plot()`, `points()`, `predict()`, `print()`, `residuals()`, and `summary()`. These functions are briefly described below. Often these functions will be appropriate only for preliminary analyses and it will often be necessary to write application specific functions. Graphics are mainly handled for the case of $\boldsymbol{x}_i$ in (1) having dimension two or less, the plot method allows a call to `coplot()` for higher dimensions.

When $\boldsymbol{x}_i$ in (1) has dimension two or more methods using interpolation (`fitted()`, `predict()`, and `residuals()`) requires the fitting points to be placed in a grid similar to the grid generated when the argument `n.points` is used. However, the grid need not to be rectangular, it is sufficient that for all $i = 1, \ldots, N$ a (hyper) cube of fitting points containing $\boldsymbol{x}_i$ can be identified.

Below the methods are described.

`print()` only print some of the components of the object. Since `print()` is called when an object is returned to the top level, or if the name of an object is just typed at the prompt, `unclass()` must be used to see the full content of the object.

`summary()` returns a list containing the formula, the degrees of freedom (if calculated), the number of local parameters (length of $\boldsymbol{\phi}$ in (5)), the `degree` and `CP` arguments from the call to `lflm()`, information of how weights were constructed, summary information of the estimates (incl. fitting points and bandwidth), and information on whether rank deficiencies were detected.

`coef()` returns a data frame, in which the first column(s) are the fitting points, followed by a column containing the bandwidth used (possibly on the scaled coordinates), the remaining columns contain the estimates of the functions at the fitting points.

`fitted()` returns the fitted values of the response. These are calculated using interpolation between fitting points.

`residuals()` returns the residuals using `fitted()`.

`predict()` returns predictions based on the fitted model. Three types of predictions are available through the `type` argument; (i) `"response"` (default) predictions of the dependent variable, (ii) `"terms"` each element of $\boldsymbol{z}^T\boldsymbol{\theta}(\boldsymbol{x})$ in (1) separately, and (iii) `"coefficients"` each element of $\boldsymbol{\theta}(\boldsymbol{x})$ separately. The calculations are performed using interpolation. By default missing values are returned if interpolation is not possible, it is possible to parse arguments to the function carrying out the interpolation.

`plot()`: By default, when $\boldsymbol{x}_i$ in (1) is one dimensional this function plots all the estimated functions; the user should set up the graphics devise to contain more than one plot before calling this function. Arguments can be parsed to the builtin plot function of S-PLUS / R. When $\boldsymbol{x}_i$ has dimension two, and S-PLUS is used, surface plotting using Trellis Graphics are available, also `coplot()` may be called. In other cases only `coplot()` may be called.

`points()` / `lines()` adds points / lines to the current plot, the argument `what` is used to specify the estimate to add and therefore should be among `names(coef(obj))`, where `obj` is a list of class `lflm`. These methods are only implemented for the case where $\boldsymbol{x}_i$ in (1) is one dimensional.

In the case that the data frame (`data`) contain missing values and `lflm()` is called with argument `na.action=na.omit` the functions `fitted()` and `residuals()` will return vectors of the same length as the number of rows in the data frame, but with missing values inserted as appropriate. This is not consistent with other regression functions in S-PLUS, but we find it more convenient for general use.

3.3 Surface estimation

When $\boldsymbol{x}$ in (1) has dimension two or more, using the argument `n.points` to `lflm()` will result in a rectangular grid of fitting points being spanned parallel to the coordinate axes. Often, no observations will be present in the corners of this grid. In this situation we suggest that the fitting points are supplied directly through the `x.points` argument to `lflm()`.

To facilitate this process it is possible to use `lflm.data.grid()` to generate the rectangular grid and `in.chull()` to delete the fitting points not inside the convex hull spanned by the observations $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N$. However, `in.chull()` works only when $\boldsymbol{x}$ has dimension two.

Note: The function `in.chull()`, supplied together with the software, is not used by other functions. Therefore, it can safely be deleted or renamed.

4 Example

The ethanol example of (Chambers & Hastie 1991, Section 8.2.2) is used since this clearly illustrates some of the differences between the standard S-PLUS function used of locally weighted regression, `loess()`, and `lflm()`. However, the strength of `lflm()` is mainly when, conditioning on one or two variables, there are more explanatory variables in the linear model than here. The example uses data stored in the data frame `ethanol` included in S-PLUS. Below "`>`" indicates the S-PLUS prompt and "`+`" indicates the secondary prompt.

The data are from an experiment with a single-cylinder automobile test engine using ethanol as fuel (Brinkman 1981). The dependent variable `NOx` is the amount of nitric oxide and nitrogen dioxide in the exhaust, normalized by the work done by the engine, and the unit is μg per joule. There are two predictors (i) the compression ratio `C` and (ii) the equivalence ratio `E`, a measure of the air to fuel ratio. There were 88 runs of the experiment.

The purpose of the analysis is to estimate the dependence of the expected value of `NOx` on `C` and `E`, without an a priory assumption of a specific parametric form. Since the convex hull spanned by the observations of the predictors is almost rectangular we will use fitting points spanning a rectangular grid of the predictors. In (Chambers & Hastie 1991, pp. 331-335) it is shown that substantial curvature exists in the direction of the equivalence ratio `E`. For this reason a local quadratic approximation seems appropriate. Using a 50% nearest neighbour bandwidth the surface can be estimated using the command

```
> loess(NOx ~ C*E, span=0.5, degree=2, data=ethanol)
```

this will scale the predictors, by dividing them by their 10% trimmed sample standard deviation (Chambers & Hastie 1991, p. 315). With the exception of the fitting points used this can also be reproduced by use of `lflm()`. The command

```
> eth.surf <- lflm(NOx ~ (1)|(C*E), bt="nn", alpha=0.5,
+ degree=2, scale=sqrt(c(var(ethanol$C),var(ethanol$E))),
+ data=ethanol, n.points=15)
```

will scale by the untrimmed sample standard deviations and store the result as `eth.surf`. The estimated surface can be viewed by issuing the command

```
> plot(eth.surf, pt="wireframe", what="Intercept",
+ zlab="NOx")
```

The plot is displayed in Figure 1. It seems that the “hill” runs parallel to the direction of the compression ratio `C`. As a first step we could then drop the cross-products between `C` and `E` by adding the argument `CP=F` in the call to `lflm`. However we will go directly to a conditional parametric model, in which the surface is linear in `C`, i.e. the expected value of `NOx` is modelled as $\theta_0(\mathtt{E})+\theta_1(\mathtt{E})\mathtt{C}$, where $\theta_0(\cdot)$ and $\theta_1(\cdot)$ are smooth functions. Such a model is fitted by the command

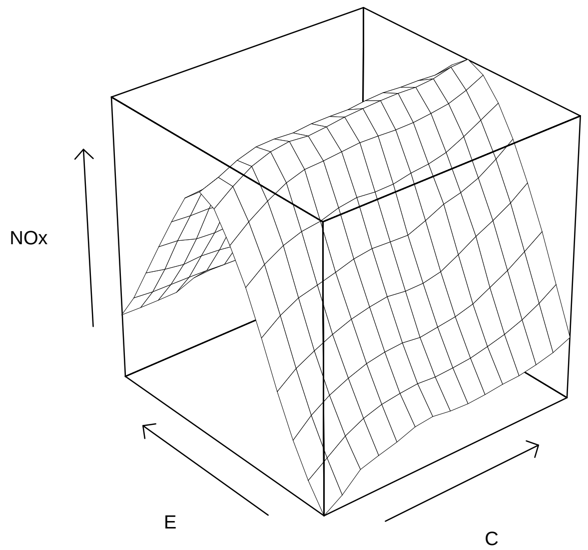


Figure 1: Wire-frame plot of intercept in `eth.surf`.

```
> eth.cpm1 <- lflm(NOx ~ (1+C)|(E), bt="nn", alpha=0.5,
+ degree=2, data=ethanol, n.points=50)
```

Since an intercept term is included by default the formula `NOx ~ (C)|(E)` will be equivalent. As usual, a model without an intercept can be requested by replacing `1` with `-1`. The model just fitted includes third order terms in the local design matrix used. Hence, to mimic

```
> loess(NOx ~ C*E, span=0.5, degree=2, parametric="C",
+ drop.square="C", data=ethanol)
```

which is an example of how a conditional parametric model is specified in `loess()` (Chambers & Hastie 1991, p. 344), the command

```
> eth.cpm2 <- lflm(NOx ~ (C)|(E), bt="nn", alpha=0.5,
+ degree=c(2,1), data=ethanol, n.points=50)
```

should be used. The fit `eth.cpm1` and the coefficient corresponding to `C` in `eth.cpm2` may be plotted by the commands

```
> par(mfrow=c(1,2))
> plot(eth.cpm1, type="l")
> lines(eth.cpm2, what="C", lty=2)
```

The resulting plots are shown in Figure 2 and as expected `eth.cpm2` results in more smooth estimates. Printing the bandwidth component of `eth.cpm1` or `eth.cpm2` reveals that at the leftmost point the bandwidth spans 56% of the axis and at the rightmost point the corresponding number is 43%.

Comparing with (Chambers & Hastie 1991, Section 8.2.2) it is seen that the results are presented quite differently from when `loess()` are used; `loess()` focus on the surface, while `lflm()` focus on the coefficient functions.

A page containing some simple diagnostics is obtained by the commands

```
> par(mfrow=c(2,2))
> plot(fitted(eth.cpm2),residuals(eth.cpm2))
> qqnorm(residuals(eth.cpm2))
> qqline(residuals(eth.cpm2))
> plot(ethanol$C,residuals(eth.cpm2))
> plot(ethanol$E,residuals(eth.cpm2))
```

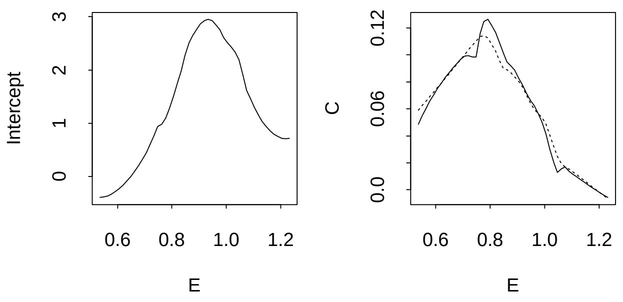


Figure 2: Estimated coefficient functions from `eth.cpm1` (solid) and the coefficient function corresponding to `C` from `eth.cpm2` (dotted).

The plots are shown in Figure 3. The lower left plot indicates that the dependence on `C` is not strictly linear given `E`. Actually, in this case an additive model fitted by use of `gam()` is probably more appropriate, see (Chambers & Hastie 1991, Section 7.2.5).

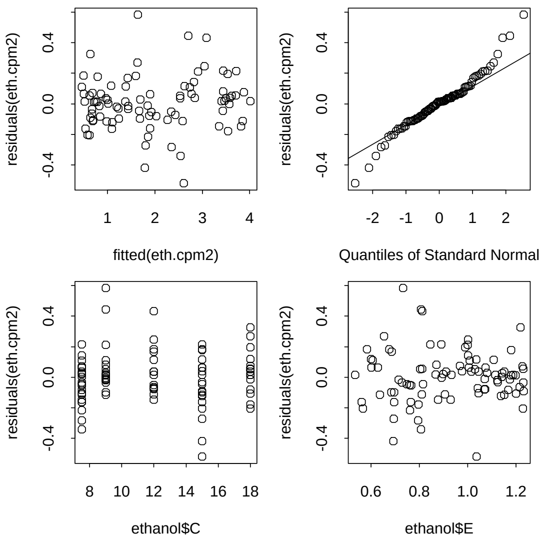


Figure 3: Simple diagnostics for the fit corresponding to `eth.cpm2`.

To gain some understanding of the uncertainty associated with the estimates, bootstrapping of the residuals can be applied (Efron & Tibshirani 1993). In Appendix A.1 a program which will generate 200 bootstrap replicates of `eth.cpm2` calculated at 30 points of equal distance along the

E axis is shown. The program also calculates pointwise estimates of the mean and the standard deviation.

The actual bootstrapping (the for-loop) took 141 seconds on a HP 9000/800. Figure 4 shows the 95% standard normal intervals based on the bootstrap replicates. To interpret these as confidence intervals we need to assume that the model is correct and estimated without bias. As argued above this is somewhat doubtful. The plot was generated by the program shown in Appendix A.2.

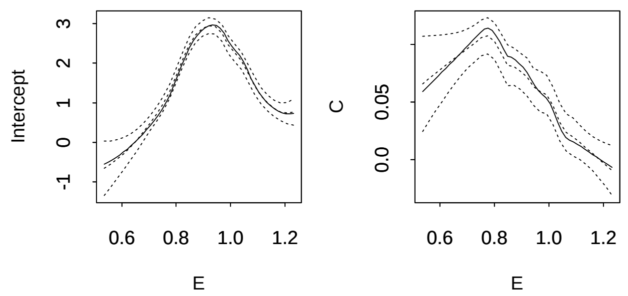


Figure 4: Mean and 95% standard normal intervals based on 200 bootstrap replicates of eth.cpm2 (dotted), together with the original estimates (solid).

5 Obtaining the code and installation

The source code is found at `http://www.imm.dtu.dk/~han/lflm.tgz`

On UNIX systems: Place `lflm.tgz` in a temporary directory. Uncompress the file by executing `gunzip lflm.tgz` and unpack the resulting tape archive file by executing `tar -xvf lflm.tar`. Hereafter; follow the instructions in the file called `README`.

The program is known to compile and run under HP-UX 9 and 10, with S-PLUS 3.4 installed, and under RedHat Linux 4.0 (kernel version 2.0.18) with R 0.50 installed.

6 Conclusion

The conditional parametric model is reviewed, together with estimation using locally weighted fitting of a linear model derived from the original model. Furthermore, a software package for S-PLUS / R is described. The software can also be used for well known methods like kernel regression and locally weighted polynomial regression.

Since the user interface is similar to other S-PLUS / R functions for regression the software is easy to use. Furthermore, to increase speed the core of the program is written in the ANSI-C programming language. The software is flexible enough to allow experimentation with variable bandwidth selection procedures and evaluation structures. Also, the size of the regression problem which can be handled is solely determined by the hardware / operation system configuration.

A Sample S-PLUS programs

A.1 Bootstrapping

The following S-PLUS program was used to generate 200 bootstrap replicates of `eth.cpm2` produced by `lflm()` as shown on page 64. The program also calculates pointwise estimates of the mean and standard deviation.

```
eth.cpm2.resid <- residuals(eth.cpm2)
eth.cpm2.fitted <- fitted(eth.cpm2)
eth.cpm2.boot.1 <- matrix(nrow=30, ncol=200)
eth.cpm2.boot.C <- matrix(nrow=30, ncol=200)
eth.cpm2.boot.E <-
  seq(min(ethanol$E),max(ethanol$E), length=30)

for(b in 1:200) {
  tmp <-
    lflm(NOx ~ (C)|(E), degree=c(2,1), bt="nn", alpha=0.5,
         x.points=data.frame(E=eth.cpm2.boot.E),
         data=data.frame(C=ethanol$C, E=ethanol$E,
```

```
           NOx=eth.cpm2.fitted +
           sample(eth.cpm2.resid, length(eth.cpm2.resid), T)))
  eth.cpm2.boot.1[,b] <- coef(tmp)[, "Intercept"]
  eth.cpm2.boot.C[,b] <- coef(tmp)[, "C"]
}

eth.cpm2.summ <- vector("list", 0)
eth.cpm2.summ$Intercept <-
  data.frame(mean=apply(eth.cpm2.boot.1, 1, "mean"),
             sd=sqrt(apply(eth.cpm2.boot.1, 1, "var")))
eth.cpm2.summ$C <-
  data.frame(mean=apply(eth.cpm2.boot.C, 1, "mean"),
             sd=sqrt(apply(eth.cpm2.boot.C, 1, "var")))
```

A.2 Plotting

To produce the plot of the pointwise 95% confidence intervals shown in Figure 4 on page 66 the following S-PLUS program was used:

```
par(mfrow=c(1, 2))
for(what in c("Intercept", "C")) {
  matplot(eth.cpm2.boot.E,
          cbind(qnorm(p=0.025,
                      mean=get(what, eth.cpm2.summ)$mean,
                      sd=get(what, eth.cpm2.summ)$sd),
                get(what, eth.cpm2.summ)$mean,
                qnorm(p=0.975,
                      mean=get(what, eth.cpm2.summ)$mean,
                      sd=get(what, eth.cpm2.summ)$sd)),
          type="l", lty=2, col=1, xlab="E", ylab=what)
  axis(1)
  axis(2)
  box()
  lines(eth.cpm2, what=what)
}
```

References

Anderson, T. W., Fang, K. T. & Olkin, I., eds (1994), *Multivariate Analysis and Its Applications*, Institute of Mathematical Statistics, Hayward, chapter Coplots, Nonparametric Regression, and conditionally Parametric Fits, pp. 21–36.

Brinkman, N. D. (1981), 'Ethanol fuel—a single-cylinder engine study of efficiency and exhaust emissions', *SAE transactions* **90**(810345), 1410–1424.

Chambers, J. M. & Hastie, T. J., eds (1991), *Statistical Models in S*, Wadsworth, Belmont, CA.

Cleveland, W. S. & Devlin, S. J. (1988), 'Locally weighted regression: An approach to regression analysis by local fitting', *Journal of the American Statistical Association* **83**, 596–610.

Efron, B. & Tibshirani, R. J. (1993), *An Introduction to the Bootstrap*, Chapman & Hall, London/New York.

Härdle, W. (1990), *Applied Nonparametric Regression*, Cambridge University Press, Cambridge, UK.

Hastie, T. & Tibshirani, R. (1990), *Generalized Additive Models*, Chapman & Hall, London/New York.

Hastie, T. & Tibshirani, R. (1993), 'Varying-coefficient models', *Journal of the Royal Statistical Society, Series B, Methodological* **55**, 757–796.

Jørgensen, B. (1993), *The Theory of Linear Models*, Chapman & Hall, London/New York.

Miller, A. J. (1992), '[Algorithm AS 274] Least squares routines to supplement those of Gentleman', *Applied Statistics* **41**, 458–478. (Correction: 94V43 p678).

Statistical Sciences (1993), *S-PLUS Programmer's Manual, Version 3.2*, StatSci, a division of MathSoft, Inc., Seattle.

Paper C

C

Tracking time-varying coefficient-functions

Preliminary accepted for publication in *Int. J. of Adaptive Control and Signal Processing*. A version with more details is available as IMM technical report number 1999-9.

The application to district heating systems mentioned in the introduction of the paper is described in Paper A.

Tracking time-varying coefficient-functions

Henrik Aa. Nielsen[1], Torben S. Nielsen[1], Alfred K. Joensen[1], Henrik Madsen[1], and Jan Holst[2]

Abstract

A method for adaptive and recursive estimation in a class of non-linear autoregressive models with external input is proposed. The model class considered is conditionally parametric ARX-models (CPARX-models), which is conventional ARX-models in which the parameters are replaced by smooth, but otherwise unknown, functions of a low-dimensional input process. These coefficient-functions are estimated adaptively and recursively without specifying a global parametric form, i.e. the method allows for on-line tracking of the coefficient-functions. Essentially, in its most simple form, the method is a combination of recursive least squares with exponential forgetting and local polynomial regression. It is argued, that it is appropriate to let the forgetting factor vary with the value of the external signal which is the argument of the coefficient-functions. Some of the key properties of the modified method are studied by simulation.

Keywords: Adaptive and recursive estimation; Non-linear models; Time-varying functions; Conditional parametric models; Non-parametric method.

1 Introduction

The conditional parametric ARX-model (CPARX-model) is a non-linear model formulated as a linear ARX-model in which the parameters are replaced by smooth, but otherwise unknown, functions of one or more explanatory variables. These functions are called coefficient-functions. In

[1]Department of Mathematical Modelling, Technical University of Denmark, DK-2800 Lyngby, Denmark

[2]Department of Mathematical Statistics, Lund University, Lund Institute of Technology, S-211 00 Lund, Sweden

(Nielsen, Nielsen & Madsen 1997) this class of models is used in relation to district heating systems to model the non-linear dynamic response of network temperature on supply temperature and flow at the plant. A particular feature of district heating systems is, that the response on supply temperature depends on the flow. This is modelled by describing the relation between temperatures by an ARX-model in which the coefficients depend on the flow.

For on-line applications it is advantageous to allow the function estimates to be modified as data become available. Furthermore, because the system may change slowly over time, observations should be down-weighted as they become older. For this reason a time-adaptive and recursive estimation method is proposed. Essentially, the estimates at each time step are the solution to a set of weighted least squares regressions and therefore the estimates are unique under quite general conditions. For this reason the proposed method provides a simple way to perform adaptive and recursive estimation in a class of non-linear models. The method is a combination of the recursive least squares with exponential forgetting (Ljung & Söderström 1983) and locally weighted polynomial regression (Cleveland & Devlin 1988). In the paper *adaptive estimation* is used to denote, that old observations are down-weighted, i.e. in the sense of *adaptive in time*. Some of the key properties of the method are discussed and demonstrated by simulation.

Cleveland & Devlin (1988) gives an excellent account for non-adaptive estimation of a regression function by use of local polynomial approximations. Non-adaptive recursive estimation of a regression function is a related problem, which has been studied recently by Thuvesholmen (1997) using kernel methods and by Vilar-Fernández & Vilar-Fernández (1998) using local polynomial regression. Since these methods are non-adaptive one of the aspects considered in these papers is how to decrease the bandwidth as new observations become available. This problem do not arise for adaptive estimation since old observations are down-weighted and eventually disregarded as part of the algorithm. Hastie & Tibshirani (1993) considered varying-coefficient models which are similar in structure to conditional parametric models and have close resemblance to additive models (Hastie & Tibshirani 1990) with respect to estimation. However, varying-coefficient models include additional assumptions on the structure. Some specific time-series counterparts of these models are the functional-coefficient autoregressive models (Chen & Tsay 1993*a*) and the non-linear additive ARX-models (Chen & Tsay 1993*b*).

The paper is organized as follows. In Section 2 the conditional parametric model is introduced and a procedure for estimation is described. Adaptive and recursive estimation in the model are described in Section 3, which also contains a summary of the method. To illustrate the method some simulated examples are included in Section 4. Further topics, such as optimal bandwidths and optimal forgetting factors are considered in Section 5. Finally, we conclude on the paper in Section 6.

2 Conditional parametric models and local polynomial estimates

When using a conditional parametric model to model the response y_s the explanatory variables are split in two groups. One group of variables $\mathbf{x}_s$ enter globally through coefficients depending on the other group of variables $\mathbf{u}_s$, i.e.

$$y_s = \mathbf{x}_s^T \boldsymbol{\theta}(\mathbf{u}_s) + e_s, \tag{1}$$

where $\boldsymbol{\theta}(\cdot)$ is a vector of coefficient-functions to be estimated and e_s is the noise term. Note that $\mathbf{x}_s$ may contain lagged values of the response. The dimension of $\mathbf{x}_s$ can be quite large, but the dimension of $\mathbf{u}_s$ must be low (1 or 2) for practical purposes (Hastie & Tibshirani 1990, pp. 83-84). In (Nielsen et al. 1997) the dimensions 30 and 1 is used. Estimation in (1), using methods similar to the methods by Cleveland & Devlin (1988), is described for some special cases in (Anderson, Fang & Olkin 1994) and (Hastie & Tibshirani 1993). A more general description can be found in (Nielsen et al. 1997). To make the paper self-contained the method is outlined below.

The functions $\boldsymbol{\theta}(\cdot)$ in (1) are estimated at a number of distinct points by approximating the functions using polynomials and fitting the resulting linear model locally to each of these *fitting points*. To be more specific let $\mathbf{u}$ denote a particular fitting point. Let $\theta_j(\cdot)$ be the j'th element of $\boldsymbol{\theta}(\cdot)$ and let $\mathbf{p}_{d(j)}(\mathbf{u})$ be a column vector of terms in the corresponding d-order polynomial evaluated at $\mathbf{u}$, if for instance $\mathbf{u} = [u_1 \;\, u_2]^T$ then

$\mathbf{p}_2(\mathbf{u}) = [1\ u_1\ u_2\ u_1^2\ u_1u_2\ u_2^2]^T$. Furthermore, let $\mathbf{x}_s = [x_{1,s} \dots x_{p,s}]^T$. With

$$\mathbf{z}_s^T = \left[x_{1,s}\mathbf{p}_{d(1)}^T(\mathbf{u}_s) \dots x_{j,s}\mathbf{p}_{d(j)}^T(\mathbf{u}_s) \dots x_{p,s}\mathbf{p}_{d(p)}^T(\mathbf{u}_s)\right] \tag{2}$$

and

$$\boldsymbol{\phi}_u^T = [\boldsymbol{\phi}_{u,1}^T \dots \boldsymbol{\phi}_{u,j}^T \dots \boldsymbol{\phi}_{u,p}^T], \tag{3}$$

where $\boldsymbol{\phi}_{u,j}$ is a column vector of local coefficients at $\mathbf{u}$ corresponding to $x_{j,s}\mathbf{p}_{d(j)}(\mathbf{u}_s)$. The linear model

$$y_s = \mathbf{z}_s^T \boldsymbol{\phi}_u + e_s; \quad i = 1, \dots, N, \tag{4}$$

is then fitted locally to $\mathbf{u}$ using weighted least squares (WLS), i.e.

$$\hat{\boldsymbol{\phi}}(\mathbf{u}) = \underset{\boldsymbol{\phi}_u}{\operatorname{argmin}} \sum_{s=1}^{N} w_u(\mathbf{u}_s)(y_s - \mathbf{z}_s^T \boldsymbol{\phi}_u)^2, \tag{5}$$

for which a unique closed-form solution exists provided the matrix with rows $\mathbf{z}_s^T$ corresponding to non-zero weights has full rank. The weights are assigned as

$$w_u(\mathbf{u}_s) = W\left(\frac{||\mathbf{u}_s - \mathbf{u}||}{\hbar(\mathbf{u})}\right), \tag{6}$$

where $||\cdot||$ denotes the Euclidean norm, $\hbar(\mathbf{u})$ is the bandwidth used for the particular fitting point, and $W(\cdot)$ is a weight function taking non-negative arguments. Here we follow Cleveland & Devlin (1988) and use

$$W(u) = \begin{cases} (1-u^3)^3, & u \in [0; 1) \\ 0, & u \in [1; \infty) \end{cases} \tag{7}$$

i.e. the weights are between 0 and 1. The elements of $\boldsymbol{\theta}(\mathbf{u})$ are estimated by

$$\hat{\theta}_j(\mathbf{u}) = \mathbf{p}_{d(j)}^T(\mathbf{u})\,\hat{\boldsymbol{\phi}}_j(\mathbf{u}); \quad j = 1, \dots p, \tag{8}$$

where $\hat{\boldsymbol{\phi}}_j(\mathbf{u})$ is the WLS estimate of $\boldsymbol{\phi}_{u,j}$. The estimates of the coefficient-functions obtained as outlined above are called *local polynomial estimates*. For the special case where all coefficient-functions are approximated by constants we use the term local constant estimates.

If $\hbar(\mathbf{u})$ is constant for all values of $\mathbf{u}$ it is denoted a fixed bandwidth. If $\hbar(\mathbf{u})$ is chosen so that a certain fraction α of the observations fulfill $||\mathbf{u}_s - \mathbf{u}|| \leq \hbar(\mathbf{u})$ then α is denoted a nearest neighbour bandwidth. A bandwidth specified according to the nearest neighbour principle is often used as a tool to vary the actual bandwidth with the local density of the data.

Interpolation is used for approximating the estimates of the coefficient-functions for other values of the arguments than the fitting points. This interpolation should only have marginal effect on the estimates. Therefore, it sets requirements on the number and placement of the fitting points. If a nearest neighbour bandwidth is used it is reasonable to select the fitting points according to the density of the data as it is done when using *k-d* trees (Chambers & Hastie 1991, Section 8.4.2). However, in this paper the approach is to select the fitting points on an equidistant grid and ensure that several fitting points are within the (smallest) bandwidth so that linear interpolation can be applied safely.

3 Adaptive estimation

As pointed out in the previous section local polynomial estimation can be viewed as local constant estimation in a model derived from the original model. This observation forms the basis of the method suggested. For simplicity the adaptive estimation method is described as a generalization of exponential forgetting. However, the more general forgetting methods described by Ljung & Söderström (1983) could also serve as a basis.

3.1 The proposed method

Using exponential forgetting and assuming observations at time $s = 1, \ldots, t$ are available, the adaptive least squares estimate of the parameters $\boldsymbol{\phi}$ relating the explanatory variables $\mathbf{z}_s$ to the response y_s using the linear model $y_s = \mathbf{z}_s^T \boldsymbol{\phi} + e_s$ is found as

$$\hat{\boldsymbol{\phi}}_t = \underset{\boldsymbol{\phi}}{\operatorname{argmin}} \sum_{s=1}^{t} \lambda^{t-s} (y_s - \mathbf{z}_s^T \boldsymbol{\phi})^2, \tag{9}$$

where $0 < \lambda < 1$ is called the forgetting factor, see also (Ljung & Söderström 1983). The estimate can be seen as a local constant approximation in the direction of time. This suggests that the estimator may also be defined locally with respect to some other explanatory variables $\mathbf{u}_t$. If the estimates are defined locally to a fitting point $\mathbf{u}$, the adaptive estimate corresponding to this point can be expressed as

$$\hat{\boldsymbol{\phi}}_t(\mathbf{u}) = \underset{\boldsymbol{\phi}_u}{\operatorname{argmin}} \sum_{s=1}^{t} \lambda^{t-s} w_u(\mathbf{u}_s)(y_s - \mathbf{z}_s^T \boldsymbol{\phi}_u)^2, \tag{10}$$

where $w_u(\mathbf{u}_s)$ is a weight on observation s depending on the fitting point $\mathbf{u}$ and $\mathbf{u}_s$, see Section 2.

In Section 3.2 it will be shown how the estimator (10) can be formulated recursively, but here we will briefly comment on the estimator and its relations to non-parametric regression. A special case is obtained if $\mathbf{z}_s = 1$ for all s, then simple calculations show that

$$\hat{\phi}_t(\mathbf{u}) = \frac{\sum_{s=1}^{t} \lambda^{t-s} w_u(\mathbf{u}_s) y_s}{\sum_{s=1}^{t} \lambda^{t-s} w_u(\mathbf{u}_s)}, \tag{11}$$

and for $\lambda = 1$ this is a kernel estimator of $\phi(\cdot)$ in $y_s = \phi(\mathbf{u}_s) + e_s$, cf. (Härdle 1990, p. 30). For this reason (11) is called an adaptive kernel estimator of $\phi(\cdot)$ and the estimator (10) may be called an adaptive local constant estimator of the coefficient-functions $\boldsymbol{\phi}(\cdot)$ in the conditional parametric model $y_s = \mathbf{z}_s^T \boldsymbol{\phi}(\mathbf{u}_s) + e_s$. Using the same techniques as in Section 2 this can be used to implement adaptive local polynomial estimation in models like (1).

3.2 Recursive formulation

Following the same arguments as in Ljung & Söderström (1983) it is readily shown that the adaptive estimates (10) can be found recursively as

$$\hat{\boldsymbol{\phi}}_t(\mathbf{u}) = \hat{\boldsymbol{\phi}}_{t-1}(\mathbf{u}) + w_u(\mathbf{u}_t) \mathbf{R}_{u,t}^{-1} \mathbf{z}_t \left[y_t - \mathbf{z}_t^T \hat{\boldsymbol{\phi}}_{t-1}(\mathbf{u}) \right] \tag{12}$$

and

$$\mathbf{R}_{u,t} = \lambda \mathbf{R}_{u,t-1} + w_u(\mathbf{u}_t) \mathbf{z}_t \mathbf{z}_t^T \ . \tag{13}$$

It is seen that existing numerical procedures implementing adaptive recursive least squares for linear models can be applied, by replacing $\mathbf{z}_t$ and y_t in the existing procedures with $\mathbf{z}_t\sqrt{w_u(\mathbf{u}_t)}$ and $y_t\sqrt{w_u(\mathbf{u}_t)}$, respectively. Note that $\mathbf{z}_t^T\hat{\boldsymbol{\phi}}_{t-1}(\mathbf{u})$ is a predictor of y_t locally with respect to $\mathbf{u}$ and for this reason it is used in (12). To predict y_t a predictor like $\mathbf{z}_t^T\hat{\boldsymbol{\phi}}_{t-1}(\mathbf{u}_t)$ is appropriate.

3.3 Modified updating formula

When $\mathbf{u}_t$ is far from the particular fitting point $\mathbf{u}$ it is clear from (12) and (13) that $\hat{\boldsymbol{\phi}}_t(\mathbf{u}) \approx \hat{\boldsymbol{\phi}}_{t-1}(\mathbf{u})$ and $\mathbf{R}_{u,t} \approx \lambda\mathbf{R}_{u,t-1}$, i.e. old observations are down-weighted without new information becoming available. This may result in abruptly changing estimates if $\mathbf{u}$ is not visited regularly, since the matrix $\mathbf{R}$ is decreasing exponentially in this case. Hence it is proposed to modify (13) to ensure that the past is weighted down only when new information becomes available, i.e.

$$\mathbf{R}_{u,t} = \lambda v(w_u(\mathbf{u}_t);\lambda)\mathbf{R}_{u,t-1} + w_u(\mathbf{u}_t)\mathbf{z}_t\mathbf{z}_t^T, \tag{14}$$

where $v(\cdot\,;\lambda)$ is a nowhere increasing function on $[0;1]$ fulfilling $v(0;\lambda) = 1/\lambda$ and $v(1;\lambda) = 1$. Note that this requires that the weights span the interval ranging from zero to one. This is fulfilled for weights generated as described in Section 2. In this paper we consider only the linear function $v(w;\lambda) = 1/\lambda - (1/\lambda - 1)w$, for which (14) becomes

$$\mathbf{R}_{u,t} = (1-(1-\lambda)w_u(\mathbf{u}_t))\mathbf{R}_{u,t-1} + \ w_u(\mathbf{u}_t)\mathbf{z}_t\mathbf{z}_t^T. \tag{15}$$

It is reasonable to denote

$$\lambda_{eff}^u(t) = 1-(1-\lambda)w_u(\mathbf{u}_t) \tag{16}$$

the *effective forgetting factor* for point $\mathbf{u}$ at time t.

When using (14) or (15) it is ensured that $\mathbf{R}_{u,t}$ can not become singular because the process $\{\mathbf{u}_t\}$ moves away from the fitting point for a longer period. However, the process $\{\mathbf{z}_t\}$ should be persistently excited as for linear ARX-models. In this case, given the weights, the estimates define a global minimum corresponding to (10).

3.4 Nearest neighbour bandwidth

Assume that $\mathbf{u}_t$ is a stochastic variable and that the pdf $f(\cdot)$ of $\mathbf{u}_t$ is known and constant over t. Based on a nearest neighbour bandwidth the actual bandwidth can then be calculated for a number of fitting points $\mathbf{u}$ placed within the domain of $f(\cdot)$ and used to generate the weights $w_u(\mathbf{u}_t)$. The actual bandwidth $\hbar(\mathbf{u})$ corresponding to the point $\mathbf{u}$ will be related to the nearest neighbour bandwidth α by

$$\alpha = \int_{\mathbb{D}_u} f(\boldsymbol{\nu}) d\boldsymbol{\nu}, \tag{17}$$

where $\mathbb{D}_u = \{\boldsymbol{\nu} \in \mathbb{R}^d \mid ||\boldsymbol{\nu} - \mathbf{u}|| \leq \hbar(\mathbf{u})\}$ is the neighbour-hood, d is the dimension of $\mathbf{u}$, and $||\cdot||$ is the Euclidean norm. In applications the density $f(\cdot)$ is often unknown. However, $f(\cdot)$ can be estimated from data, e.g. by the empirical pdf.

3.5 Effective number of observations

In order to select an appropriate value for α the effective number of observations used for estimation must be considered. In Appendix A it is shown that under certain conditions, when the modified updating (15) is used,

$$\tilde{\eta}_u = \frac{1}{1 - E[\lambda^u_{eff}(t)]} = \frac{1}{(1-\lambda)E[w_u(\mathbf{u}_t)]} \tag{18}$$

is a lower bound on the effective number of observations (in the direction of time) corresponding to a fitting point $\mathbf{u}$. Generally (18) can be considered an approximation. When selecting α and λ it is then natural to require that the number of observations within the bandwidth, i.e. $\alpha\tilde{\eta}_u$, is sufficiently large to justify the complexity of the model and the order of the local polynomial approximations.

As an example consider $u_t \sim N(0,1)$ and $\lambda = 0.99$ where the effective number of observations within the bandwidth, $\alpha\tilde{\eta}_u$, is displayed in Figure 1. It is seen that $\alpha\tilde{\eta}_u$ depends strongly on the fitting point u but only moderately on α. When investigating the dependence of $\alpha\tilde{\eta}_u$ on λ and α it turns out that $\alpha\tilde{\eta}_u$ is almost solely determined by λ. In conclusion, for the example considered, the effective forgetting factor $\lambda^u_{eff}(t)$ will be affected by the nearest neighbour bandwidth, so that the effective number of observations within the bandwidth will be strongly dependent on λ, but only weakly dependent on the bandwidth (α). The ratio between the rate at which the weights on observations goes to zero in the direction of time and the corresponding rate in the direction of u_t will be determined by α.

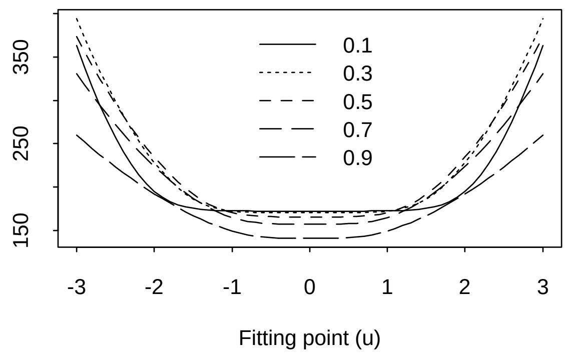


Figure 1: Effective number of observations within the bandwidth ($\alpha\tilde{\eta}_u(u)$) for $\alpha = 0.1, \ldots, 0.9$ and $\lambda = 0.99$.

As it is illustrated by Figure 1 the effective number of observations behind each of the local approximations depends on the fitting point. This is contrary to the non-adaptive nearest neighbour method, cf. Section 2, and may result in a somewhat unexpected behaviour of the estimates. If the system follows a linear ARX-model and if the coefficients of the system are estimated as coefficient-functions then both adaptive and non-adaptive nearest neighbour approaches will be unbiased. However, for this example the variance of local constant estimates will decrease for increasing values of $|u|$. This is verified by simulations, which also show that local linear and quadratic approximations results in increased variance for large $|u|$. Note that, when the true function is not a constant, the local constant approximation may result in excess bias, see e.g. (Nielsen et al. 1997).

If λ is varied with the fitting point as $\lambda(\mathbf{u}) = 1 - 1/(T_0 E[w_u(\mathbf{u}_t)])$ then $\tilde{\eta}_u = T_0$. Thus, the effective number of observations within the bandwidth is constant across fitting points. Furthermore, T_0 can be interpreted as the memory time constant. To avoid highly variable estimates of $E[w_u(\mathbf{u}_t)]$ in the tails of the distribution of $\mathbf{u}_t$ the estimates should be based on a parametric family of distributions. However, in the remaining part of this paper λ is not varied across fitting points.

3.6 Summary of the method

To clarify the method the actual algorithm is briefly described in this section. It is assumed that at each time step t measurements of the output y_t and the two sets of inputs $\mathbf{x}_t$ and $\mathbf{u}_t$ are received. The aim is to obtain adaptive estimates of the coefficient-functions in the non-linear model (1).

Besides λ in (15), prior to the application of the algorithm a number of fitting points $\mathbf{u}^{(i)}$; $i = 1, \ldots, n_{fp}$ in which the coefficient-functions are to be estimated has to be selected. Furthermore the bandwidth associated with each of the fitting points $\hbar^{(i)}$; $i = 1, \ldots, n_{fp}$ and the degrees of the approximating polynomials $d(j)$; $j = 1, \ldots, p$ have to be selected for each of the p coefficient-functions. For simplicity the degree of the approximating polynomial for a particular coefficient-function will be fixed across fitting points. Finally, initial estimates of the coefficient-functions in the model corresponding to local constant estimates, i.e. $\hat{\boldsymbol{\phi}}_0(\mathbf{u}^{(i)})$, must be chosen. Also, the matrices $\mathbf{R}_{u^{(i)},0}$ must be chosen. One possibility is $\text{diag}(\epsilon, \ldots, \epsilon)$, where ϵ is a small positive number.

In the following description of the algorithm it will be assumed that $\mathbf{R}_{u^{(i)},t}$ is non-singular for all fitting points. In practice we would just stop updating the estimates if the matrix become singular. Under the assumption mentioned the algorithm can be described as:

For each time step t: Loop over the fitting points $\mathbf{u}^{(i)}$; $i = 1, \ldots, n_{fp}$ and for each fitting point:

- Construct the explanatory variables corresponding to local constant estimates using (2):
  $\mathbf{z}_t^T = [x_{1,t}\mathbf{p}_{d(1)}^T(\mathbf{u}_t) \ldots x_{p,t}\mathbf{p}_{d(p)}^T(\mathbf{u}_t)]$.
- Calculate the weight using (6) and (7):
  $w_{u^{(i)}}(\mathbf{u}_t) = (1 - (||\mathbf{u}_t - \mathbf{u}^{(i)}||/\hbar^{(i)})^3)^3$, if $||\mathbf{u}_t - \mathbf{u}^{(i)}|| < \hbar^{(i)}$ and zero

otherwise.

- Find the effective forgetting factor using (16):
  $\lambda_{eff}^{(i)}(t) = 1 - (1-\lambda) w_{u^{(i)}}(\mathbf{u}_t)$.

- Update $\mathbf{R}_{u^{(i)},t-1}$ using (15):
  $\mathbf{R}_{u^{(i)},t} = \lambda_{eff}^{(i)}(t)\mathbf{R}_{u^{(i)},t-1} + w_{u^{(i)}}(\mathbf{u}_t)\mathbf{z}_t\mathbf{z}_t^T$.

- Update $\hat{\boldsymbol{\phi}}_{t-1}(\mathbf{u}^{(i)})$ using (12):
  $\hat{\boldsymbol{\phi}}_t(\mathbf{u}^{(i)}) = \hat{\boldsymbol{\phi}}_{t-1}(\mathbf{u}^{(i)}) + w_{u^{(i)}}(\mathbf{u}_t)\mathbf{R}_{u^{(i)},t}^{-1}\mathbf{z}_t \left[y_t - \mathbf{z}_t^T \hat{\boldsymbol{\phi}}_{t-1}(\mathbf{u}^{(i)})\right]$.

- Calculate the updated local polynomial estimates of the coefficient-functions using (8):
  $\hat{\theta}_{jt}(\mathbf{u}^{(i)}) = \mathbf{p}_{d(j)}^T(\mathbf{u}^{(i)})\, \hat{\boldsymbol{\phi}}_{j,t}(\mathbf{u}^{(i)}); \quad j = 1, \dots p$

The algorithm could also be implemented using the matrix inversion lemma as in (Ljung & Söderström 1983).

4 Simulations

Aspects of the proposed method are illustrated in this section. When the modified updating formula (15) is used the general behaviour of the method for different bandwidths is illustrated in Section 4.1. In Section 4.2 results obtained using the two updating formulas (13) and (15) are compared.

The simulations are performed using the non-linear model

$$y_t = a(t, u_{t-1})y_{t-1} + b(t, u_{t-1})x_t + e_t, \tag{19}$$

where $\{x_t\}$ is the input process, $\{u_t\}$ is the process controlling the coefficients, $\{y_t\}$ is the output process, and $\{e_t\}$ is a white noise standard Gaussian process. The coefficient-functions are simulated as

$$a(t, u) = 0.3 + (0.6 - \frac{1.5}{N}t) \exp\left(-\frac{(u - \frac{0.8}{N}t)^2}{2(0.6 - \frac{0.1}{N}t)^2}\right)$$

and

$$b(t,u) = 2 - \exp\left(-\frac{(u+1-\frac{2}{N}t)^2}{0.32},\right)$$

where $t = 1, \ldots, N$ and $N = 5000$, i.e. $a(t,u)$ ranges from -0.6 to 0.9 and $b(t,u)$ ranges from 1 to 2. The functions are displayed in Figure 2. As indicated by the figure both coefficient-functions are based on a Gaussian density in which the mean and variance varies linearly with time.

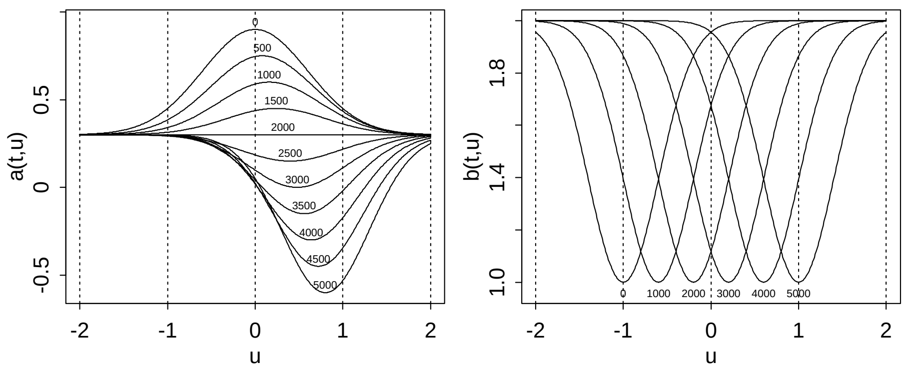


Figure 2: The time-varying coefficient-functions plotted for equidistant points in time as indicated on the plots.

Local linear adaptive estimates of the functions $a()$ and $b()$ are then found using the proposed procedure with the model

$$y_t = a(u_{t-1})y_{t-1} + b(u_{t-1})x_t + e_t. \tag{20}$$

In all cases initial estimates of the coefficient-functions are set to zero and during the initialization the estimates are not updated, for the fitting point considered, until ten observations have received a weight of 0.5 or larger.

4.1 Highly correlated input processes

In the simulation presented in this section a strongly correlated $\{\mathbf{u}_t\}$ process is used and also the $\{\mathbf{x}_t\}$ process is quite strongly correlated. This allows us to illustrate various aspects of the method. For less correlated series the performance is much improved. The data are generated using (19) where $\{x_t\}$ and $\{u_t\}$ are zero mean $AR(1)$-processes with poles in

0.9 and 0.98, respectively. The variance for both series is one and the series are mutually independent. In Figure 3 the data are displayed. Based on these data adaptive estimation in (20) are performed using nearest neighbour bandwidths, calculated assuming a standard Gaussian distribution for u_t.

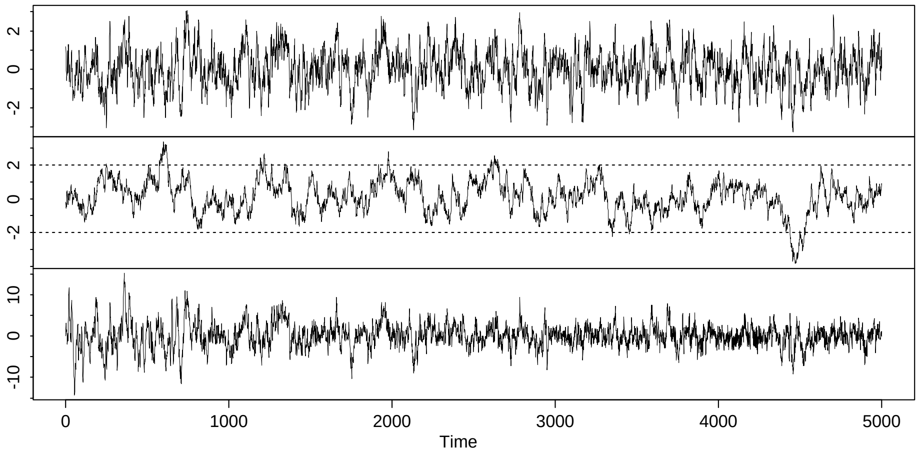


Figure 3: Simulated output (bottom) when x_t (top) and u_t (middle) are $AR(1)$-processes.

The results obtained using the modified updating formula (15) are displayed for fitting points $u = -2, -1, 0, 1, 2$ in Figures 4 and 5. For the first 2/3 of the period the estimates at $u = -2$, i.e. $\hat{a}(-2)$ and $\hat{b}(-2)$, only gets updated occasionally. This is due to the correlation structure of $\{u_t\}$ as illustrated by the realization displayed in Figure 3.

For both estimates the bias is most pronounced during periods in which the true coefficient-function changes quickly for values of u_t near the fitting point considered. This is further illustrated by the true functions in Figure 2 and it is, for instance clear that adaption to $a(t, 1)$ is difficult for $t > 3000$. Furthermore, $u = 1$ is rarely visited by $\{u_t\}$ for $t > 3000$, see Figure 3. In general, the low bandwidth ($\alpha = 0.3$) seems to result in large bias, presumably because the effective forgetting factor is increased on average, cf. Section 3.5. Similarly, the high bandwidth ($\alpha = 0.7$) result in large bias for $u = 2$ and $t > 4000$. A nearest neighbour bandwidth of 0.7 corresponds to an actual bandwidth of approximately 2.5 at $u = 2$ and since most values of u_t are below one, it is clear that the estimates at $u = 2$ will be highly influenced by the actual function values for u

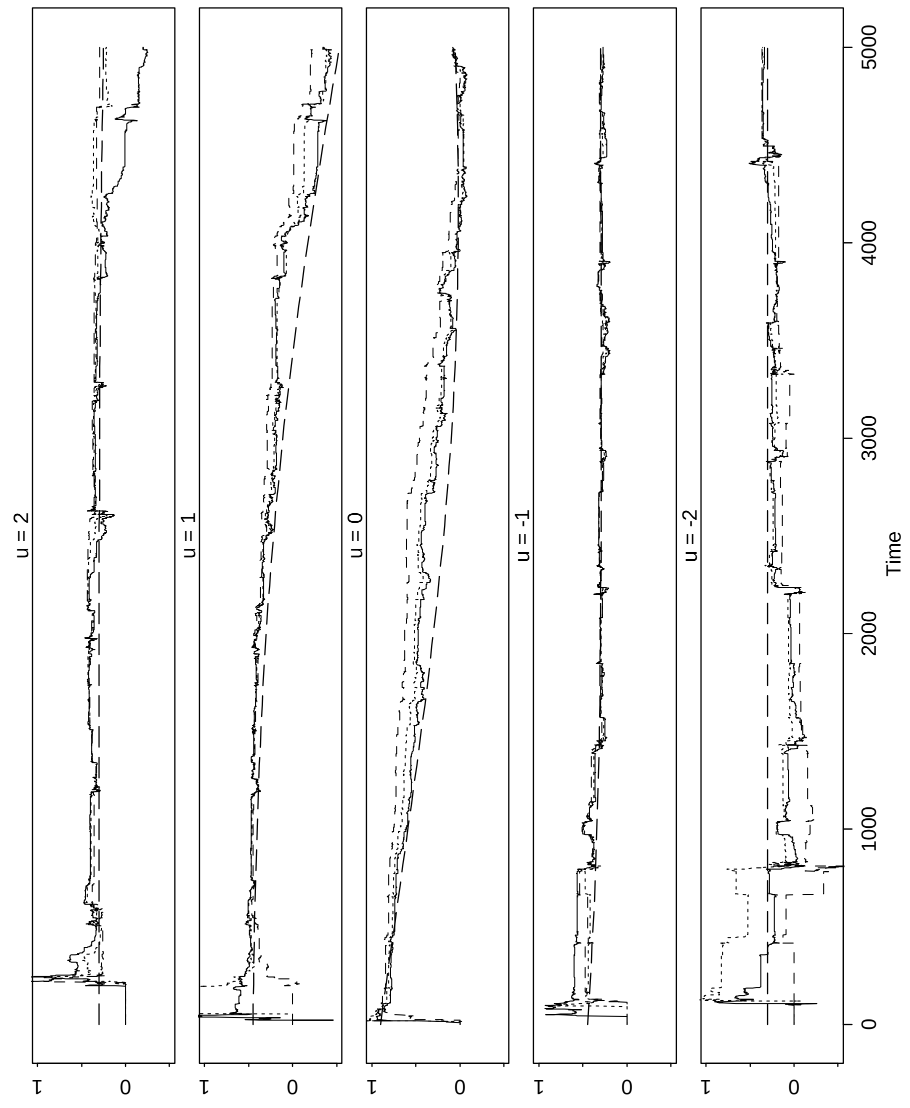


Figure 4: Adaptive estimates of $a(u)$ using local linear approximations and nearest neighbour bandwidths 0.3 (dashed), 0.5 (dotted), and 0.7 (solid). True values are indicated by smooth dashed lines.

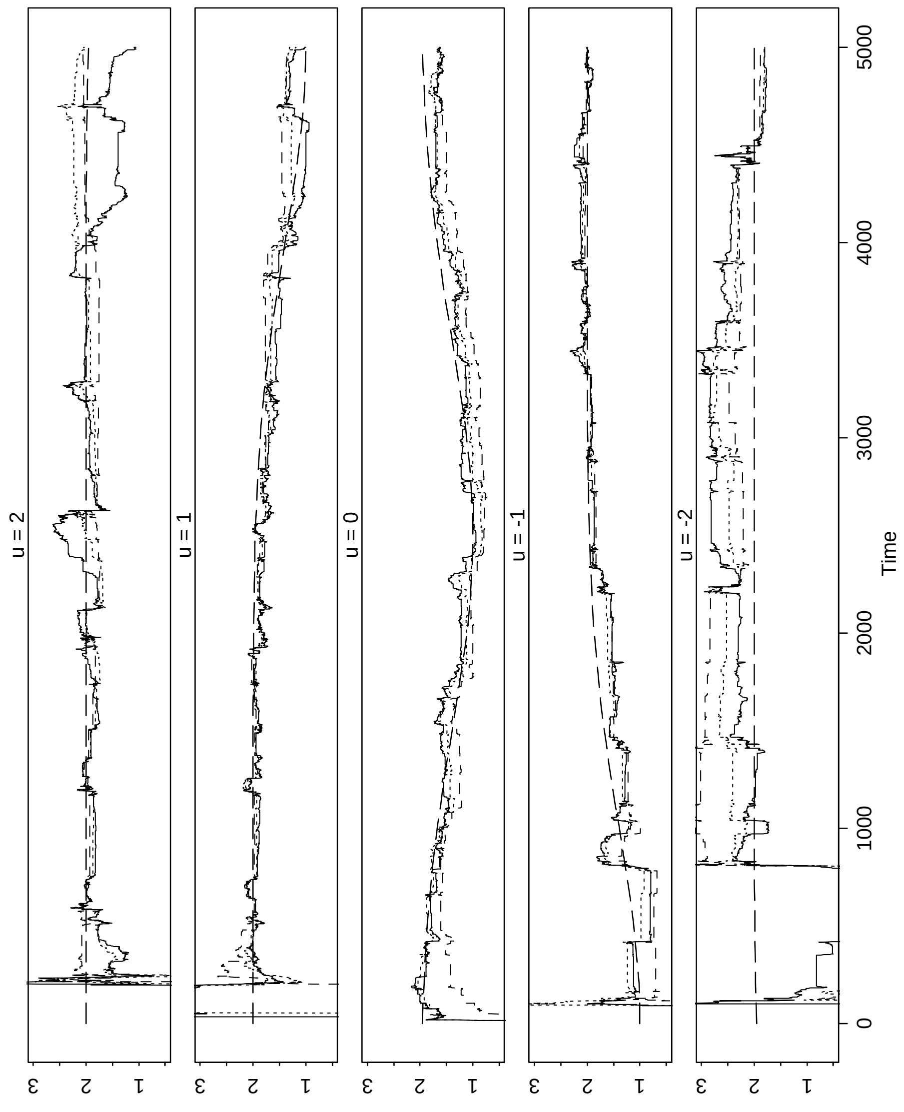


Figure 5: Adaptive estimates of $b(u)$ using local linear approximations and nearest neighbour bandwidths 0.3 (dashed), 0.5 (dotted), and 0.7 (solid). True values are indicated by smooth dashed lines.

near one. From Figure 2 it is seen that for $t > 4000$ the true values at $u = 1$ is markedly lower that the true values at $u = 2$. Together with the fact that $u = 2$ is not visited by $\{u_t\}$ for $t > 4000$ this explains the observed bias at $u = 2$, see Figure 6.

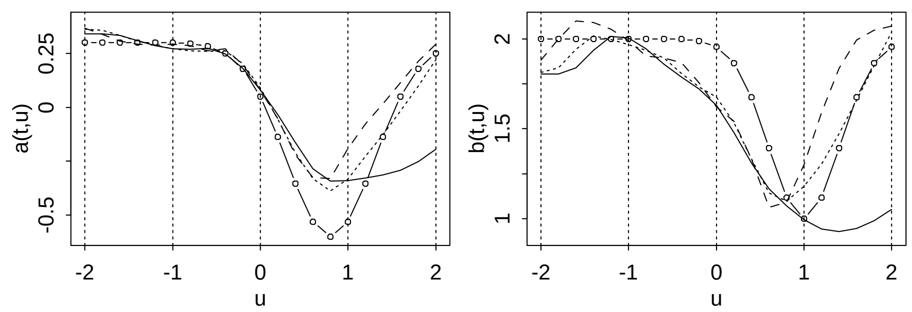


Figure 6: Adaptive estimates for the example considered in Section 4.1 at $t = 5000$ for $\alpha = 0.3$ (dashed), 0.5 (dotted), 0.7 (solid). True values are indicated by circles and fitting points ranging from -2 to 2 in steps of 0.2 are used.

4.2 Abrupt changes in input signals

One of the main advantages of the modified updating formula (15) over the normal updating formula (13) is that it does not allow fast changes in the estimates at fitting points which has not been visited by the process $\{u_t\}$ for a longer period. If, for instance, we wish to adaptively estimate the stationary relation between the heat consumption of a town and the ambient air temperature then $\{u_t\}$ contains an annual fluctuation and at some geographical locations the transition from, say, warm to cold periods may be quite fast. In such a situation the normal updating formula (13) will, essentially, forget the preceding winter during the summer, allowing for large changes in the estimate at low temperatures during some initial period of the following winter. Actually, it is possible that, using the normal updating formula will result in a nearly singular $\mathbf{R}_t$.

To illustrate this aspect 5000 observations are simulated using the model (19). The sequence $\{x_t\}$ is simulated as a standard Gaussian $AR(1)$-process with a pole in 0.9. Furthermore, $\{u_t\}$ is simulated as an iid process where

$$u_t \sim \begin{cases} N(0,1), & t = 1, \ldots, 1000 \\ N(3/2, 1/6^2), & t = 1001, \ldots, 4000 \\ N(-3/2, 1/6^2), & t = 4001, \ldots, 5000 \end{cases}$$

To compare the two methods of updating, i.e. (13) and (15), a fixed λ is used in (15) across the fitting points and the effective forgetting factors are designed to be equal. If $\tilde{\lambda}$ is the forgetting factor corresponding to (13) it can be varied with u as

$$\tilde{\lambda}(u) = E[\lambda^u_{eff}(t)] = 1 - (1 - \lambda)E[w_u(u_t)],$$

where $E[w_u(u_t)]$ is calculated assuming that u_t is standard Gaussian, i.e. corresponding to $1 \leq t \leq 1000$. A nearest neighbour bandwidth of 0.5 and $\lambda = 0.99$ are used, which results in $\tilde{\lambda}(0) = 0.997$ and $\tilde{\lambda}(\pm 2) = 0.9978$.

The corresponding adaptive estimates obtained for the fitting point $u = -1$ are shown in Figure 7. The figure illustrates that for both methods the updating of the estimates stops as $\{u_t\}$ leaves the fitting point $u = -1$. Using the normal updating (13) of $\mathbf{R}_t$ its value is multiplied by $\tilde{\lambda}(-1)^{3000} \approx 0.00015$ as $\{u_t\}$ returns to the vicinity of the fitting point. This results in large fluctuations of the estimates, starting at $t = 4001$. As opposed to this, the modified updating (15) does not lead to such fluctuations after $t = 4000$.

5 Further topics

Optimal bandwidth and forgetting factor: So far in this paper it has been assumed that the bandwidths used over the range of $\mathbf{u}_t$ is derived from the nearest neighbour bandwidth α and it has been indicated how it can be ensured that the average forgetting factor is large enough.

However, the adaptive and recursive method is well suited for forward validation (Hjorth 1994) and hence tuning parameters can be selected by minimizing, e.g. the root mean square of the one-step prediction error (using observed $\mathbf{u}_t$ and $\mathbf{x}_t$ to predict y_t, together with interpolation between fitting points to obtain $\hat{\boldsymbol{\theta}}_{t-1}(\mathbf{u}_t)$).

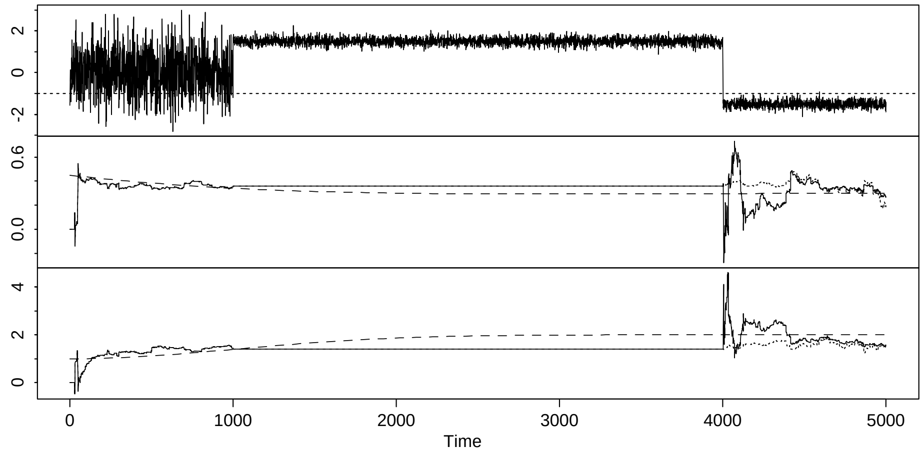


Figure 7: Realization of $\{u_t\}$ (top) and adaptive estimates of $a(-1)$ (middle) and $b(-1)$ (bottom), using the normal updating formula (solid) and the modified updating formula (dotted). True values are indicated by dashed lines.

There are numerous ways to define the tuning parameters. A simple approach is to use (λ, α), cf. (15) and (17). A more ambiguous approach is to use both λ and $\hbar$ for each fitting point $\mathbf{u}$. Furthermore, tuning parameters controlling scaling and rotation of $\mathbf{u}_s$ and the degree of the local polynomial approximations may also be considered.

If n fitting points are used this amounts to $2n$, or more, tuning parameters. To make the dimension of the (global) optimization problem independent of n and to have $\lambda(\mathbf{u})$ and $\hbar(\mathbf{u})$ vary smoothly with $\mathbf{u}$ we may choose to restrict $\lambda(\mathbf{u})$ and $\hbar(\mathbf{u})$, or appropriate transformations of these (logit for λ and log for $\hbar$), to follow a spline basis (de Boor 1978, Lancaster & Salkauskas 1986). This is similar to the smoothing of spans described by Friedman (1984).

Local time-polynomials: In this paper local polynomial approximations in the direction of time is not considered. Such a method is proposed for usual ARX-models by Joensen, Nielsen, Nielsen & Madsen (1999). This method can be combined with the method described here and will result in local polynomial approximations where cross-products between time and the conditioning variables ($\mathbf{u}_t$) are excluded.

6 Conclusion and discussion

In this paper methods for adaptive and recursive estimation in a class of non-linear autoregressive models with external input are proposed. The model class considered is conditionally parametric ARX-models (CPARX-model), which is a conventional ARX-model in which the parameters are replaced by smooth, but otherwise unknown, functions of a low-dimensional input process. These functions are estimated adaptively and recursively without specifying a global parametric form. One possible application of CPARX-models is the modelling of varying time delays, cf. (Nielsen et al. 1997).

The methods can be seen as generalizations or combinations of recursive least squares with exponential forgetting (Ljung & Söderström 1983), local polynomial regression (Cleveland & Devlin 1988), and conditional parametric fits (Anderson et al. 1994). Hence, the methods constitutes an extension to the notion of local polynomial estimation. The so called modified method is suggested for cases where the process controlling the coefficients are highly correlated or exhibit seasonal behaviour. The estimates at each time step can be seen as solutions to a range of weighted least squares regressions and therefore the solution is unique for well behaved input processes. A particular feature of the modified method is that the effective number of observations behind the estimates will be almost independent of the actual bandwidth. This is accomplished by varying the effective forgetting factor with the bandwidth. The bandwidth mainly controls the rate at which the weights corresponding to exponential forgetting goes to zero relatively to the rate at which the remaining weights goes to zero.

For some applications it may be possible to specify global polynomial approximations to the coefficient-functions of a CPARX-model. In this situation the adaptive recursive least squares method can be applied for tracking the parameters defining the coefficient-functions for all values of the input process. However, if the argument(s) of the coefficient-functions only stays in parts of the space corresponding to the possible values of the argument(s) for longer periods this may seriously affect the estimates of the coefficient-functions for other values of the argument(s), as it corresponds to extrapolation using a fitted polynomial. This problem is effectively solved using the conditional parametric model in combination with the modified updating formula.

A Effective number of observations

Using the modified updating formula, as described in Section 3.3, the estimates at time t can be written as

$$\hat{\boldsymbol{\phi}}_t(\mathbf{u}) = \operatorname*{argmin}_{\boldsymbol{\phi}_u} \sum_{s=1}^{t} \beta(t,s) w_u(\mathbf{u}_s)(y_s - \mathbf{z}_s^T \boldsymbol{\phi}_u)^2,$$

where

$$\beta(t,t) = 1,$$

and, for $s < t$

$$\beta(t,s) = \prod_{j=s+1}^{t} \lambda_{eff}^u(j) = \lambda_{eff}^u(t)\beta(t-1,s),$$

where $\lambda_{eff}^u(t)$ is given by (16). It is then obvious to define the effective number of observations (in the direction of time) as

$$\eta_u(t) = \sum_{i=0}^{\infty} \beta(t,t-i) = 1 + \lambda_{eff}^u(t) + \lambda_{eff}^u(t)\lambda_{eff}^u(t-1) + \ldots \quad \text{(A.1)}$$

Suppose that the fitting point $\mathbf{u}$ is chosen so that $E[\eta_u(t)]$ exists. Consequently, when $\{\lambda_{eff}^u(t)\}$ is i.i.d. and when $\bar{\lambda}_u \in [0,1)$ denotes $E[\lambda_{eff}^u(t)]$, the average effective number of observations is

$$\bar{\eta}_u = 1 + \bar{\lambda}_u + \bar{\lambda}_u^2 + \ldots = \frac{1}{1-\bar{\lambda}_u}.$$

When $\{\lambda_{eff}^u(t)\}$ is not i.i.d., it is noted that since the expectation operator is linear, $E[\eta_u(t)]$ is the sum of the expected values of each summand in (A.1). Hence, $E[\eta_u(t)]$ is independent of t if $\{\lambda_{eff}^u(t)\}$ is strongly stationary, i.e. if $\{\mathbf{u}_t\}$ is strongly stationary. From (A.1)

$$\eta_u(t) = 1 + \lambda_{eff}^u(t)\eta_u(t-1) \quad \text{(A.2)}$$

is obtained, and from the definition of covariance it then follows, that

$$\bar{\eta}_u = \frac{1 + Cov[\lambda_{eff}^u(t), \eta_u(t-1)]}{1-\bar{\lambda}_u} \geq \frac{1}{1-\bar{\lambda}_u}, \quad \text{(A.3)}$$

since $0 < \lambda < 1$ and assuming, that the covariance between $\lambda_{eff}^u(t)$ and $\eta_u(t-1)$ is positive. Note that, if the process $\{\mathbf{u}_t\}$ behaves such that if it has been near $\mathbf{u}$ for a longer period up to time $t-1$ it will tend to be near $\mathbf{u}$ at time t also, a positive covariance is obtained. It is the experience of the authors that such a behaviour of a stochastic process is often encountered in practice.

As an alternative to the calculations above $\lambda_{eff}^{u}(t)\eta_u(t-1)$ may be linearized around $\bar{\lambda}_u$ and $\bar{\eta}_u$. From this it follows, that if the variances of $\lambda_{eff}^{u}(t)$ and $\eta_u(t-1)$ are small then

$$\bar{\eta}_u \approx \frac{1}{1-\bar{\lambda}_u}.$$

Therefore we may use $1/(1-\bar{\lambda}_u)$ as an approximation to the effective number of observations, and in many practical applications it will be an lower bound, c.f. (A.3). By assuming a stochastic process for $\{\mathbf{u}_t\}$ the process $\{\eta_u(t)\}$ can be simulated using (A.2) whereby the validity of the approximation can be addressed.

References

Anderson, T. W., Fang, K. T. & Olkin, I., eds (1994), *Multivariate Analysis and Its Applications*, Institute of Mathematical Statistics, Hayward, chapter Coplots, Nonparametric Regression, and conditionally Parametric Fits, pp. 21–36.

Chambers, J. M. & Hastie, T. J., eds (1991), *Statistical Models in S*, Wadsworth, Belmont, CA.

Chen, R. & Tsay, R. S. (1993*a*), 'Functional-coefficient autoregressive models', *Journal of the American Statistical Association* **88**, 298–308.

Chen, R. & Tsay, R. S. (1993*b*), 'Nonlinear additive ARX models', *Journal of the American Statistical Association* **88**, 955–967.

Cleveland, W. S. & Devlin, S. J. (1988), 'Locally weighted regression: An approach to regression analysis by local fitting', *Journal of the American Statistical Association* **83**, 596–610.

de Boor, C. (1978), *A Practical Guide to Splines*, Springer Verlag, Berlin.

Friedman, J. H. (1984), A variable span smoother, Technical Report 5, Laboratory for Computational Statistics, Dept. of Statistics, Stanford Univ., California.

Härdle, W. (1990), *Applied Nonparametric Regression*, Cambridge University Press, Cambridge, UK.

Hastie, T. J. & Tibshirani, R. J. (1990), *Generalized Additive Models*, Chapman & Hall, London/New York.

Hastie, T. & Tibshirani, R. (1993), 'Varying-coefficient models', *Journal of the Royal Statistical Society, Series B, Methodological* **55**, 757–796.

Hjorth, J. S. U. (1994), *Computer Intensive Statistical Methods: Validation Model Selection and Bootstrap*, Chapman & Hall, London/New York.

Joensen, A. K., Nielsen, H. A., Nielsen, T. S. & Madsen, H. (1999), 'Tracking time-varying parameters with local regression', *Automatica* . To be published.

Lancaster, P. & Salkauskas, K. (1986), *Curve and Surface Fitting: An Introduction*, Academic Press, New York/London.

Ljung, L. & Söderström, T. (1983), *Theory and Practice of Recursive Identification*, MIT Press, Cambridge, MA.

Nielsen, H. A., Nielsen, T. S. & Madsen, H. (1997), ARX-models with parameter variations estimated by local fitting, *in* Y. Sawaragi & S. Sagara, eds, '11th IFAC Symposium on System Identification', Vol. 2, pp. 475–480.

Thuvesholmen, M. (1997), 'An on-line crossvalidation bandwidth selector for recursive kernel regression', Lic thesis, Department of Mathematical Statistics, Lund University, Sweden.

Vilar-Fernández, J. A. & Vilar-Fernández, J. M. (1998), 'Recursive estimation of regression functions by local polynomial fitting', *Annals of the Institute of Statistical Mathematics* **50**, 729–754.

Paper D

Model output statistics applied to wind power prediction

D

Originally published as

> A. K. Joensen, G. Giebel, L. Landberg, H. Madsen, and H. Aa. Nielsen. Model output statistics applied to wind power prediction. In *Wind Energy for the Next Millenium*, European Wind Energy Conference, pages 1177–1180, Nice, France, 1–5 March 1999.

Model output statistics applied to wind power prediction

Alfred Karsten Joensen[1,2], Gregor Giebel[1], Lars Landberg[1],
Henrik Madsen[2] Henrik Aalborg Nielsen[2]

Abstract

Being able to predict the output of a wind farm online for a day or two in advance has significant advantages for utilities, such as better possibility to schedule fossil fuelled power plants and a better position on electricity spot markets.
In this paper prediction methods based on Numerical Weather Prediction (NWP) models are considered. The spatial resolution used in NWP models implies that these predictions are not valid locally at a specific wind farm, furthermore, due to the non-stationary nature and complexity of the processes in the atmosphere, and occasional changes of NWP models, the deviation between the predicted and the measured wind will be time dependent. If observational data is available, and if the deviation between the predictions and the observations exhibits systematic behaviour, this should be corrected for; if statistical methods are used, this approach is usually referred to as MOS (Model Output Statistics). The influence of atmospheric turbulence intensity, topography, prediction horizon length and auto-correlation of wind speed and power is considered, and to take the time-variations into account, adaptive estimation methods are applied.
Three estimation techniques are considered and compared, Extended Kalman Filtering, recursive least squares and a new modified recursive least squares algorithm.

Keywords: Forecasting Methods; Wind Energy; Statistical Analysis; Performance

[1]Department of Wind Energy and Atmospheric Physics, Risø National Laboratory, DK-4000 Roskilde, Denmark
[2]Department of Mathematical Modelling, Technical University of Denmark, DK-2800 Lyngby, Denmark

1 Introduction

Several models for predicting the output from wind farms have already been developed, some based on observations from the wind farms (Madsen 1996), others based on numerical weather predictions (Landberg 1997), and again others on combination of both (Joensen, Madsen & Nielsen 1997).

This paper describes how statistical methods, usually referred to as model output statistics (MOS), can be used in models that combine observations and NWP model predictions, and the approach taken here is slightly different from the approach in (Joensen et al. 1997). The NWP model, HIRLAM (Machenhauer 1988), is run by the Danish Meteorological Institute (DMI). The observations, wind speed w_t and power p_t, are from four sites in Denmark: The Risø mast at Risø National Laboratory, and the Avedøre, Kappel and Østermarie wind farms.

The NWP model predicts several meteorological variables, such as temperature, surface fluxes and pressure, wind speed ω_t and direction θ_t at 31 levels/heights, see (Machenhauer 1988) for definition of the levels, and at the surface, i.e. 10 m a.g.l. The NWP model is run four times at day, at 00:00, 06:00, 12:00 and 18:00 UTC, and the predictions are given in 3 hourly steps 36 hours ahead.

2 Finding the right NWP model level

In (Landberg 1997) it was found that the NWP predicted wind from level 27 gave the best results when used as input to the neutral geostrophic drag law to determine u_*, and the neutral logarithmic profile was used to calculate the wind at hub height. It was concluded that the reason why the stability dependant relations did not improve the results, was that the heat fluxes were not predicted accurately enough. Since HIRLAM has been updated several times since the investigation in (Landberg 1997), it is reasonable to re-evaluate these results.

Based on the results in (Landberg 1997) the neutral relations are used to transform the NWP wind down to the surface, and the prediction performance is compared to the performance of the surface wind calculated by the NWP model. HIRLAM takes the stability into account when the surface wind is calculated, and this comparison will therefore show if it is advantageous to include the stability. Furthermore, in order to make a fair comparison, the predictions should be corrected for any bias and offset, i.e. the simple MOS model

$$w_{t+k} = a_k \omega_{t+k} + b_k + \epsilon_{t+k} \tag{1}$$

where ϵ_{t+k} is assumed to be white noise and k is the prediction horizon, is fitted to the observations using the least squares method.

Observations from 44, 76 and 125 m above the surface from the Risø mast are used in the comparison, because the wind which should be used might not be the same depending on which height above the surface it is compared to.

To evaluate the performance of the predictions

$$\rho = \frac{VAR(w_{t+k}) - MSE_k}{VAR(w_{t+k})} \tag{2}$$

is used, where VAR is the estimated variance of the observations and MSE_k is the mean square error of the predictions k hours ahead. The interpretation of ρ is that it measures how much of the total variation in the observations is explained by the predictions, i.e. a value of 1 means that the predictions are perfect and 0 means that predictions are useless.

Figure 1 shows the results for each model height and each prediction horizon compared to the 44 m observations at the Risø mast. It is clearly seen that for all prediction horizons the best result is obtained using the surface wind, and the corresponding figures for the 76 and 125 m show the same results. The conclusion is therefore that it is advantageous to take the stability into account, and hence the surface wind calculated by the NWP model should be used.

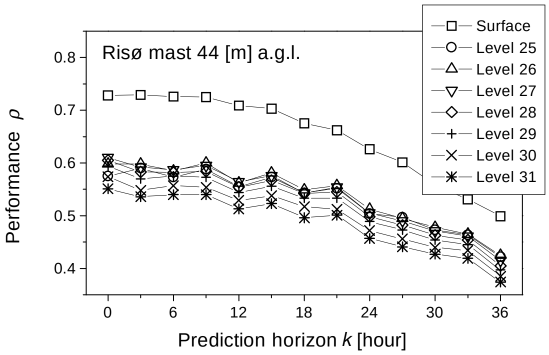


Figure 1: Performance for various NWP model levels

3 Wind direction dependency

Due to the spatial resolution of any NWP model it should be expected that some kind of wind direction dependant fine tuning to a specific site should be possible. One way to do this fine tuning is to apply a MOS model

$$w_{t+k} = a_k(\theta_{t+k})\omega_{t+k} + b_k(\theta_{t+k}) + \epsilon_{t+k} \tag{3}$$

From a physical point of view the adjustment due to the topography should be a wind direction dependant factor, but this model also includes a wind direction dependant offset. The reason for this is purely statistical, e.g. if the prediction accuracy of the NWP model is not the same for all directions the inclusion of the offset will increase the performance.

Local regression (Hastie & Tibshirani 1993) has been used to estimate the coefficient functions in (3). When using this method it has been assumed that for a given wind direction sector the coefficient functions are well approximated by second order polynomials. Using the terminology of local regression, the nearest neighbour bandwidth was chosen to include 40% of the observations at each fitting point. A physical way to take the topography into account is to perform a high resolution analysis of the site and the surroundings, and use this analysis to correct the NWP wind for local topography effects, which obviously are not included by the NWP model resolution. To see if this is advantageous the NWP surface wind has been corrected by matrixes calculated by WAsP (Mortensen, Landberg, Troen & Petersen 1993).

Again Risø mast data from the last half of 1997 and first half of 1998 has been used for the estimation, while validation was performed with data from the last half of 1998. In order to make a fair comparison, the MOS model (1) has been applied after the WAsP correction, and the performance has also been calculated for the MOS model applied to the raw NWP surface predictions.

Surprisingly, Figure 2 shows that the performance of the WAsP corrected forecast is worse than without, although the difference is only minor. The reason is most likely that the physical assumptions behind WASP are not satisfied when WAsP is used for predictions of wind speed and direction which contain errors. Nevertheless, it seems as if there is some dependency on the topography, because the statistical correction (3) improves the performance, but this is not the only reason. This follows from the fact that the wind speed distribution depends on the wind direction, i.e. the wind speed in Denmark is usually higher when coming from west compared to e.g. north or east. This is a feature of the overall flow, and can not be prescribed to the local topography. Because the wind speed is not perfectly predicted by the NWP model this is incorporated by the wind direction dependent offset in (3).

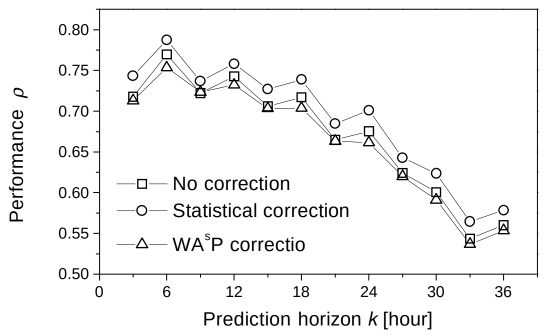


Figure 2: Performance for direction dependent models

4 Diurnal variation

Surprisingly, as seen in Figure 2, it seems easier to predict e.g. 6 hours ahead than 3 hours ahead. The reason is that the prediction horizon k aliases with the time of day, and therefore, as shown in Figure 3, with the diurnal variation in the wind speed/atmospheric stability. This is because the NWP model is update 4 times a day, hence odd prediction horizons correspond to the following times of day: 03:00, 09:00, 15:00 and 21:00, and even horizons correspond to: 00:00, 06:00, 12:00 and 18:00.

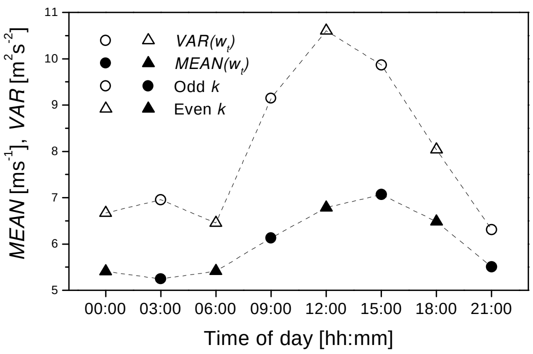


Figure 3: Diurnal variation

Furthermore, Figure 3 is only based on measurements from the 1998 summer period, for latitudes like those of Denmark, there is also an annual variation in the diurnal variation (Nielsen, Joensen, Madsen, Landberg & Giebel 1999). When the data used for the estimation is not from the same seasons of the year as the data used for the validation, which is the case in Section 3, the parameters of (3) become biased towards the diurnal variation in that specific period. This is the main reason for the effect seen in Figure 2.

5 Adaptive estimation

In the previous sections only wind speed models have been considered, we now turn to wind power models which are slightly more complicated due to the non-linear relation between the wind speed and the power.

Furthermore, due to the non-stationary nature of the atmosphere, it must be expected that the parameters of a MOS model will be time-varying. Hence adaptive estimation methods are considered. Two widely used methods for this purpose is Kalman filtering and recursive least squares, see (Ljung & Söderström 1983) and the references therein. The key idea behind these methods is the same, which is to discard old information as new becomes available, or to be more specific, the methods slide a time-window of a specific width over the observations, where only the newest observations are seen. This approach has its drawbacks and advantages. If the true system is non-stationary and if this non-stationarity is not described by the model, the approach implies that the model adapts to the current state of the underlying system. But, on the other hand, because less observations are used to determine the parameters of the model, the parameters might become poorly determined, resulting in large parameter and prediction variance. The optimal model choice is therefore a model which balances simplicity and flexibility.

5.1 Extended Kalman filter

One way to simplify the model for predicting the power is, to in some way, include a known relation between wind speed and direction and power, i.e. the power curve, in the model that is to be estimated adaptively. One solution is to apply the model

$$p_{t+k} = a_k pow(b_k \omega_{t+k} + c_k, \theta_{t+k}) + d_k p_t + l_k + \epsilon_{t+k} \tag{4}$$

where $pow(\cdot,\cdot)$ is the wind farm power curve derived by the PARK application (Sanderhoff 1993). See (Landberg 1998) for a similar analysis. The reason for the scaling of the NWP wind speed inside the power curve comes from the observation that the ratio between the measured wind speed and the NWP wind speed is different from one and time dependent. The constant inside the power curve lets the estimation determine the cut-in and cut-out wind speeds. The observed power at time t is include because for short prediction horizon the power observations are auto-correlated (Nielsen et al. 1999), this is also the reason for including the scaling of the power curve, because for short horizons more emphasis will be on the auto-correlation, i.e. p_t, and for larger horizons more weight will be put on the NWP and hence the power curve. Because this model is not linear in the parameters, the Extended Kalman Filter has been used for the estimation in this model.

5.2 Recursive least squares

A way to avoid the non-linearity is to use a polynomial approximation of the power curve, i.e.

$$p_{t+k} = a_k \omega_{t+k} + b_k \omega_{t+k}^2 + c_k \omega_{t+k}^3 + d_k p_t + l_k + \epsilon_{t+k} \tag{5}$$

This model has the same number of parameters as model (4), but it does not incorporate any knowledge about the wind farm power curve, apart from the fact that most power curves are very well approximated by a third order polynomial in the wind speed. The parameters of this model have been estimated using the standard recursive least squares algorithm.

5.3 Recursive local regression

So far we have not mentioned how the parameters in the MOS models depend on the prediction horizon. Actually we have just estimated one set of parameters for each prediction horizon. Addressing the variance problem of the Kalman Filter and the usual recursive least squares algorithm, it might be advantageous to make assumptions about how the parameters depend on k. Furthermore, in Section 3 and 4 it was shown that the NWP model predicted wind direction improved the performance for the wind speed predictions, and that there was an aliasing effect with the time of day/atmospheric stability, caused by the update frequency of the NWP model. To take all these findings into account the following model is proposed

$$\begin{aligned} p_{t+k} = {} & a(k, \theta_{t+k}, t_{day})\omega_{t+k} + b(k, \theta_{t+k}, t_{day})\omega_{t+k}^2 + \\ & c(k, \theta_{t+k}, t_{day})\omega_{t+k}^3 + d(k, \theta_{t+k}, t_{day})p_t + \\ & m(k, \theta_{t+k}, t_{day}) + \epsilon_{t+k} \end{aligned} \tag{6}$$

This model is similar in structure to (5) except that the parameters/ coefficients now are assumed to be functions of the prediction horizon, the wind direction and the time of day.

To take the stability into account it was found sufficient to estimate two sets of coefficient functions, one set for the following times of day: 00:00, 03:00, 06:00, and 21:00, i.e. mainly neutral/stable conditions, and one set for the times: 09:00, 12:00, 15:00, 18:00, i.e. mainly unstable conditions. For the wind direction and the prediction horizon, the approach described in (Nielsen, Nielsen, Joensen, Madsen & Holst 1998) have been used for the estimation of the coefficient functions. This approach is best described as recursive local regression, and it is an extension of the usual recursive least squares algorithm, where the functional shape is found by estimating the parameters locally over a grid spanning the variables, e.g. for a given wind direction θ in the grid, only observations close to this direction are used when the value of the coefficient function for this particular value of θ is estimated.

In the actual estimation the coefficient functions were estimated in a fine grid spanning the NWP model predicted wind direction, using a fixed bandwidth of 100 Deg, and for the prediction horizon an increasing bandwidth was used, i.e. for the 3 hour prediction a bandwidth spanning only the 3 hour prediction was used, increasing to a bandwidth spanning the 12 hour up to the 36 hour prediction for the 36 hour prediction, this choice reflects the fact the variation of the parameters with the prediction horizon was found to be small for large prediction horizons.

When only one set of coefficient functions were estimated for all times of day, the assumption about the variation of the coefficient functions with k failed, because in this case a 6 hourly variation is introduced in the coefficient functions with k.

Because some wind directions are rare it was found important to use a different degree of time adaptation depending on the wind direction (Nielsen et al. 1998). For frequent wind directions the optimal time window was found to be about 2-3 months, while for rare wind direction it was not to use any adaptation at all. This indicates that the variation of the coefficient functions with the wind direction is larger than the time-variation.

6 Results

Figure 4 shows the performance for the three adaptive approaches that have been described in the previous sections. It clearly seen that model (6) gives the best results, the non-linear model (4) and the linear model (5) are close in performance, neither model performs best on all prediction horizons, but overall the linear model seems to perform best. This suggests that the polynomial approximation of the power curve is adequate.

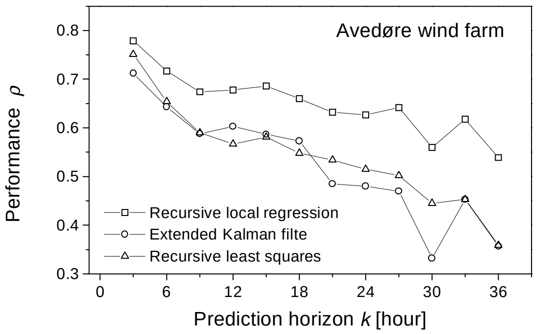


Figure 4: Performance of adaptive approaches

Figure 5 shows the prediction performance of model (6) for the three wind farms, and it is seen that there is a pronounced variation in the performance for the individual wind farms, which can be due to many factors, e.g. the NWP model accuracy depends on the specific location, or another factor that might be of importance in this study is that the quality of the observations from the wind farms was rather poor, about 30 % of the observations were missing.

7 Summary

In this paper various MOS approaches have been proposed for wind power prediction models, which are based on numerical weather predictions and observations.

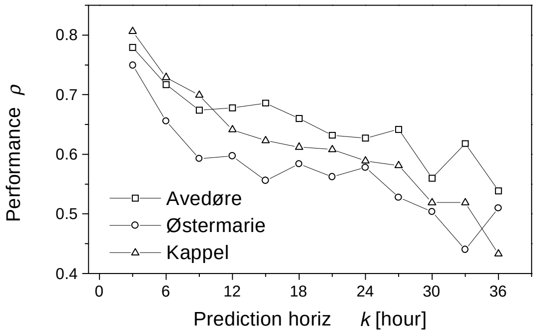


Figure 5: Performance for various wind farms

Three estimation methods have been considered: Extended Kalman filtering, recursive least squares and a new modified recursive least squares algorithm. The results indicate that the best MOS approach, is one which takes the wind direction, the time of day, the prediction horizon, and auto-correlation of the observations into account when using a wind speed polynomial approximation of the power curve to predict the future power from a wind farm. Furthermore it was found that the surface wind from the NWP model, which in this case is HIRLAM (Machenhauer 1988), should be used.

8 Acknowledgements

This work is partially funded by the European Commission (EC) under JOULE (JOR3-CT95-0008). A. Joensen is partly funded by the Danish Research Academy. G. Giebel is funded by EC Training through Research (JOR3-CT97-5004).

References

Hastie, T. & Tibshirani, R. (1993), 'Varying-coefficient models', *Journal of the Royal Statistical Society, Series B, Methodological* **55**, 757–796.

Joensen, A. K., Madsen, H. & Nielsen, T. S. (1997), Non-parametric statistical methods for wind power prediction, *in* 'Proceedings of the EWEC97', Ireland, pp. 788–792.

Landberg, L. (1997), 'Short-term prediction of local wind conditions', *Boundary-Layer Meteorology* **70**, 171–180.

Landberg, L. (1998), 'A mathematical look at a physical power prediction model', *Wind Energy* **1**, 23–30.

Ljung, L. & Söderström, T. (1983), *Theory and Practice of Recursive Identification*, MIT Press, Cambridge, MA.

Machenhauer, B., ed. (1988), *HIRLAM Final Report*, Danish Meteorological Institute, Copenhagen, Denmark.

Madsen, H., ed. (1996), *Models and Methods for Predicting Wind Power*, Department of Mathematical Modelling, Technical University of Denmark, Denmark.

Mortensen, N. G., Landberg, L., Troen, I. & Petersen, E. L. (1993), 'Wind atlas analysis and application program (WAsP)', RisøNational Laboratory, Roskilde, Denmark.

Nielsen, H. A., Nielsen, T. S., Joensen, A. K., Madsen, H. & Holst, J. (1998), '*Tracking time-varying coefficient-functions*'. To be published.

Nielsen, T. S., Joensen, A., Madsen, H., Landberg, L. & Giebel, G. (1999), 'A new reference for predicting wind power', *Wind Energy* **1**, 29–34.

Sanderhoff, P. (1993), 'PARK - User's Guide. A PC-program for calculation of wind turbine park performance', RisøNational Laboratory, Roskilde, Denmark. Risø-I-668(EN).

Paper E

Tracking time-varying parameters with local regression

E

Accepted for publication in *Automatica*.

Tracking time-varying parameters with local regression

Alfred Joensen[1,2], Henrik Madsen[1],
Henrik Aa. Nielsen[1], and Torben S. Nielsen[1]

Abstract

This paper shows that the recursive least squares (RLS) algorithm with forgetting factor is a special case of a varying-coefficient model, and a model which can easily be estimated via simple local regression. This observation allows us to formulate a new method which retains the RLS algorithm, but extends the algorithm by including polynomial approximations. Simulation results are provided, which indicates that this new method is superior to the classical RLS method, if the parameter variations are smooth.

Keywords: Recursive estimation; varying-coefficient; conditional parametric; polynomial approximation; weighting functions.

1 Introduction

The *RLS* algorithm with forgetting factor (Ljung & Söderström 1983) is often applied in on-line situations, where time variations are not modeled adequately by a linear model. By sliding a time-window of a specific width over the observations where only the newest observations are seen, the model is able to adapt to slow variations in the dynamics. The width, or the bandwidth $\hbar$, of the time-window determines how fast the model adapts to the variations, and the most adequate value of $\hbar$ depends on how fast the parameters actually vary in time. If the time variations are fast, $\hbar$ should be small, otherwise the estimates will be seriously biased. However, fast adaption means that only few observations are used for the estimation, which results in a noisy estimate. Therefore the choice of $\hbar$ can be seen as a bias/variance trade off.

[1]Department of Mathematical Modelling, Technical University of Denmark, DK-2800 Lyngby, Denmark

[2]Department of Wind Energy and Atmospheric Physics, Risø National Laboratory, DK-4000 Roskilde, Denmark

In the context of local regression (Cleveland & Devlin 1988) the parameters of a linear model estimated by the *RLS* algorithm can be interpreted as zero order local time polynomials, or in other words local constants. However, it is well known that polynomials of higher order in many cases provide better approximations than local constants. The objective of this paper is thus to illustrate the similarity between the *RLS* algorithm and local regression, which leads to a natural extension of the *RLS* algorithm, where the parameters are approximated by higher order local time polynomials. This approach does, to some degree, represent a solution to the bias/variance trade off. Furthermore, viewing the *RLS* algorithm as local regression, could potentially lead to development of new and refined *RLS* algorithms, as local regression is an area of current and extensive research. A generalisation of models with varying parameters is presented in (Hastie & Tibshirani 1993), and, as will be shown in this paper, the *RLS* algorithm is an estimation method for one of these models.

Several extensions of the RLS algorithm have been proposed in the literature, especially to handle situations where the parameter variations are not the same for all the parameters. Such situations can be handled by assigning individual bandwidths to each parameter, e.g. *vector forgetting*, or by using the *Kalman Filter* (Parkum, Poulsen & Holst 1992). These approaches all have drawbacks, such as assumptions that the parameters are uncorrelated and/or are described by a random walk. Polynomial approximations and local regression can to some degree take care of these situations, by approximating the parameters with polynomials of different degrees. Furthermore, it is obvious that the parameters can be functions of other variables than time. In (Nielsen, Nielsen, Madsen & Joensen 1999) a recursive algorithm is proposed, which can be used when the parameters are functions of time and some other explanatory variables.

Local regression is adequate when the parameters are functions of the same explanatory variables. If the parameters depend on individual explanatory variables, estimation methods for additive models should be used (Fan, Hardle & Mammen 1998, Hastie & Tibshirani 1990). Unfortunately it is not obvious how to formulate recursive versions of these estimation methods, and to the authors best knowledge no such recursive methods exists. Early work on additive models and recursive regression dates back to (Holt 1957) and (Winters 1960), which developed recursive estimation methods for models related to the additive models, where individual forgetting factors are assigned to each additive component, and the trend is approximated by a polynomial in time.

2 The varying-coefficient approach

Varying-coefficient models are considered in (Hastie & Tibshirani 1993). These models can be considered as linear regression models in which the parameters are replaced by smooth functions of some explanatory variables. This section gives a short introduction to the varying-coefficient approach and a method of estimation, local regression, which becomes the background for the proposed extension of the *RLS* algorithm.

2.1 The model

We define the varying-coefficient model

$$y_i = \boldsymbol{z}_i^T \boldsymbol{\theta}(\boldsymbol{x}_i) + e_i; \quad i = 1, \ldots, N, \tag{1}$$

where y_i is a response, $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$ are explanatory variables, $\boldsymbol{\theta}(\cdot)$ is a vector of unknown but smooth functions with values in $\mathbf{R}$, and N is the number of observations. If ordinary regression is considered e_i should be identically distributed (i.d.), but if i denotes at time index and $\boldsymbol{z}_i^T$ contains lagged values of the response variable, e_i should be independent and identically distributed (i.i.d).

The definition of a varying-coefficient model in (Hastie & Tibshirani 1993) is somewhat different than the one given by Eq. 1, in the way that the individual parameters in $\boldsymbol{\theta}(\cdot)$ depend on individual explanatory variables. In (Anderson, Fang & Olkin 1994), the model given by Eq. 1 is denoted a conditional parametric model, because when $\boldsymbol{x}_i$ is constant the model reduces to an ordinary linear model

2.2 Local constant estimates

As only models where the parameters are functions of time are considered, only $\boldsymbol{x}_i = i$ is considered in the following. Estimation in Eq. 1 aims at estimating the functions $\boldsymbol{\theta}(\cdot)$, which in this case are the one-dimensional functions $\boldsymbol{\theta}(i)$. The functions are estimated only for distinct values of the argument t. Let t denote such a point and $\hat{\boldsymbol{\theta}}(t)$ the estimated coefficient functions, when the coefficients are evaluated at t.

One solution to the estimation problem is to replace $\boldsymbol{\theta}(i)$ in Eq. 1 with a constant vector $\boldsymbol{\theta}(i) = \boldsymbol{\theta}$ and fit the resulting model locally to t, using weighted least squares, i.e.

$$\hat{\boldsymbol{\theta}}(t) = \mathbf{arg}\min_{\theta} \sum_{i=1}^{t} w_i(t)(y_i - \boldsymbol{z}_i^T \boldsymbol{\theta})^2. \tag{2}$$

Generally, using a nowhere increasing weight function $W : \mathbf{R}_0 \rightarrow \mathbf{R}_0$ and a spherical kernel the actual weight $w_i(t)$ allocated to the ith observation is determined by the Euclidean distance, in this case $|i - t|$, as

$$w_i(t) = W\left(\frac{|i-t|}{\hbar(t)}\right). \tag{3}$$

The scalar $\hbar(t)$ is called the bandwidth, and determines the size of the neighbourhood that is spanned by the weight function. If e.g. $\hbar(t)$ is constant for all values of t it is denoted a fixed bandwidth. In practice, however, also the nearest neighbour bandwidth, which depends on the distribution of the explanatory variable, is used (Cleveland & Devlin 1988). Although, in this case where $\boldsymbol{x}_i = i$, i.e. the distribution of the explanatory variable is rectangular, a fixed bandwidth and a nearest neighbour bandwidth are equivalent.

2.3 Local polynomial estimation

If the bandwidth $\hbar(t)$ is sufficiently small the approximation of $\boldsymbol{\theta}(t)$ as a constant vector near t is good. This implies, however, that a relatively low number of observations is used to estimate $\boldsymbol{\theta}(t)$, resulting in a noisy estimate. On the contrary a large bias may appear if the bandwidth is large.

It is, however, obvious that locally to t the elements of $\boldsymbol{\theta}(t)$ may be better approximated by polynomials, and in many cases polynomials will provide good approximations for larger bandwidths than local constants. Local polynomial approximations are easily included in the method described. Let $\theta_j(t)$ be the jth element of $\boldsymbol{\theta}(t)$ and let $\boldsymbol{p}_d(t)$ be a column vector of terms in a d-order polynomial evaluated at t, i.e. $\boldsymbol{p}_d(t) = [t^d \; t^{d-1} \; \cdots \; 1]$. Furthermore, introduce $\boldsymbol{z}_i = [z_{1i} \; \cdots \; z_{pi}]^T$,

$$\boldsymbol{u}_{i,t}^T = \left[z_{1i}\boldsymbol{p}_{d_1}^T(t-i) \; \cdots \; z_{ji}\boldsymbol{p}_{d_j}^T(t-i) \; \cdots \; z_{pi}\boldsymbol{p}_{d_p}^T(t-i) \right], \tag{4}$$

$$\hat{\boldsymbol{\phi}}^T(t) = [\hat{\boldsymbol{\phi}}_1^T(t) \; \cdots \; \hat{\boldsymbol{\phi}}_j^T(t) \; \cdots \; \hat{\boldsymbol{\phi}}_p^T(t)], \tag{5}$$

where $\hat{\boldsymbol{\phi}}_j(t)$ is a column vector of local constant estimates at t, i.e.

$$\hat{\boldsymbol{\phi}}_j^T(t) = [\hat{\phi}_{jd_j+1}(t) \; \cdots \; \hat{\phi}_{j1}(t)] \tag{6}$$

corresponding to $z_{ji}\boldsymbol{p}_{d_j}^T(t-i)$. Now weighted least squares estimation is applied as described in Section 2.2, but fitting the linear model

$$y_i = \boldsymbol{u}_{i,t}^T\boldsymbol{\phi} + e_i; \quad i = 1, \ldots, t, \tag{7}$$

locally to t, i.e. the estimate $\hat{\boldsymbol{\phi}}(t)$ of the parameters $\boldsymbol{\phi}$ in Eq. 7 becomes a function of t as a consequence of the weighting. Estimates of the elements of $\boldsymbol{\theta}(t)$ can now be obtained as

$$\hat{\theta}_j(t) = \boldsymbol{p}_{d_j}^T(0)\hat{\boldsymbol{\phi}}_j(t) = [\underbrace{0 \; \cdots \; 0 \; 1}_{d_j+1}]\hat{\boldsymbol{\phi}}_j(t) = \hat{\phi}_{j1}(t); \quad j = 1, \ldots, p. \tag{8}$$

3 Recursive least squares with forgetting factor

In this section the well known *RLS* algorithm with forgetting factor is compared to the proposed method of estimation for the varying-coefficient approach. Furthermore, it is shown how to include local polynomial approximations in the *RLS* algorithm.

3.1 The weight function

The *RLS* algorithm with forgetting factor aims at estimating the parameters in the linear model

$$y_i = \boldsymbol{z}_i^T \boldsymbol{\theta} + e_i \tag{9}$$

which corresponds to Eq. 1 when $\boldsymbol{\theta}(\boldsymbol{x}_i)$ is replaced by a constant vector $\boldsymbol{\theta}$. The parameter estimate $\hat{\boldsymbol{\theta}}(t)$, using the *RLS* algorithm with constant forgetting factor λ, is given by

$$\hat{\boldsymbol{\theta}}(t) = \mathbf{arg}\min_{\theta} \sum_{i=1}^{t} \lambda^{t-i} (y_i - \boldsymbol{z}_i^T \boldsymbol{\theta})^2. \tag{10}$$

In this case the weight which is assigned to the ith observation in Eq. 10 can be written as

$$w_i(t) = \lambda^{t-i} = \left[\exp\left(\frac{i-t}{(\ln\lambda)^{-1}}\right)\right]^{-1} = \left[\exp\left(\frac{|i-t|}{-(\ln\lambda)^{-1}}\right)\right]^{-1} \tag{11}$$

where the fact that $i \leq t$ in Eq. 10 is used. Now it is easily seen that Eq. 11 corresponds to Eq. 3 with a fixed bandwidth $\hbar(t) = \hbar = -(\ln\lambda)^{-1}$, which furthermore shows how the bandwidth and the forgetting factor are related. By also comparing Eq. 9 and Eq. 1 it is thus verified that the *RLS* algorithm with forgetting factor corresponds to local constant estimates in the varying-coefficient approach, with the specific choice Eq. 11 of the weight function.

3.2 Recursive local polynomial approximation

The *RLS* algorithm is given by (Ljung & Söderström 1983)

$$\mathbf{R}(t) = \sum_{i=1}^{t} \lambda^{t-i} \boldsymbol{z}_t \boldsymbol{z}_t^T = \lambda \mathbf{R}(t-1) + \boldsymbol{z}_t \boldsymbol{z}_t^T, \tag{12}$$

$$\hat{\boldsymbol{\theta}}(t) = \hat{\boldsymbol{\theta}}(t-1) + \mathbf{R}^{-1}(t)\boldsymbol{z}_t \left[y_t - \boldsymbol{z}_t^T \hat{\boldsymbol{\theta}}(t-1)\right], \tag{13}$$

with initial values

$$\mathbf{R}^{-1}(0) = \alpha \mathbf{I}, \quad \boldsymbol{\theta}(0) = \mathbf{0},$$

where α is large (Ljung & Söderström 1983). Hence, the recursive algorithm is only asymptotically equivalent to solving the least squares criteria Eq. 10, which on the other hand does not give a unique solution for small values of t.

In Section 2.3 it was shown how to include local polynomial approximation of the parameters in the varying-coefficient approach, and that this could be done by fitting the linear model Eq. 7 and calculating the parameters from Eq. 8. It is thus obvious to use the same approach in an extension of the *RLS* algorithm, replacing $\boldsymbol{z}_t$ by $\boldsymbol{u}_{i,t}$. However, the explanatory variable $\boldsymbol{u}_{i,t}$ is a function of t, which means that as we step forward in time,

$$\mathbf{R}(t-1) = \sum_{i=1}^{t-1} \lambda^{t-1-i} \boldsymbol{u}_{i,t-1} \boldsymbol{u}_{i,t-1}^T$$

can not be used in the updating formula for $\mathbf{R}(t)$, as $\mathbf{R}(t)$ depends on $\boldsymbol{u}_{i,t}$. To solve this problem a linear operator which is independent of t, and maps $\boldsymbol{p}_{d_j}(s)$ to $\boldsymbol{p}_{d_j}(s+1)$ has to be constructed. Using the coefficients of the relation

$$(s+1)^d = s^d + ds^{d-1} + \frac{d(d-1)}{2!} s^{d-2} + \cdots + 1. \tag{14}$$

it follows that

$$\begin{aligned}\boldsymbol{p}_{d_j}(s+1) &= \begin{bmatrix} 1 & d_j & \frac{d_j(d_j-1)}{2!} & \frac{(d_j-1)(d_j-2)}{3!} & \cdots & 1 \\ 0 & 1 & d_j-1 & \frac{(d_j-1)(d_j-2)}{2!} & \cdots & 1 \\ & & 1 & d_j-2 & \cdots & 1 \\ & & & 1 & & \\ \vdots & & & & \ddots & \vdots \\ 0 & \cdots & & & \cdots & 1 \end{bmatrix} \begin{bmatrix} s^{d_j} \\ s^{d_j-1} \\ \vdots \\ 1 \end{bmatrix} \\ &= \mathbf{L}_j \boldsymbol{p}_{d_j}(s)\end{aligned} \tag{15}$$

Since $\mathbf{L}_j$ is a linear operator it can be applied directly to $\boldsymbol{u}_{i,t} = \mathbf{L}\boldsymbol{u}_{i,t-1}$, where

$$\mathbf{L} = \begin{bmatrix} \mathbf{L}_1 & 0 & 0 & 0 & 0 \\ 0 & \mathbf{L}_2 & 0 & 0 & 0 \\ \vdots & & & & \vdots \\ 0 & \cdots & & \cdots & \mathbf{L}_p \end{bmatrix} \tag{16}$$

Which, when applied to the recursive calculation Eq. 12 of $\mathbf{R}(t)$, yields

$$\mathbf{R}(t) = \lambda \mathbf{L}\mathbf{R}(t-1)\mathbf{L}^T + \boldsymbol{u}_t \boldsymbol{u}_t^T , \tag{17}$$

and the updating formula for the parameters Eq. 13 is left unchanged. The proposed algorithm will be denoted *POLRLS* (Polynomial *RLS*) in the following.

Note that if the polynomials in Eq. 4 were calculated for the argument i instead of $t-i$, then $\boldsymbol{u}_{i,t} = \boldsymbol{u}_{i,t-1}$, and it is seen that the recursive calculation in Eq. 12 could be used without modification, but now there would be a numerical problem for $t \to \infty$.

4 Simulation study

Simulation is used to compare the *RLS* and *POLRLS* algorithms. For this purpose we have generated $N = 11$ samples of $n = 1000$ observations from the time-varying ARX-model

$$y_i = a y_{i-1} + b(i) z_i + e_i, \quad e_i \in N(0,1),$$

where

$$a = 0.7, \quad b(i) = 5 + 4 \sin\left(\frac{2\pi}{1000} i\right), \quad z_i \in N(0,1).$$

The estimation results are compared using the sample mean of the mean square error (*MSE*) of the deviation between the true and the estimated parameters

$$\begin{aligned} MSE_a &= \tfrac{1}{N-1} \sum_{j=2}^{N} \left\{ \tfrac{1}{n-s+1} \sum_{i=s}^{n} (a - \hat{a}(i))^2 \right\} \\ MSE_b &= \tfrac{1}{N-1} \sum_{j=2}^{N} \left\{ \tfrac{1}{n-s+1} \sum_{i=s}^{n} (b(i) - \hat{b}(i))^2 \right\} \end{aligned}$$

and the sample mean of the *MSE* of the predictions

$$MSE_p = \frac{1}{N-1} \sum_{j=2}^{N} \left\{ \frac{1}{n-s+1} \sum_{i=s}^{n} (y_i - \hat{a}(i-1) y_{i-1} - \hat{b}(i-1) z_i)^2 \right\} . \tag{18}$$

Only observations for which $i \geq s = 350 > max(\hbar_{opt})$, where $\hbar_{opt}$ is the optimal bandwidth, are used in the calculation of the *MSE*, to make sure that the effect of the initialisation has almost vanished. The observations used for the prediction in Eq. 18, has not been used for the estimation of the parameters, therefore the optimal bandwidth, $\hbar_{opt}$, can be found by minimizing Eq. 18 with respect to the bandwidth $\hbar$, i.e. forward validation. The optimal bandwidth is found using the first sample, $j = 1$, the 10 following are used for the calculation of the sample means.

The *POLRLS* method was applied with two different sets of polynomial orders. The results are shown in Figure 1 and Table 1. Obviously, knowing the true model, a zero order polynomial approximation of a and a second order polynomial approximation of b, should be the most adequate choice. In a true application such knowledge might not be available, i.e. if no preliminary analysis of data is performed. Therefore, a second order polynomial approximation is used for both parameters, as this could be the default or standard choice. In both cases the *POLRLS* algorithm performs significantly better than the *RLS* algorithm, and, as expected, using a second order approximation of a increases the *MSE* because in this case the estimation is disturbed by non-significant explanatory variables.

Method	Pol. order	$\hbar_{opt}$	MSE_p	MSE_a	MSE_b
POLRLS	$d_1 = 2, d_2 = 2$	62	1.0847	0.0024	0.0605
POLRLS	$d_1 = 0, d_2 = 2$	57	1.0600	0.0005	0.0580
RLS	$d_1 = 0, d_2 = 0$	11	1.1548	0.0044	0.0871

Table 1: *MSE* results using the *RLS* and *POLRLS* algorithms.

In the figure it is seen, that it is especially when the value of $b(i)$ is small, that the variance of $\hat{a}$ is large. In this case the signal to noise ratio is low, and the fact that a larger bandwidth can be used in the new algorithm, means that the variance can be significantly reduced. Furthermore, it is seen that the reduction of the parameter estimation variance is greater for the fixed parameter than the time varying parameter. The reason for this is that the optimal bandwidth is found by minimising the *MSE* of the predictions, and bias in the estimate of b contributes relatively more to the *MSE* than variance in the estimate of a, i.e. the optimal value of $\hbar$ balances bias in the estimate of b and variance in the estimate

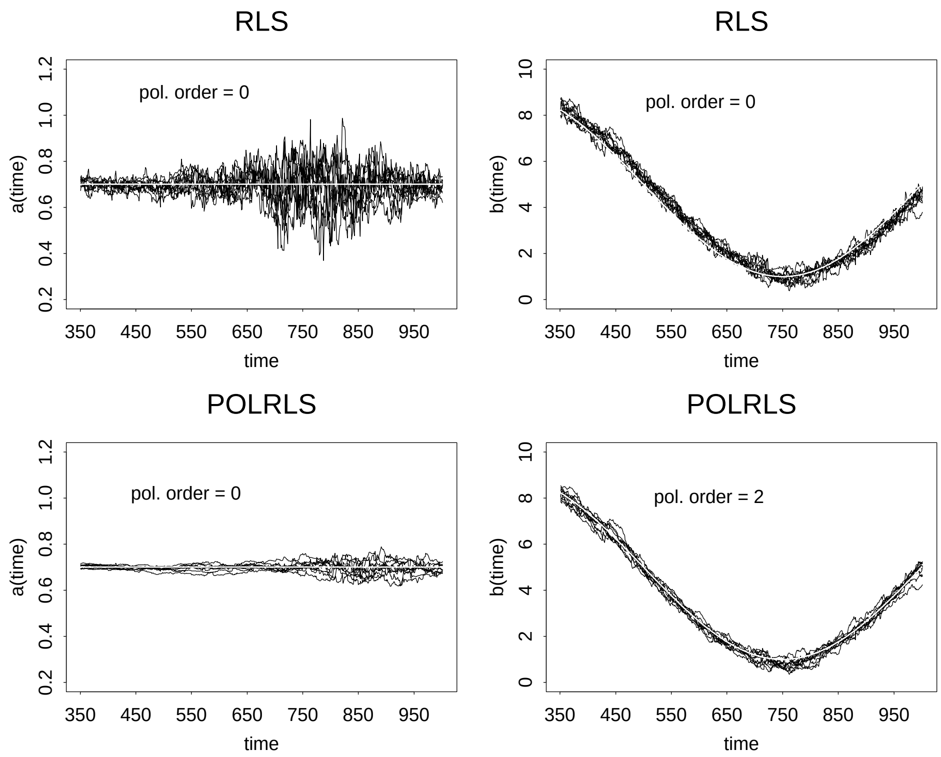


Figure 1: Estimated parameter trajectories. The first row shows the trajectories from the *RLS* algorithm, the second row shows the result from the *POLRLS* algorithm where a has been approximated by a zero order polynomial, and b by a second order polynomial.

of a. When a second order polynomial is used instead of a zero order polynomial, for the estimation of b, it is possible to avoid bias even when a significantly larger bandwidth is used.

5 Summary

In this paper the similarity between the varying-coefficient approach and the *RLS* algorithm with forgetting factor has been demonstrated. Furthermore an extension of the *RLS* algorithm, along the lines of the varying-coefficient approach is suggested. Using an example it is shown that the new algorithm leads to an significantly improvement of the estimation performance, if the variation of the true parameters is smooth.

References

Anderson, T. W., Fang, K. T. & Olkin, I., eds (1994), *Multivariate Analysis and Its Applications*, Institute of Mathematical Statistics, Hayward, chapter Coplots, Nonparametric Regression, and conditionally Parametric Fits, pp. 21–36.

Cleveland, W. S. & Devlin, S. J. (1988), 'Locally weighted regression: An approach to regression analysis by local fitting', *Journal of the American Statistical Association* **83**, 596–610.

Fan, J., Hardle, W. & Mammen, E. (1998), 'Direct estimation of low dimensional components in additive models', *The Annals of Statistics* **26**, 943–971.

Hastie, T. J. & Tibshirani, R. J. (1990), *Generalized Additive Models*, Chapman & Hall, London/New York.

Hastie, T. & Tibshirani, R. (1993), 'Varying-coefficient models', *Journal of the Royal Statistical Society, Series B, Methodological* **55**, 757–796.

Holt, C. (1957), 'Forecasting trends and sesonals by exponentially weighted moving averages', *O.N.R. Memorandum 52* . Carnegie Institute of Technology.

Ljung, L. & Söderström, T. (1983), *Theory and Practice of Recursive Identification*, MIT Press, Cambridge, MA.

Nielsen, H. A., Nielsen, T. S., Madsen, H. & Joensen, A. (1999), 'Tracking time-varying coefficient-functions'. To be published.

Parkum, J. E., Poulsen, N. K. & Holst, J. (1992), 'Recursive forgetting algorithms', *Int. J. Control* **55**, 109–128.

Winters, P. (1960), 'Forecasting sales by exponentially weighted movings averages', *Man. Sci.* **6**, 324–342.

Paper F

A Semi-parametric approach for decomposition of absorption spectra in the presence of unknown components

F

Originally published as

> Payman Sadegh, Henrik Aalborg Nielsen, and Henrik Madsen. A semi-parametric approach for decomposition of absorption spectra in the presence of unknown components. Technical Report 17, Department of Mathematical Modelling, Technical University of Denmark, DK-2800 Lyngby, Denmark, 1999.

In the version included here some erroneous cross-references are corrected.

A Semi-parametric approach for decomposition of absorption spectra in the presence of unknown components

Payman Sadegh[1], Henrik Aalborg Nielsen[1] and Henrik Madsen[1]

Abstract

Decomposition of absorption spectra using linear regression has been proposed for calculating concentrations of mixture compounds. The method is based on projecting the observed mixture spectrum onto the linear space generated by the reference spectra that correspond to the individual components comprising the mixture. The computed coefficients are then used as estimates for concentration of the components that comprise the mixture. Existence of unknown components in the mixture, however, introduces bias on the obtained concentration estimates. We extend the usual linear regression model to an additive semi-parametric model to take the unknown component into account, estimate the absorption profile of the unknown component, and obtain concentration estimates of the known compounds. A standard back-fitting method as well as a mean weighted least squares criterion are applied. The techniques are illustrated on simulated absorption spectra.

Keywords: Parameter estimation, non-parametric methods, unbiased estimates, chemometry, absorption spectra, additive models, semi-parametric models, mean weighted least squares, back-fitting.

1 Introduction

Chemometric spectroscopy is a simple way for examination of gases and liquids. UV examination of wastewater for instance has been proposed for quality control purposes (Thomas, Theraulaz & Suryani 1996). The technique is based on the analysis of the absorption spectrum obtained

[1]Department of Mathematical Modelling, Technical University of Denmark, DK-2800 Lyngby, Denmark

from the sample of interest. Depending on the concentrations of comprising compounds, the spectral absorption of the mixtures vary at different wavelengths. This information may in principle be used to *encode* the concentration of an existing compound, given information about the absorption pattern of the compound of interest. The functional dependency of absorption spectra upon concentrations and absorption spectra of comprising compounds is in general unknown. Several simple models have been proposed, most notably, a linear regression model (Gallot & Thomas 1993). In this model, it is assumed that the absorption spectrum of a mixture is a linear combination of absorption spectra of the comprising compounds where each coefficients determines the concentration of the corresponding compound. Hence, if spectral absorption measurements are performed at minimum p wavelengths, where p is the number of existing compounds, the concentrations may be estimated by the least squares technique (Gallot & Thomas 1993). The technique fails when unknown compounds are present. Even though, it is not of interest to estimate the concentration of the unknown components, the presence of such will introduce bias on the concentration estimates for the known ones unless the spectrum of the unknown component is orthogonal to the spectra of the known ones. Such a situation is unlikely to occur for most decomposition problems of interest in chemometry. We propose a semi-parametric model to account for the presence of unknown compounds. We apply both a standard back-fitting method for estimation in additive models and a novel technique based on a mean weighted least squares criterion (MWLS). MWLS provides a promising easy way to embed prior information into kernel estimation schemes. While kernel estimation of a function at N data points may be regarded as N independent weighted least squares problems, the MWLS combines the N optimization problems into one. The advantage is that any global information about the behavior of the process may be imposed as hard or soft constraints in the resulting single optimization criterion.

The rest of the paper is organized as follows. In Section 2, we present the semi-parametric formulation of spectral absorption decomposition problem, we review techniques from the theory of general additive models, and introduce the MWLS technique. In Section 3, we present a simple numerical study based on simulated data, and finally, Section 4 presents concluding remarks.

2 Problem formulation

Consider the problem of decomposing the observed function $f(t)$, $t \in T$, into known functions $f_i(t)$, $i = 1, ..., p$, by estimating the parameter $\theta = [\theta_1, \cdots, \theta_p]^\top$ of the linear regression

$$f(t) = \sum_{i=1}^{p} \theta_i f_i(t) + R(t) + e(t), \tag{1}$$

where $R(t)$ accounts for the superposition of all unknown components comprising $f(t)$, and $\{e(t)\}$ is a sequence of zero mean independently distributed random variables representing the measurement noise. Disregarding $R(t)$, the usual least squares estimate of θ is given by

$$\arg\min_{\theta} \sum_{t \in T} [f(t) - \sum_{i=1}^{p} \theta_i f_i(t)]^2. \tag{2}$$

where $T = \{t_1, \cdots, t_N\}$ is the set of N sampled observations. Inserting $f(t)$ from (1) in the solution to (2) shows that the bias on the least squares estimate is given by

$$(X^\top X)^{-1} X^\top \begin{bmatrix} R(t_1) \\ \vdots \\ R(t_N) \end{bmatrix}$$

where

$$X = \begin{bmatrix} f_1(t_1) & \dots & f_p(t_1) \\ f_1(t_2) & \dots & f_p(t_2) \\ \vdots & \vdots & \vdots \\ f_1(t_N) & \dots & f_p(t_N) \end{bmatrix}.$$

Hence depending on $R(t)$, the bias might be arbitrarily large.

For chemometric data, the function $f(t)$ is the measured absorption spectrum at various wavelengths t and $f_i(t)$ is the known absorption spectrum for component i. The presence of unknown components introduces the remainder term $R(t)$ which should be simultaneously with the coefficients θ_i estimated from data. The back-fitting algorithm (Hastie & Tibshirani 1990) can be applied under such circumstances. Back-fitting

is an iterative method for decomposition of a number of unknown functions in an additive model. Starting from an initial guess, the algorithm iteratively estimates each one of the functions, fixing all others to their corresponding latest updated values. The algorithm has an explicit solution for problems of the type (1), see (Hastie & Tibshirani 1990), page 118. The solution involves a smoother function for estimation of $R(\cdot)$ and a weighted least squares criterion to estimate θ. Only in the case the smoother is a spline smoother, the back-fitting can be explicitly related to an optimization criterion (Hastie & Tibshirani 1990).
Another approach, which is novel to the best of our knowledge, is based on a mean weighted least squares criterion. The approach is particularly appealing since its solution is explicitly related to an optimization criterion which is a missing element for back-fitting using other smoothers than splines. The MWLS approach is as follows. Consider the following optimization

$$\min_{\theta,\{\phi(\tau)\}} \sum_{\tau\in T}\sum_{t\in T} w_h(|t-\tau|)[f(t)-\sum_{i=1}^{p}\theta_i f_i(t)-q\,(t-\tau;\phi(\tau))]^2, \tag{3}$$

where $q\,(t-\tau;\phi(\tau))$ and $\phi(\tau)$ respectively denote a local approximation to $R(\cdot)$ around τ and its corresponding (τ-dependent) parameter and $\{w_h(|d|)\}$ is a weight sequence that falls monotonically with $|d|$. One typical choice for $q\,(t-\tau;\phi(\tau))$ is a low order polynomial in $t-\tau$. The weight sequence is obtained from the kernel $K_h(|d|)$ according to

$$w_h(|d|) = \frac{K_h(|d|)}{\sum\limits_{d} K_h(|d|)}.$$

Some typical selections for the kernel $K_h(|d|)$ are Gaussian kernel

$$K_h(|d|) = \frac{1}{h\sqrt{2\pi}}\exp\left(-\frac{d^2}{2h^2}\right),$$

and Epanechnikov kernel

$$K_h(|d|) = \frac{3}{4h}\left(1-\frac{d^2}{h^2}\right) I(|d|\le h),$$

where $I(|d|\le h)=1$ if $|d|\le h$ and zero otherwise.

The criterion (3) may be explained as follows. The optimization problem obtained by considering the inner sum in (3) as the cost function, i.e.

$$\min_{\theta,\phi(\tau)} \sum_{t\in T} w_h(|t-\tau|)[f(t) - \sum_{i=1}^{p} \theta_i f_i(t) - q\,(t-\tau;\phi(\tau))]^2, \tag{4}$$

provides the usual weighted least squares problem for non-parametric estimation of $R(\tau)$, $\tau \in T$, based on the local approximation $R(t) \approx q(t-\tau;\phi(\tau))$ around τ. Hence a non-parametric estimate for $R(\tau)$ is obtained by inserting the optimal value of $\phi(\tau)$ in $q(0;\phi(\tau))$. In connection with estimating θ, on the other hand, (4) is of no immediate use since the obtained estimates of θ vary with τ. This dependency is eliminated in (3) by the outer summation over τ. This may be thought of as restricting the solutions to the independent optimization problems (4) to yield a common estimate for θ.

Now assume that the local approximation $q(t-\tau;\phi(\tau))$ is linear in $\phi(\tau)$, i.e.

$$q\,(t-\tau;\phi(\tau)) = \sum_{i=1}^{m} \phi_i(\tau) g_i(t-\tau) \tag{5}$$

where $\phi(\tau) = [\phi_1(\tau), \cdots, \phi_m(\tau)]^\top$. Let W_τ denote a diagonal $N \times N$ matrix with the (i,i) element being equal to $w_h(|t_i-\tau|)$. Further denote

$$X_\tau = \begin{bmatrix} g_1(t_1-\tau) & \dots & g_m(t_1-\tau) \\ g_1(t_2-\tau) & \dots & g_m(t_2-\tau) \\ \vdots & \vdots & \vdots \\ g_1(t_N-\tau) & \dots & g_m(t_N-\tau) \end{bmatrix},$$

and finally $Y = [f(t_1), \cdots, f(t_N)]^\top$.

Proposition 1 Assume that the local approximation $q(t-\tau;\phi(\tau))$ is linear in $\phi(\tau)$ (see (5)). The optimal value of θ according to (3) is equivalent to the solution to the weighted least squares problem

$$\min_{\theta} (Y - X\theta)^\top W (Y - X\theta)$$

where

$$W = \sum_{\tau\in T} \left(W_\tau - W_\tau X_\tau (X_\tau^\top W_\tau X_\tau)^{-1} X_\tau^\top W_\tau \right)$$

PROOF: Since $\phi(\tau)$ varies with τ in (3), $\phi(\tau)$ may be simply computed by finding the optimal value of $\phi(\tau)$ in (4) as a function of θ. Inserting the optimal values for $\phi(\tau)$ in (3) and collecting terms yields the desired result.

3 Numerical example

In this section, we apply the techniques discussed earlier to a spectral decomposition problem. Consider two absorption spectra $f_1(t)$ and $f_2(t)$ as given in Figure 1. These spectra contain COD, TOC, TSS, and BOD with concentrations 63, 0, 15, and 15 for f_1 and 36, 12.5, 0, 11.5 for f_2. The "unknown component" is assumed to consist of concentrations of nitrates with spectrum $R(t)$ as illustrated in Figure 2. The spectra of Figure 1 and Figure 2 are taken experimentally from real samples.

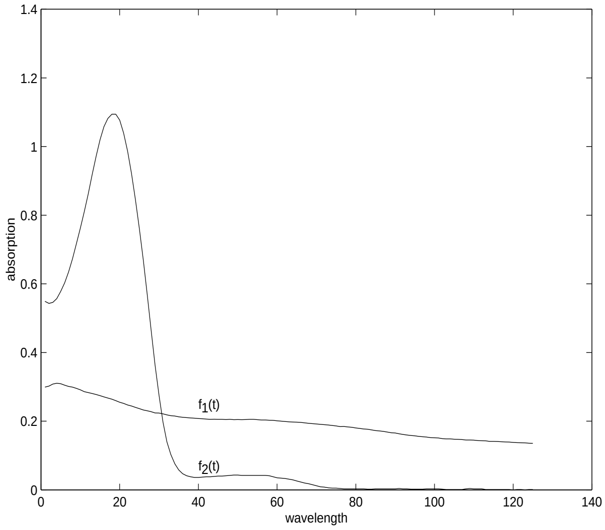


Figure 1: Reference absorption spectra.

We simulate the spectrum illustrated in Figure 3 by the linear combination:

$$f(t) = f_1(t) + f_2(t) + R(t).$$

The least squares solution yields coefficient estimates 1.35 and 1.81 for f_1

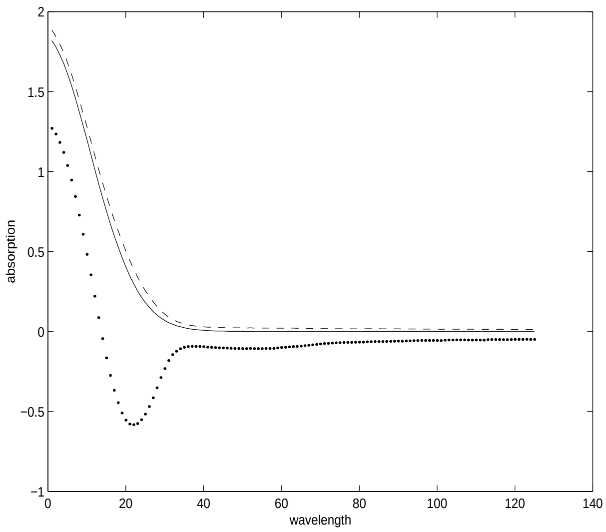


Figure 2: Spectrum for the unknown component. The solid curve, dashed, and dotted curves respectively represent the true spectrum, the estimated spectrum using a mean weighted least squares criterion, and the estimated spectrum using regular least squares.

and f_2, and an estimate for $R(t)$ as illustrated in Figure 2. These results clearly indicate the inappropriateness of the least squares solution.

We apply the result presented in Proposition 1 where the local approximators are polynomials of second order and the weights are computed according to a unit bandwidth Gaussian kernel. The coefficients of f_1 and f_2 are respectively estimated to 0.91 and 0.93.

We further apply the back-fitting solution to the above estimation problem. The best result is obtained by applying spline smoother with a large degree of freedom, yielding estimates of 1.25 and 0.73 for the coefficients of f_1 and f_2 respectively. These estimates are noticeably more biased than the MWLS solution.

Finally, we investigate the effect of measurement noise. We simulate 25 independent samples according to

$$f(t) = f_1(t) + f_2(t) + R(t) + e(t)$$

where $\{e(t)\}$ is a sequence of i.i.d. $N(0, 0.001^2)$ random variables, and

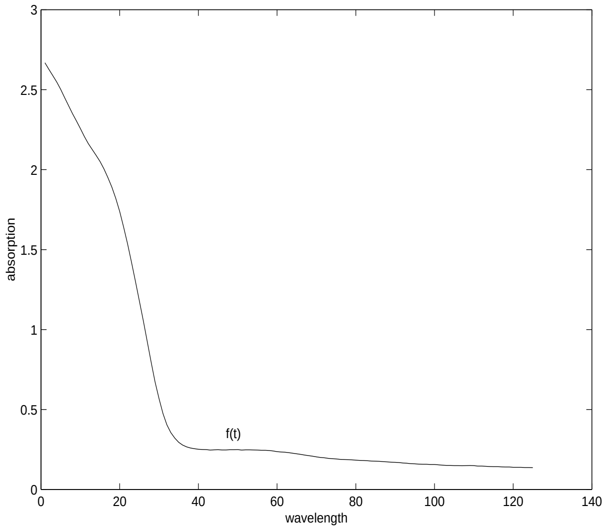


Figure 3: Simulated sample.

empirically compute mean and covariance of the concentration estimates $\hat{\theta}$. These quantities are computed to

$$E\{\hat{\theta}\} = \begin{bmatrix} 0.9039 \\ 0.9306 \end{bmatrix}^\top, \; COV\{\hat{\theta}\} = \begin{bmatrix} 0.0576 & -0.0007 \\ -0.0007 & 0.0030 \end{bmatrix}$$

where $E\{\cdot\}$ and $COV\{\cdot\}$ as usual denote mean value and covariance. Numerical experimentation indicates that the estimation procedure fails for noise variances above 0.01^2.

4 Discussion and conclusion

We have proposed a solution to decomposition of absorption spectra in presence of correlated error (e.g. due to existence of unknown components). The underlying assumption throughout the paper is a linear additive model. We have applied both the back-fitting solution and a mean weighted least squares criterion. Numerical experience with back-fitting iterations for typical chemometric spectra fails due to high correlation among data. The back-fitting method yields reasonable estimates only if the explicit end solution of the iterations, which exists for a model linear in parameters, is applied. Contrary to back-fitting, the mean weighted least squares solution is not tied to an iterative algorithm but to a well defined optimization problem. The mean weighted least squares solution performs reasonably well for decomposition of absorption spectra in the presence of unknown components. The solution is rather sensitive to measurement noise which is again due to high correlation among reference spectra on the one hand and high correlation between the unmodelled error spectrum and the reference spectra on the other.

To further elaborate on the discussion above, Figure 4 on page 136 shows a scatter-plot matrix of the wavelength and five typical absorption spectra f_i, $i = 1, \cdots, 5$. The actual spectra are shown in the left column and, with the axes swapped, in the bottom row. From these it is seem that $f_2(t)$ and $f_3(t)$ are very similar, correspondingly the plot of $f_2(t)$ against $f_3(t)$ shows an almost linear relation. Consequently, concentrations of substances corresponding to these spectra will be difficult to distinguish, i.e. the estimated concentrations will be highly correlated.

For data simulated according to some arbitrary linear combination of the illustrated spectra and some typical spectrum for the unknown component $R(t)$ (simulated as a Guassian bell curve around some bandwidth), the R-squared value is above 0.9999 when omitting $R(t)$ from the model and replacing it with a intercept term. This indicate that the simulated spectrum $f(t)$ lies almost entirely in the space spanned by the reference spectra, making estimation of $R(t)$ difficult if $f(t)$ is measured with noise. Consequently, to reduce bias on the estimates of concentrations $R(t)$ must, to some extend which is determined by the level of noise, lie in an other space than the one spanned by the reference spectra.

The above considerations indicate that, if possible, reference spectra should be chosen so that (1) all explain different aspects of the unknown spectra (as opposed to $f_2(t)$ and $f_3(t)$ above), and (2) the unknown $R(t)$ lies, to some extend, in an other space than the one spanned by the reference spectra. Furthermore, measurement noise should be reduced as much as possible, e.g. by performing several measurements on the sample of interest and averaging.

Finally, the mean weighted least squares criterion introduced in this paper has application potentials far beyond the scope of the present paper.

The approach provides a simple yet powerful tool to embed "global" information about the process of interest in local estimation techniques, hence combining the noise reduction and interpretability of global models with the generality, minimal model reliance, and convenience of non-parametric methods. Contrast this to usual ways of embedding prior information in non-parametric methods which concern local or smoothness properties such as selection of a suitable kernel, bandwidth, and degree of local approximators. This poses an interesting direction for future research and forthcoming publications.

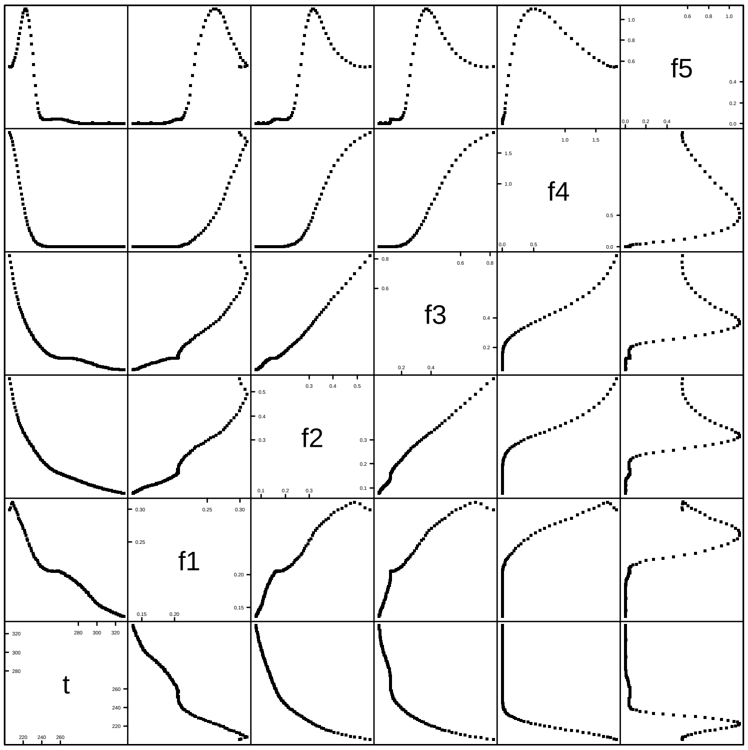


Figure 4: Scatter-plot matrix of wavelength (t) and reference spectra ($f_1(t), \ldots, f_5(t)$.

References

Gallot, S. & Thomas, O. (1993), 'Fast and easy interpretation of a set of absorption spectra: theory and qualitative applications for UV examination of waters and wastewatres', *Fresenius Journal of Analytical Chemistry* **346**, 976–983.

Hastie, T. J. & Tibshirani, R. J. (1990), *Generalized Additive Models*, Chapman & Hall.

Thomas, O., Theraulaz, F. & Suryani, S. (1996), 'Advanced UV examination of wastewaters', *Environmental Technology* **17**, 251–261.

Paper G

A generalization of some classical time series tools

Submitted to *Computational Statistics and Data Analysis*. A previous version is available as IMM technical report number 1998-16.

G

A generalization of some classical time series tools

Henrik Aa. Nielsen[1] and Henrik Madsen[1]

Abstract

In classical time series analysis the sample autocorrelation function ($SACF$) and the sample partial autocorrelation function ($SPACF$) has gained wide application for structural identification of linear time series models. We suggest generalizations, founded on smoothing techniques, applicable for structural identification of non-linear time series models. A similar generalization of the sample cross correlation function is discussed. Furthermore, a measure of the departure from linearity is suggested. It is shown how bootstrapping can be applied to construct confidence intervals under independence or linearity. The generalizations do not prescribe a particular smoothing technique. In fact, when the smoother are replaced by linear regressions the generalizations reduce to close approximations of $SACF$ *and* $SPACF$*. For this reason a smooth transition form the linear to the non-linear case can be obtained by varying the bandwidth of a local linear smoother. By adjusting the flexibility of the smoother the power of the tests for independence and linearity against specific alternatives can be adjusted. The generalizations allow for graphical presentations, very similar to those used for* $SACF$ *and* $SPACF$*. In this paper the generalizations are applied to some simulated data sets and to the Canadian lynx data. The generalizations seem to perform well and the measure of the departure from linearity proves to be an important additional tool.*

Keywords: Lagged scatter plot; R-squared; Non-linear time series; Smoothing; Non-parametric; Independence; Bootstrap.

[1]Department of Mathematical Modelling, Technical University of Denmark, DK-2800 Lyngby, Denmark

1 Introduction

The sample autocorrelation function and the sample partial autocorrelation function have gained wide application for structural identification of linear time series models. For non-linear time series these tools are not sufficient because they only address linear dependencies.

During the last couple of decades a number of results on properties of, and estimation and testing in, nonlinear models have been obtained. For an overview (Priestley 1988, Tong 1990, Tjøstheim 1994) can be consulted. However, considerable fewer results have been seen on the problem of structural identification. Tjøstheim & Auestad (1994) have suggested a method based on kernel estimates to select the significant lags in a non-linear model, and Granger & Lin (1994) used the mutual information coefficient and Kendall's τ as generalizations of the correlation coefficient and Kendall's partial τ as a generalization of the partial correlation coefficient. Chen & Tsay (1993) have considered a best subset modelling procedure and the ACE and BRUTO algorithms, see e.g. (Hastie & Tibshirani 1990), for identification of non-linear additive ARX models. Recently, Lin & Pourahmadi (1998) have used the BRUTO algorithm to identify the lags needed in a semi-parametric non-linear model. Multivariate adaptive regression splines (Friedman 1991) was introduced for modelling of non-linear autoregressive time series by Lewis & Stevens (1991). Teräsvirta (1994) suggested a modelling procedure for non-linear autoregressive time series in which a (parametric) smooth threshold autoregressive model is used in case a linear model proves to be inadequate. For the case of non-linear transfer functions Hinich (1979) considered the case where the impulse response function of the transfer function depends linearly on the input process.

In this paper we suggest the new tools LDF (Lag Dependence Function), $PLDF$ (Partial Lag Dependence Function), and $NLDF$ (Non-linear Lag Dependence Function) for structural identification of non-linear time series. The tools can be applied in a way very similar to the sample autocorrelation function and the sample partial autocorrelation function. Smoothing techniques are used, but the tools are not dependent on any particular smoother, see e.g. (Hastie & Tibshirani 1990, Chapter 3) for an overview of smoothing techniques. For some smoothers an (almost) continuous transition from the linear to the non-linear case can be obtained by varying the smoothing parameter. Also, smoothers applying optimal selection of the bandwidth may be used; however, see e.g. (Chen & Tsay 1993) for a discussion of the potential problems in applying criteria such as generalized cross validation to time series data. Under a hypothesis of independence bootstrap confidence intervals (Efron & Tibshirani 1993) of the lag dependence function are readily calculated, and we propose that these can also be applied for the partial lag dependence function. Furthermore, under a specific linear hypothesis, bootstrapping can be used to construct confidense intervals for the non-linear lag dependence function. The lag dependence function and the non-linear lag dependence function are readily calculated in that only univariate smoothing are needed, whereas multivariate smoothing or backfitting are required for the application of the partial lag dependence function.

It is noted that the tools suggested does not claim to estimate any underlying property of the stochastic process by which the data are generated. Instead they, essentially, measure the in-sample variance reduction of a specific model compared to a reduced model, see also (Anderson-Sprecher 1994). The models are specified both in terms of the lags included and the smoothers applied. The lags identified are thus conditional on the generality of the non-linearity allowed for. Since the size of the confidence intervals depend on the flexibility of the smoother used it is informative to apply the tools using a range of smoothing parameters.

The tools are illustrated both by using simulated linear and non-linear time series models, and by considering the Canadian lynx data (Moran 1953), which have attained a bench-mark status in time series literature. Using the Canadian lynx data results very similar to those found by Lin & Pourahmadi (1998) are obtained.

In Section 2 the study is motivated by considering a simple deterministic non-linear process for which the sample autocorrelation function is non-significant. Section 3 describes the relations between multiple linear regression, correlation, and partial correlation with focus on aspects leading to the generalization. The proposed tools are described in Sections 4, 5, and 6 and bootstrapping is considered in Section 7. Examples of application by considering simulated linear and non-linear processes and the Canadian lynx data (Moran 1953) are found in Section 8. In Section 9 a generalization of the sample cross correlation function is briefly discussed. Finally, in Section 10 some further remarks are given.

2 Motivation

The sample autocorrelation function (Brockwell & Davis 1987), commonly used for structural identification in classical time series analysis, measures only the degree of linear dependency. In fact deterministic series exists for which the sample autocorrelation function is almost zero, see also (Granger 1983). One such example is $x_t = 4x_{t-1}(1 - x_{t-1})$ for which Figure 1 shows 1000 values using $x_1 = 0.8$ and the corresponding sample autocorrelation function $SACF$ together with an approximative 95% confidence interval of the estimates under the hypothesis that the underlying process is i.i.d. Furthermore lagged scatter plots for lag one and two are shown. From the plot of the series and the $SACF$ the deterministic structure is not revealed. However, the lagged scatter plots clearly reveals that the series contains a non-linear dynamic dependency.

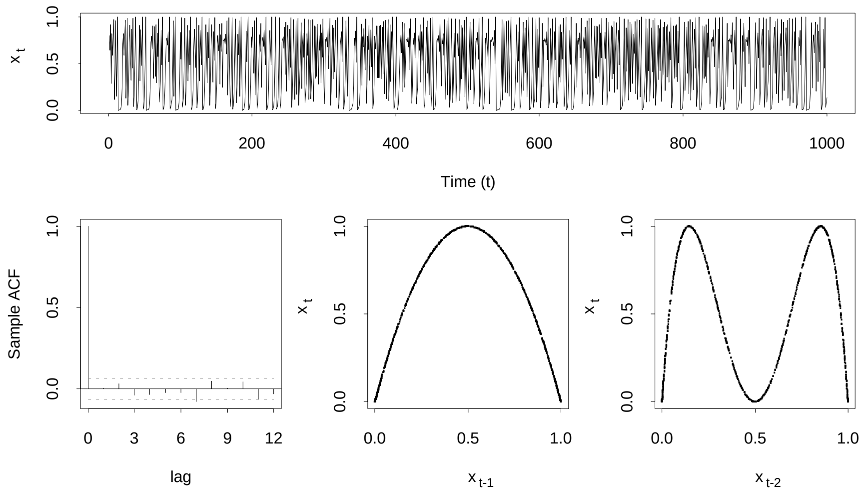


Figure 1: The time series (top), $SACF$ (bottom, left), x_t versus x_{t-1} (bottom, middle), and x_t versus x_{t-2} (bottom, right) for 1000 values from the recursion $x_t = 4x_{t-1}(1 - x_{t-1})$.

In practice the series will often be contaminated with noise and it is then difficult to judge from the lagged scatter plots whether any dependence is present. Smoothing the lagged scatter plots will aid the interpretation but different smoothing parameters may result in quite different estimates. Therefore it is important to separate the variability of the smooth from the underlying dependence.

From Figure 1 it is revealed that, in principle, x_t can be regarded as a function of x_{t-k} for any $k > 0$, but $k = 1$ is sufficient, since x_t can be predicted exactly from x_{t-1} alone. This indicates that there may exist a non-linear equivalent to the partial autocorrelation function (Brockwell & Davis 1987) and reveals that substantial information can be obtained by adjusting for the dependence of lag $1, \ldots, k-1$ when x_t and x_{t-k} are addressed. The sample partial autocorrelation function amounts to a linear adjustment.

3 Preliminaries

Estimates of correlation and partial correlation are closely related to values of the coefficient of determination (R-squared) obtained using linear regression models. The generalizations of the sample autocorrelation function $SACF$ and the sample partial autocorrelation function $SPACF$ are based on similar R-squared values obtained using non-linear models. In this section the relations between multiple linear regression, correlation, and partial correlation are presented.

Consider the multivariate stochastic variable $(Y, X_1, \ldots, X_k)$. The squared multiple correlation coefficient $\rho^2_{0(1\ldots k)}$ between Y and $(X_1, \ldots, X_k)$ can be written (Kendall & Stuart 1961, p. 334, Eq. (27.56))

$$\rho^2_{0(1\ldots k)} = \frac{V[Y] - V[Y \mid X_1, \ldots, X_k]}{V[Y]}. \tag{1}$$

If the variances are estimated using a maximum likelihood estimator, assuming normality, it then follows that an estimate of $\rho^2_{0(1\ldots k)}$ is

$$R^2_{0(1\ldots k)} = \frac{SS_0 - SS_{0(1\ldots k)}}{SS_0}, \tag{2}$$

where $SS_0 = \sum (y_i - \sum y_i/N)^2$ and $SS_{0(1\ldots k)}$ is the sum of squares of the least squares residuals when regressing y_i linearly on $x_{1i}, \ldots, x_{ki}$ ($i =$

$1, \ldots, N$). $R^2_{0(1\ldots k)}$ is also called the coefficient of determination of the regression and can be interpreted as the relative reduction in variance due to the regressors.

Hence it follows that when regressing y_i linearly on x_{ki} the coefficient of determination $R^2_{0(k)}$ equals the squared estimate of correlation between Y and X_k, and furthermore it follows that $R^2_{0(k)} = R^2_{k(0)}$.

The partial correlation coefficient $\rho_{(0k)|(1\ldots k-1)}$ between Y and X_k given $X_1, \ldots, X_{k-1}$ measures the extend to which, by using linear models, the variation in Y, which cannot be explained by $X_1, \ldots, X_{k-1}$, can be explained by X_k. Consequently, the partial correlation coefficient is the correlation between $(Y \mid X_1, \ldots, X_{k-1})$ and $(X_k \mid X_1, \ldots, X_{k-1})$, see also (Rao 1965, p. 270). Using (Whitaker 1990, p. 140) we obtain

$$\rho^2_{(0k)|(1\ldots k-1)} = \frac{V[Y \mid X_1, \ldots, X_{k-1}] - V[Y \mid X_1, \ldots, X_k]}{V[Y \mid X_1, \ldots, X_{k-1}]}. \tag{3}$$

For $k = 1$ it is readily seen that $\rho^2_{(0k)|(1\ldots k-1)} = \rho^2_{0(1)}$. If the variances are estimated using the maximum likelihood estimator, assuming normality, it follows that an estimate of $\rho^2_{(0k)|(1\ldots k-1)}$ is

$$R^2_{(0k)|(1\ldots k-1)} = \frac{SS_{0(1\ldots k-1)} - SS_{0(1\ldots k)}}{SS_{0(1\ldots k-1)}}. \tag{4}$$

Besides an estimate of $\rho^2_{(0k)|(1\ldots k-1)}$ this value can also be interpreted as the relative decrease in the variance when including x_{ki} as an additional predictor in the linear regression of y_i on $x_{1i}, \ldots, x_{k-1,i}$. Note that (4) may also be derived from (Ezekiel & Fox 1959, p. 193).

Interpreting $R^2_{0(1\ldots k)}$, $R^2_{0(k)}$, and $R^2_{(0k)|(1\ldots k-1)}$ as measures of variance reduction when comparing models (Anderson-Sprecher 1994), these can be calculated and interpreted for a wider class of models such as smooths and additive models. For non-linear models Kvålseth (1985, p. 282) suggests the use of a statistic like the square root of (2) as what is called "a generalized correlation coefficient or index suitable for both linear and non-linear models". In the remaining part of this paper "$\sim$" will be used above values of SS and R^2 obtained from models other than linear models.

4 Lag dependence

Assume that observations $\{x_1, \dots, x_N\}$ from a stationary stochastic process $\{X_t\}$ exists. It is readily shown that except for minor differences in the denominators the estimate of the autocorrelation function in lag k is equal to the estimate of the correlation coefficient between X_t and X_{t-k} using the observations $\{x_1, \dots, x_N\}$. Hence, the squared $SACF(k)$ can be closely approximated by the coefficient of determination when regressing x_t linearly on x_{t-k}, i.e. $R^2_{0(k)}$.

This observation leads to a generalization of $SACF(k)$, based on $\tilde{R}^2_{0(k)}$ obtained from a smooth of the k-lagged scatter plot, i.e. a plot of x_t against x_{t-k}. The smooth is an estimate of the conditional mean $f_k(x) = E[X_t \mid X_{t-k} = x]$. Thus, the Lag Dependence Function in lag k, $LDF(k)$, is calculated as

$$LDF(k) = \operatorname{sign}\left(\hat{f}_k(b) - \hat{f}_k(a)\right)\sqrt{(\tilde{R}^2_{0(k)})_+} \tag{5}$$

where a and b is the minimum and maximum over the observations and the subscript "+" indicates truncation of negative values. The truncation is necessary to ensure that (5) is defined. However, the truncation will only become active in extreme cases. Using a local linear smoother with a nearest neighbour bandwidth of 1/3 results in a negative R-squared at lag 4 for the series considered in Figure 1. Due to the combination of bandwidth and periocity at this lag the smooth obtained is in opposite phase of the data. The negative R-squared is thus consistent with the observations made by Kvålseth (1985) for the case of gross model misspecification.

Due to the reasons mentioned in the beginning of this section, when $\hat{f}_k(\cdot)$ is restricted to be linear, $LDF(k)$ is a close approximation of $SACF(k)$ and, hence, it can be interpreted as a correlation. In the general case $LDF(k)$ can be interpreted as (the signed square-root of) the part of the overall variation in x_t which can be explained by x_{t-k}. Generally, R-squared for the non-parametric regression of x_t on x_{t-k}, $\tilde{R}_{0(k)}$ do not equal R-squared for the corresponding non-parametric regression of x_{t-k} on x_t, and consequently, unlike $SACF(k)$, the lag dependence function is not an even function. In this paper only causal models will be considered and (5) will only be used for $k > 0$ and by definition $LDF(0)$ will be set equal to one.

5 Partial lag dependence

For the time series $\{x_1, \ldots, x_N\}$ the sample partial autocorrelation function in lag k, denoted $SPACF(k)$ or $\hat{\phi}_{kk}$, is obtainable as the Yule–Walker estimate of ϕ_{kk} in the $AR(k)$ model

$$X_t = \phi_{k0} + \phi_{k1} X_{t-1} + \ldots + \phi_{kk} X_{t-k} + e_t, \tag{6}$$

where $\{e_t\}$ is i.i.d. with zero mean and constant variance, see also (Brockwell & Davis 1987, p. 235). An additive, but non-linear, alternative to (6) is

$$X_t = \varphi_{k0} + f_{k1}(X_{t-1}) + \ldots + f_{kk}(X_{t-k}) + e_t. \tag{7}$$

This model may be fitted using the backfitting algorithm (Hastie & Tibshirani 1990), see also Section 5.1. The function $f_{kk}(\cdot)$ can be interpreted as a partial dependence function in lag k when the effect of lags $1, \ldots, k-1$ is accounted for. If the functions $f_{kj}(\cdot)$, $(j = 1, \ldots, k)$ are restricted to be linear then $\hat{f}_{kk}(x) = \hat{\phi}_{kk} x$ and the function can be uniquely identified by its slope $\hat{\phi}_{kk}$.

However, since the partial autocorrelation function in lag k is the correlation between $(X_t \,|\, X_{t-1}, \ldots, X_{t-(k-1)})$ and $(X_{t-k} \,|\, X_{t-1}, \ldots, X_{t-(k-1)})$, the squared $SPACF(k)$ may also be calculated as $R^2_{(0k)|(1\ldots k-1)}$, based on linear autoregressive models of order $k-1$ and k. Using models of the type (7) $SPACF(k)$ may then be generalized using an R-squared value obtained from a comparison of models (7) of order $k-1$ and k. This value is denoted $\tilde{R}^2_{(0k)|(1\ldots k-1)}$ and we calculate the Partial Lag Dependence Function in lag k, $PLDF(k)$, as

$$PLDF(k) = \operatorname{sign}\left(\hat{f}_{kk}(b) - \hat{f}_{kk}(a)\right) \sqrt{(\tilde{R}^2_{(0k)|(1\ldots k-1)})_+}. \tag{8}$$

When (7) is replaced by (6) $PLDF(k)$ equals $SPACF(k)$. As for $LDF(k)$, generally, $PLDF(k)$ cannot be interpreted as a correlation. However, $PLDF(k)$ can be interpreted as (the signed square-root of) the relative decrease in one-step prediction variance when lag k is included as an additional predictor. For $k = 1$ the model (7) corresponding to $k-1$ reduce to an overall mean and the R-squared value in (8) is thus $\tilde{R}^2_{0(1)}$, whereby $PLDF(1) = LDF(1)$ if the same smoother is used for both functions. It can be noticed that the same relation exists between the partial autocorrelation function and the autocorrelation function. For $k = 0$ the partial lag dependence function is set equal to one.

Except for the sign $PLDF(k)$ may also be based on the completely general autoregressive model

$$x_t = g_k(x_{t-1}, \ldots, x_{t-k}) + e_t \tag{9}$$

where $g : \mathbb{R}^k \rightarrow \mathbb{R}$. However, the estimation of $g_k(\cdot, \ldots, \cdot)$ without other than an assumption of smoothness is not feasible in practice for k larger than, say, three, see also (Hastie & Tibshirani 1990). Recently, alternatives to (9) has been considered by Lin & Pourahmadi (1998).

5.1 Fitting the additive models

To fit the non-linear additive autoregressive model (7) the backfitting algorithm (Hastie & Tibshirani 1990) is suggested. However, concurvity (Hastie & Tibshirani 1990) between the lagged values of the time series may exist and, hence, the estimates may not be uniquely defined. By including the lags sequentially, this is only an indication of no additional predictive ability of the most recently included lag. For this reason it is suggested to fit models of increasing order, starting with $k = 1$ and ending with the highest lag K for which $PLDF(k)$ is to be calculated. In the calculation of the residual sum of squares only residuals corresponding to $t = K + 1, \ldots, N$ should be used.

For the numerical examples considered in this paper local polynomial regression (Cleveland & Devlin 1988) is used for smoothing. The convergence criterion used is the maximum absolute change in any of the estimates relative to the range of the fitted values. An iteration limit is applied as a simple test for convergence.

For $k = 1$ the estimation problem reduces to local polynomial regression and hence convergence is guaranteed. If for any $k = 2, \ldots, K$ convergence is not obtained, or if the residual sum of squares increases compared to the previous lag, we put $\hat{f}_{jk}(\cdot) = 0$, $(j = k, \ldots, K)$ and $\hat{f}_{kj}(\cdot) = \hat{f}_{k-1,j}(\cdot)$, $(j = 1, \ldots, k-1)$. This ensures that convergence is possible for $k + 1$.

6 Strictly non-linear lag dependence

The lag dependence function described in Section 4 measures both linear and non-linear dependence. If, in the definition of $\tilde{R}^2_{0(k)}$, the sum of squares from a overall mean SS_0 is replaced by the sum of squares from fitting a strait line to the k-lagged scatter plot, a measure of non-linearity is obtained. In this paper this will be called the strictly Non-linear Lag Dependence Function in lag k, or $NLDF(k)$.

7 Confidence intervals

Smoothers usually require one or more smoothing parameters to be selected, see e.g. (Hastie & Tibshirani 1990, Chapter 3). Therefore, in principle, smoothing parameters can be selected to obtain R-squared values arbitrarily close to one, also when the underlying process is i.i.d. (assuming no ties are present in the data). For this reason it is important to obtain confidence intervals for, e.g., the lag dependence function under the hypothesis that the underlying process is i.i.d. and for a given set of smoothing parameters. Furthermore, it is applicable to calculate a confidence interval under a hypothesis of linearity for the strictly non-linear lag dependence function. These aspects are considered in this section.

As indicated above it is clear that the range of the confidence intervals will depend on the flexibility of the smoother. To detect a general non-linearity a flexible smoother must be used whereby the range of the confidence interval will be increased compared to the case where we are only intrested in detecting minor depertures from linearity or departures in the direction of near-global higher order polynomials. Thus, the bandwidth of the smoother can be used to adjust the properties of the test. It is recomended to apply the methods using a range of bandwidths and smoothers. These aspects are exemplified in Sections 8.1 and 8.3.

Under the hypothesis that the time series $\{x_1, \dots, x_N\}$ is observations from an i.i.d. process the distribution of any of the quantities discussed in the previous sections can be approximated by generating a large number of i.i.d. time series of length N from an estimate of the density function of the process and recalculating the quantities for each of the generated time series. Methods as outlined above are often denoted bootstrap methods and in this context various approaches to the calculation of approximate confidence intervals have been addressed extensively in the literature, see e.g. (Efron & Tibshirani 1993). In all but one of the examples considered in this paper the empirical density function is used. However, for short time series it may be more appropriate to condition on a parametric form of the density function.

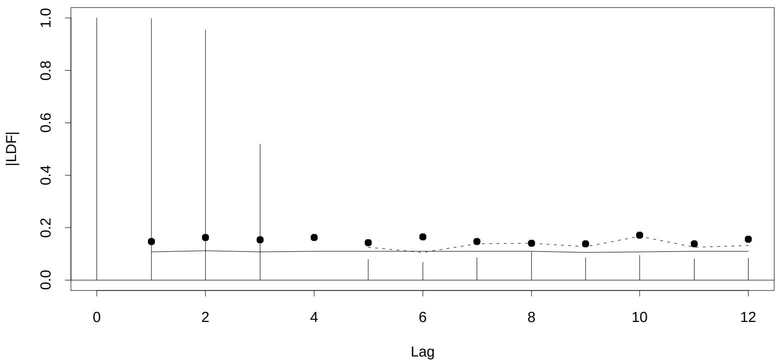
7.1 Confidence limit for $|LDF(k)|$

Calculation of $LDF(k)$ involves only scatter plot smoothing and, thus, it is faster to calculate than, e.g., $PLDF(k)$. For this reason we shall first consider $LDF(k)$ for some range $k = 1, \ldots, K$. For an i.i.d. process it is obvious that the distribution of $LDF(k)$ will depend on k only due to the fact that k affects the number of points on the k-lagged scatter plot. Hence, when $k \ll N$ the distribution of $LDF(k)$ under the hypothesis of independence is approximately independent of k.

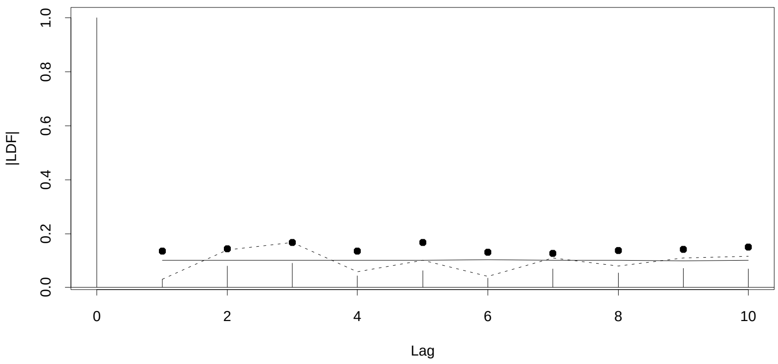
The sign in the definition of $LDF(k)$ is included only to establish an approximate equality with $SACF(k)$ when linear models are used and to include information about the sign of the average value of the slope. When the observations originates from an i.i.d. process $LDF(k)$ will be positive with probability 1/2. Consequently, when the smoother is flexible enough the null-distribution of $LDF(k)$ will be bimodal, since in this case $\tilde{R}^2_{0(k)}$ will be strictly positive. The most efficient way of handling this problem is to base the bootstrap calculations on the absolute value of $LDF(k)$. Hence, an upper confidence limit on $|LDF(k)|$ is to be approximated.

Below the standard, percentile, and BC_a methods, all defined in (Efron & Tibshirani 1993, Chapters 13 and 14), will be briefly discussed. For the series considered in Figure 1 the $LDF(k)$ were calculated for $k \leq 12$ using a local linear smoother and a nearest neighbour bandwidth of 1/3. The result is shown in Figure 2a together with 95% bootstrap confidence limits calculated separately for each lag and based on 1000 bootstrap replicates, generated under the hypothesis of independence. The BC_a limit could not be calculated for lags 1 to 4, since all the bootstrap replicates were either smaller or larger than the actual value of $|LDF(k)|$. Results corresponding to Figure 2a when the true process is standard Gaussian i.i.d. are shown in Figure 2b. For practical purposes an equality of the standard and percentile methods are observed (no difference is visible on the plots), whereas the results obtained using the BC_a method is highly dependent on the lag through the value of $|LDF(k)|$. Hence, the BC_a method cannot be used when the confidence limit is only calculated for one lag and used for the remaining lags as outlined above. The high degree of correspondence between the standard and percentile method indicates that sufficient precision can be obtained using the standard method on fever bootstrap replicates. This is highly related to the approximate normality of $|LDF(k)|$ and it is suggested that this is investigated for each application before a choice between the standard and percentile method is made.

The underlying model of the BC_a method assumes that the estimate in question may be biased and that the variance of the estimate depends linearly on an increasing transformation of the true parameter (Efron & Tibshirani 1993, p. 326-8), and furthermore the estimate is assumed to be normally distributed. The bias and the slope of the line are then estimated from the data. With λ being the fraction of the bootstrap replicates strictly below the original estimate, the bias is $\Phi^{-1}(\lambda)$ (Φ is the cumulative standard Gaussian distribution function). This explains why the BC_a limit is non-existing for lags 1-4 of the deterministic series. The slope is estimated by use of the jackknife procedure (Efron & Tibshirani 1993, p. 186). It seems that, although the underlying model of the BC_a method is a superset of the underlying model of the standard method, the estimation of bias and slope induces some additional variation in the confidence limit obtained. As a consequence it may be advantageous to average the BC_a limits over the lags and use this value instead of the individual values. However, the standard and percentile methods seem to be appropriate for this application and since significant savings of computational effort can be implemented by use of these methods it is suggested that only these are applied on a routine basis.



(a) Deterministic



(b) Standard Gaussian i.i.d.

Figure 2: Absolute value of the Lag Dependence Function of the deterministic series presented in Figure 1 and of 1000 observations from a standard Gaussian i.i.d. process. The dots indicate the maximum over the 1000 bootstrap replicates. Standard, percentile, and BC_a 95% confidence limits are indicated by lines (BC_a dotted).

7.2 Confidence limit for $|PLDF(k)|$

In Section 7.1 it is shown how bootstrapping can be used to construct an approximative confidence limit for $|LDF(k)|$. There is some indication that this limit can be used also for $|PLDF(k)|$ if the same smoother is used for calculation of $LDF(k)$ and $\hat{f}_{k1}(\cdot), \ldots, \hat{f}_{kk}(\cdot)$ (Sections 4 and 5).

For (linear) autoregressive models of order p, with i.i.d. $N(0, \sigma^2)$ errors, and fitted using N observations it holds that the residual sum of squares is distributed as $\sigma^2\chi^2(N-p)$ (Brockwell & Davis 1987, p. 251 and 254). We can relax the normality assumption and conclude that if the true process is i.i.d. with variance σ^2 the following approximations apply when linear autoregressive models are used

$$SS_0 \sim \sigma^2\chi^2(N-1) \tag{10}$$
$$SS_{0(k)} \sim \sigma^2\chi^2(N-2) \tag{11}$$
$$SS_{0(1\ldots k-1)} \sim \sigma^2\chi^2(N-k) \tag{12}$$
$$SS_{0(1\ldots k)} \sim \sigma^2\chi^2(N-k-1) \tag{13}$$

For $N \gg k$ the distribution of all four sums of squares are approximately equal.

For locally weighted regression Cleveland & Devlin (1988) stated that the distribution of the residual sum of squares can be approximated by a constant multiplied by a χ^2 variable, see also (Hastie & Tibshirani 1990, Section 3.9). Furthermore, for generalized additive models Hastie & Tibshirani (1990, Section 8.1) uses a χ^2 distribution with degrees of freedom equal to the number of observations minus a quantity depending on the flexibility of the smoothers used.

For these reasons we conjecture that when $N \gg k$ and when the same smoother is used for $LDF(k)$ and $PLDF(k)$, as outlined in the beginning of this section, then the sum of squares SS_0, $\widetilde{SS}_{0(k)}$, $\widetilde{SS}_{0(1\ldots k-1)}$, and $\widetilde{SS}_{0(1\ldots k)}$ will follow approximately the same distribution.

This conjecture leads to approximate equality of means and variances of the sums of squares. Since for both $LDF(k)$ and $PLDF(k)$ the compared models differ by an additive term, estimated by the same smoother in both cases, we also conjecture that for an i.i.d. process.

$$Cor[\widetilde{\widetilde{SS}}_{0(k)}, SS_0] \approx Cor[\widetilde{\widetilde{SS}}_{0(1...k)}, \widetilde{\widetilde{SS}}_{0(1...k-1)}]. \tag{14}$$

Using linearizations about the mean of the sums of squares it then follows from the approximate equality of means that

$$E[|LDF(k)|] \approx E[|PLDF(k)|], \tag{15}$$

and from both conjectures that

$$V[|LDF(k)|] \approx V[|PLDF(k)|]. \tag{16}$$

Eqs. (15) and (16) tell us that the approximate i.i.d. confidence limit obtained for $|LDF(k)|$ can be used also as an approximate limit for $|PLDF(k)|$. In Section 8 (Canadian lynx data) an example of the quality of the approximation is given, and the mentioned arguments seems to be confirmed by the bootstrap limits obtained in that example.

7.3 Confidence limit for $|NLDF(k)|$

Assuming a specific linear model this can be used for simulation and an approximate bootstrap confidence limit for $|NLDF(k)|$ can be obtained given this model. Consequently, the alternative contains both linear and non-linear models. To make the approach sensible the linear model needs to be appropriately selected, i.e. using the standard time series tools of identification, estimation, and validation. When the parametric bootstrap is applied the procedure outlined above is an example of the procedures considered by Tsay (1992).

Alternatively, the simulations can be performed using autocovariances only and assuming these to be zero after a specific lag. In this case an estimator of autocovariance must be used that ensures that the autocovariance function used for simulation is non-negative definite, see e.g. (Brockwell & Davis 1987, p. 27). Note that using this approach on the series considered in Figure 1 will, essentially, result in a test for independence.

Hjellvik & Tjøstheim (1996) consider a similar test for linearity and uses Akaike's information criterion (Brockwell & Davis 1987) to select an appropriate $AR(p)$-model under which the bootstrap replicates are generated. In (Theiler, Eubank, Longtin, Galdrikian & Farmer 1992) a range of alternative linear null hypotheses is considered. Especially, the random sampling in the phase spectrum described in Section 2.4.1 of this reference seems to be a relevant linear null hypothesis.

8 Examples

8.1 Linear processes

Below it is briefly illustrated how LDF and $PLDF$ behaves compared to $SACF$ and $SPACF$ in case of simple linear processes. The $AR(2)$ process

$$X_t = 1.13X_{t-1} - 0.64X_{t-2} + e_t \tag{17}$$

and the $MA(2)$ process

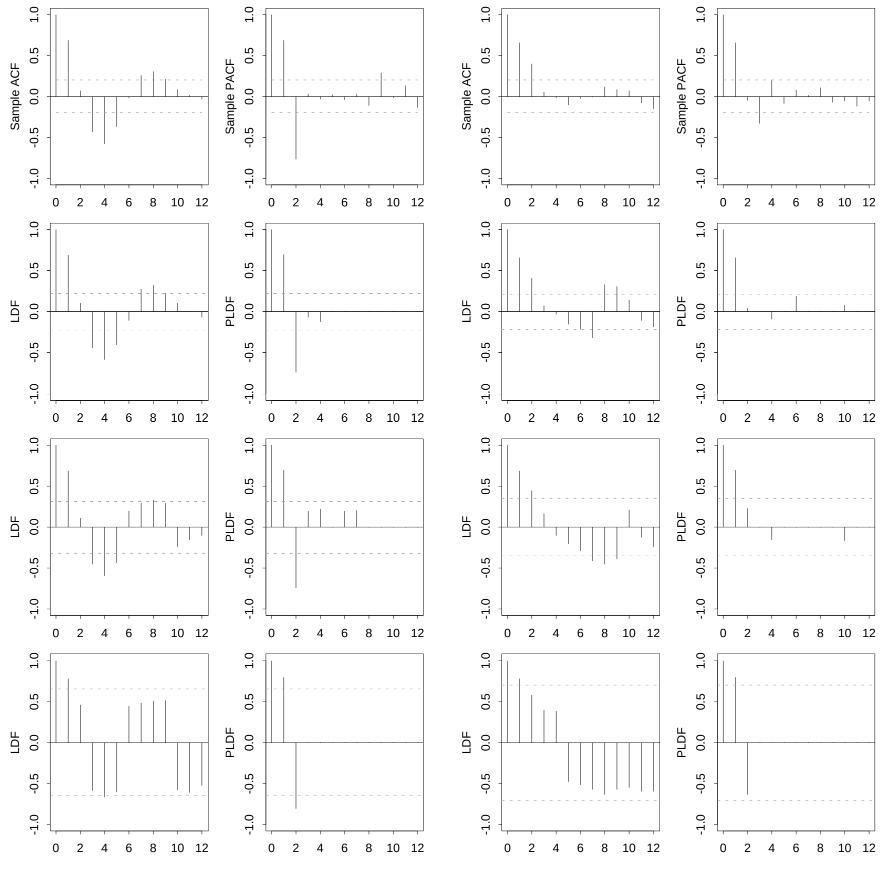
$$X_t = e_t + 0.6983e_{t-1} + 0.5247e_{t-2} \tag{18}$$

are considered, where in both cases $\{e_t\}$ is i.i.d. $N(0,1)$.

Figure 3 contain plots based on 100 simulated values from (17) and (18), respectively (the default random number generator of S-PLUS version 3.4 for HP-UX were used). Each figure show $SACF$ and $SPACF$. The remaining plots are LDF and $PLDF$ for local linear smoothers using a nearest neighbour bandwidth of 1.00 (2nd row), 0.50 (3rd row), and 0.1 (bottom row). 95% confidence intervals are indicated by dotted lines. The confidence intervals obtained for LDF are included on the plots of $PLDF$.

For the calculation of $PLDF$ a convergence criterion (see Section 5.1) of 0.01 and an iteration limit of 20 is used. Standard bootstrap intervals are calculated for LDF under the i.i.d. hypothesis using 200 replicates. For LDF the agreement with $SACF$ is large for nearest neighbour bandwidths 1.0 and 0.5. As expected, the range of the confidence interval increases with decreasing bandwidth, and, using the smallest bandwidth, it is almost not possible to reject the i.i.d. hypothesis, c.f. the arguments mentioned in the beginning of Section 7.

When a nearest neighbour bandwidth of 1.0 is used $PLDF$ agrees well with $SPACF$ for the lower half of the lags, whereas $PLDF$ is exactly zero for most of the larger half of the lags. Similar comments apply for nearest neighbour bandwidths 0.5 and 0.1. This is due to the function estimates being set equal to zero when the iteration limit is reached.



(a) $AR(2)$ process (b) $MA(2)$ process

Figure 3: Plots of autocorrelation functions and their generalizations for 100 observations from the $AR(2)$ process (17) and the $MA(2)$ process (18).

8.2 Non-linear processes

Three non-linear processes are addressed, namely (i) the non-linear autoregressive process ($NLAR(1)$)

$$X_t = \frac{1}{1 + \exp(-5X_{t-1} + 2.5)} + e_t, \tag{19}$$

where $\{e_t\}$ i.i.d. $N(0, 0.1^2)$, and (ii) the non-linear moving average process ($NLMA(1)$)

$$X_t = e_t + 2\cos(e_{t-1}), \tag{20}$$

where $\{e_t\}$ i.i.d. $N(0, 1)$ and (iii) the non-linear and deterministic process described in Section 2, called $DNLAR(1)$ in the following. For all three cases 1000 observations are generated. The starting value for $NLAR(1)$ is set to 0.5 and for $DNLAR(1)$ it is set to 0.8. Plots of the series $NLAR(1)$ and $NLMA(1)$ are shown in Figure 4. The plot of $DNLAR(1)$ is shown in Figure 1.

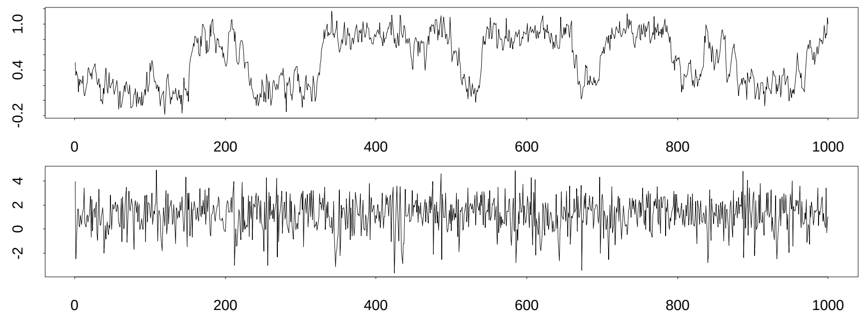


Figure 4: Plots of the series $NLAR(1)$ (top) and $NLMA(1)$ bottom.

For the calculation of LDF, $PLDF$, and $NLDF$ a local linear smoother with a nearest neighbour bandwidth of 0.5 is used. Actually, lagged scatter plots indicate that a local quadratic smoother should be applied, at least for $NLMA(1)$ and $DNLAR(1)$, but to avoid a perfect fit for the deterministic series a local linear smoother is used. Confidence intervals are constructed using standard normal intervals, since normal QQ-plots of the absolute values of the 200 bootstrap replicates showed this to be appropriate. The confidence interval obtained for LDF is included on the plots of $PLDF$.

Figure 5 shows $SACF$, $SPACF$, LDF, and $PLDF$ for the three series. For $NLMA(1)$ and $DNLAR(1)$ the linear tools, $SACF$ and $SPACF$, indicate independence and LDF shows that lag dependence is present. From these observations it can be concluded that $NLMA(1)$ and $DNLAR(1)$ are nonlinear processes. From the plots of LDF and $PLDF$ it cannot be inferred whether $NLMA(1)$ is of the autoregressive or of the moving average type. For $DNLAR(1)$ the autoregressive property is more clear since $PLDF$ drops to exactly zero after lag two. In case of $DNLAR(1)$ a more flexible smoother will result in values of LDF being significantly different from zero for lags larger than two, while, for $NLMA(1)$, LDF will be close to zero for lags larger than one independent of the flexibility of the smoother used. This is an indication of $DNLAR(1)$ being of the autoregressive type and $NLMA(1)$ being of the moving average type.

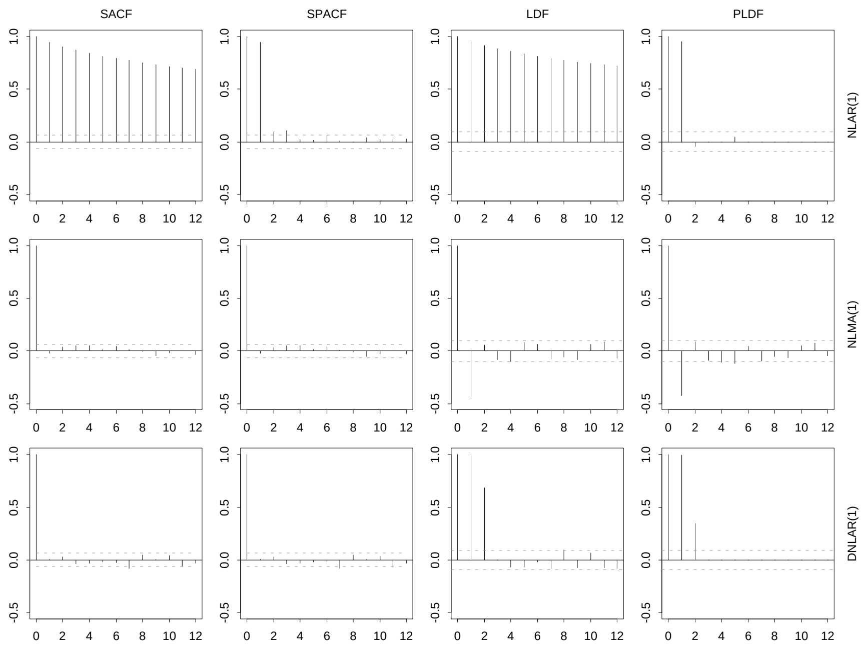


Figure 5: $SACF$, $SPACF$, LDF, and $PLDF$ (columns, left to right) for series $NLAR(1)$, $NLMA(1)$, and $DNLAR(1)$ (rows, top to bottom).

For $NLAR(1)$ the linear tools indicate that the observations come from an $AR(1)$ process. This is not seriously contradicted by LDF or $PLDF$, although LDF decline somewhat slower to zero than $SACF$. To investigate if the underlying process is linear a Gaussian $AR(1)$ model is fitted to the data and this model is used as the hypothesis under which 200 (parametric) bootstrap replicates of $NLDF$ are generated. Figure 6 shows $NLDF$ and a 95% standard normal interval, constructed under the hypothesis mentioned above. A normal QQ-plot show that the absolute values of the bootstrap replicates are approximately Gaussian. From Figure 6 it is concluded that the underlying process is not the estimated $AR(1)$-model, and based on $PLDF$ it is thus concluded that the observations originate from a non-linear process of $AR(1)$ type.

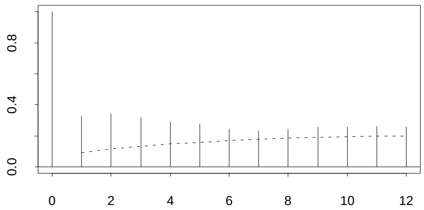


Figure 6: $NLDF$ for $NLAR(1)$, including a 95% confidence interval under the assumption of an $AR(1)$ process (dotted).

8.3 Canadian lynx data

Recently, Lin & Pourahmadi (1998) analyzed the Canadian lynx data (Moran 1953) using non-parametric methods similar to the methods presented in this paper. The data is included in the software S-PLUS (version 3.4 for HP-UX) and described in (Tong 1990, Section 7.2). In this paper a thorough analysis of the data will not be presented, but the data will be used to illustrate how the methods suggested can be applied. As in (Lin & Pourahmadi 1998) the data is $\log_{10}$-transformed prior to the analysis.

For the transformed data LDF, $PLDF$, and $NLDF$ are computed using a local quadratic smoother and nearest neighbour bandwidths of 0.5 and 1. For LDF 200 bootstrap replicates are generated under the i.i.d. hypothesis and QQ-plots indicate that standard normal intervals are appropriate. The same apply for $NLDF$ with the exception that the bootstrap replicates are generated under the hypothesis that the $AR(2)$ model of Moran (1953), also described by Lin & Pourahmadi (1998), is true. Confidence intervals are computed also for $PLDF$ for the nearest neighbour bandwidth of 1.0. The intervals are based one hundred bootstrap replicates of $PLDF$ generated under the i.i.d. hypothesis. QQ-plots indicate that the percentile method should be applied to the absolute values of $PLDF$.

In Figure 7 plots of LDF, $NLDF$, and $PLDF$ are shown. Dotted lines indicates 95% confidence intervals under the i.i.d. hypothesis (LDF) and under the $AR(2)$ model of Moran (1953) ($NLDF$). The intervals obtained for LDF are also shown on the plots of $PLDF$. Furthermore, for the nearest neighbour bandwidth of 1.0, a 95% confidence interval for white noise is included on the plot of $PLDF$ (solid lines).

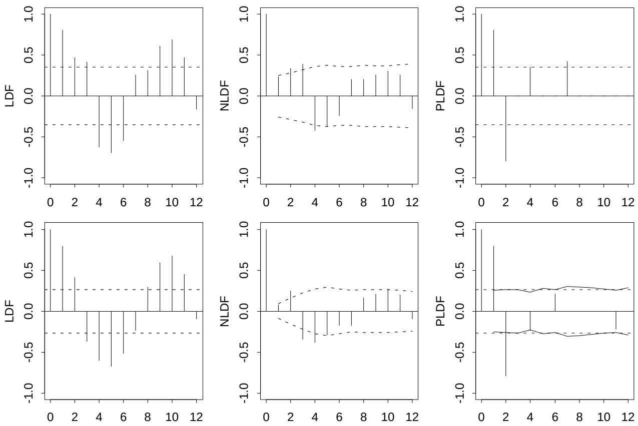


Figure 7: Canadian lynx data ($\log_{10}$-transformed). Plots of LDF, $NLDF$, and $PLDF$ using local quadratic smoothers and nearest neighbour bandwidths 0.5 (top row) and 1.0 (bottom row).

From the plots of LDF it is clearly revealed that the process is not i.i.d. The plots of $NLDF$ for a nearest neighbour bandwidth of 0.5 show hardly any significant values, but when a nearest neighbour bandwidth of 1.0 is used lags two, three, and four show weak significance. This indicates that a departure from linearity in the direction of an almost quadratic relationship is present in the data. See also the comments about the flexibility of smoothers in the beginning of Section 7. Finally, the plots of $PLDF$ clearly illustrate that lag one and two are the most important lags and that other lags are, practically, non-significant. In conclusion, an appropriate model seems to be a non-linear autoregressive model containing lag one and two, i.e. a model of the type (7) with $k = 2$.

Estimation in this model using local quadratic smoothers and a nearest neighbour bandwidth of 1.0 yields the results shown in Figure 8. The response for lag one seems to be nearly linear. This aspect should be further investigated. The results agree well with the results of Lin & Pourahmadi (1998).

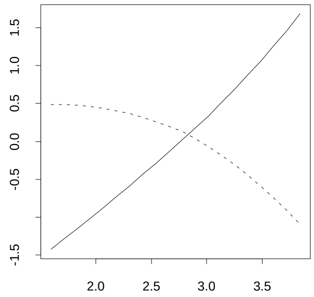


Figure 8: Non-linear additive autoregressive model for the $\log_{10}$-transformed Canadian lynx data ($\hat{f}_{21}(\cdot)$ solid, $\hat{f}_{22}(\cdot)$ dotted). The estimate of the constant term is 2.76 and the MSE of the residuals is 0.0414.

9 Lagged cross dependence

Given two time series $\{x_1, \ldots, x_N\}$ and $\{y_1, \ldots, y_N\}$ the Sample Cross Correlation Function between processes $\{X_t\}$ and $\{Y_t\}$ in lag k ($SCCF_{xy}(k)$) is an estimate of the correlation between X_{t-k} and Y_t. It is possible to generalized this in a way similar to the way LDF is constructed. Like $SCCF$ this generalization will be sensible to autocorrelation, or lag dependence, in $\{X_t\}$ in general. For $SCCF$ this problem is (approximately) solved by prewhitening (Brockwell & Davis 1987, p. 402). However, prewhitening is very dependent on the assumption of linearity, in that it relies on the impulse response function from the noise being independent on the level. For this reason, in the non-linear case, it is not possible to use prewhitening and the appropriateness of the generalization of $SCCF$ depend on $\{X_t\}$ being i.i.d.

10 Final remarks

The generalizations of the sample correlation functions reduce to their linear counterpart when the smoothers are replaced by linear models. Hence, if a local linear smoother is applied an almost continuous transition from linear to non-linear measures of dependence is obtainable via the bandwidth of the smoother. It is noted that the partial lag dependence function, and its linear counterpart, in lag k compares the residual sum of squares of a model containing lags $1, \ldots, k$ relatively to the case where lag k is omitted. When building models aimed at prediction it might be more informative to use a quantity depending on differences in residual sum of squares. Since $\tilde{R}^2_{0(1\ldots k)} - \tilde{R}^2_{0(1\ldots k-1)}$ is the normalized reduction in the (in-sample) one-step prediction error variance when including lag k as a predictor this quantity could be used instead of $\tilde{R}^2_{(0k)|(1\ldots k-1)}$ in (8).

Optimal bandwidth selection is not addressed in this paper. However, the methods can still be applied in this case, but the power against specific alternatives cannot be adjusted. Furthermore, the methods are not restricted to the use of non-parametric methods. Any procedure of generating fitted values uniquely identified by the lag(s) included may be applied. However, such procedures may require special considerations regarding confidence intervals.

If the conditional mean of the series can be modelled the methods described in this paper can be applied to the series of squared residuals and the conditional variance can, possibly, be addressed in this way. This approach is similar to the approach by Tjøstheim & Auestad (1994, Section 5).

References

Anderson-Sprecher, R. (1994), 'Model comparisons and R^2', *The American Statistician* **48**, 113–117.

Brockwell, P. J. & Davis, R. A. (1987), *Time Series: Theory and Methods*, Springer-Verlag, Berlin/New York.

Chen, R. & Tsay, R. S. (1993), 'Nonlinear additive ARX models', *Journal of the American Statistical Association* **88**, 955–967.

Cleveland, W. S. & Devlin, S. J. (1988), 'Locally weighted regression: An approach to regression analysis by local fitting', *Journal of the American Statistical Association* **83**, 596–610.

Efron, B. & Tibshirani, R. J. (1993), *An Introduction to the Bootstrap*, Chapman & Hall, London/New York.

Ezekiel, M. & Fox, K. A. (1959), *Methods of Correlation and Regression Analysis*, third edn, John Wiley & Sons, Inc., New York.

Friedman, J. H. (1991), 'Multivariate adaptive regression splines', *The Annals of Statistics* **19**, 1–67. (Discussion: p67-141).

Granger, C. & Lin, J.-L. (1994), 'Using the mutual information coefficient to identify lags in nonlinear models', *Journal of Time Series Analysis* **15**, 371–384.

Granger, C. W. J. (1983), Forecasting white noise, *in* A. Zellner, ed., 'Applied Time Series Analysis of Economic Data', U.S. Department of Commerce, Washington, pp. 308–314.

Hastie, T. J. & Tibshirani, R. J. (1990), *Generalized Additive Models*, Chapman & Hall, London/New York.

Hinich, M. J. (1979), 'Estimating the lag structure of a nonlinear time series model', *Journal of the American Statistical Association* **74**, 449–452.

Hjellvik, V. & Tjøstheim, D. (1996), 'Nonparametric statistics for testing of linearity and serial independence', *Journal of Nonparametric Statistics* **6**, 223–251.

Kendall, M. & Stuart, A. (1961), *Inference and Relationship*, Vol. 2 of *The Advanced Theory of Statistics*, first edn, Charles Griffin & Company Limited, London.

Kvålseth, T. O. (1985), 'Cautionary note about R^2', *The American Statistician* **39**, 279–285.

Lewis, P. A. W. & Stevens, J. G. (1991), 'Nonlinear modeling of time series using multivariate adaptive regression splines (MARS)', *Journal of the American Statistical Association* **86**, 864–877.

Lin, T. C. & Pourahmadi, M. (1998), 'Nonparametric and non-linear models and data mining in time series: a case-study on the Canadian lynx data', *Applied Statistics* **47**, 187–201.

Moran, P. A. P. (1953), 'The statistical analysis of the sunspot and lynx cycles', *Journal of Animal Ecology* **18**, 115–116.

Priestley, M. B. (1988), *Non-linear and Non-stationary Time Series Analysis*, Academic Press, New York/London.

Rao, C. R. (1965), *Linear Statistical Inference and its applications*, Wiley, New York.

Teräsvirta, T. (1994), 'Specification, estimation, and evaluation of smooth transition autoregressive models', *Journal of the American Statistical Association* **89**, 208–218.

Theiler, J., Eubank, S., Longtin, A., Galdrikian, B. & Farmer, J. D. (1992), 'Testing for nonlinearity in time series: the method of surrogate data', *Physica D* **58**, 77–94.

Tjøstheim, D. (1994), 'Non-linear time series: A selective review', *Scandinavian Journal of Statistics* **21**, 97–130.

Tjøstheim, D. & Auestad, B. H. (1994), 'Nonparametric identification of nonlinear time series: Selecting significant lags', *Journal of the American Statistical Association* **89**, 1410–1419.

Tong, H. (1990), *Non-linear Time Series. A Dynamical System Approach*, Oxford University Press, Oxford.

Tsay, R. S. (1992), 'Model checking via parametric bootstraps in time series analysis', *Applied Statistics* **41**, 1–15.

Whitaker, J. (1990), *Graphical Models in Applied Multivariate Statistics*, Wiley, New York.

Paper H

Wind power prediction using ARX models and neural networks

Originally published as

> Henrik Aalborg Nielsen and Henrik Madsen. Wind power prediction using ARX models and neural networks. In M. H. Hamza, editor, *Proceedings of the Fifteenth IASTED International Conference on Modelling, Identification and Control*, pages 310–313, February 1996.

More details on the application of the neural networks can be found in

> Henrik Aalborg Nielsen and Henrik Madsen. Neural networks for wind power prediction. In Henrik Madsen, editor, *Models and Methods for Predicting Wind Power*, pages 51–58. ELSAM, Skærbæk, Denmark, 1996. ISBN 87-87090-29-5.

Wind power prediction using ARX models and neural networks

Henrik Aalborg Nielsen[1] and Henrik Madsen[1]

Abstract

Prediction of the wind power production at a wind farm placed near the west coast of Denmark is considered. The wind farm consists of 27 wind mills each with a power capacity of 225kW. Based on previous work regarding autoregressive models with external signals (ARX models), models based on feed-forward neural networks with one hidden layer are formulated. The size of the network is determined by the Bayes Information Criterion. The prediction performance of the selected networks are compared with the performance of the ARX models. Furthermore the naive predictor has been used as a reference of prediction performance. The criterion used for evaluating the prediction performance is the Root Mean Square of the prediction errors.
For most horizons three to four hidden units are found optimal with respect to the Bayes Information Criterion. Comparing the optimal neural network predictors with the ARX-based and naive predictors it is concluded that the neural network type investigated is inferior in prediction performance to the other prediction procedures investigated. Finally neural networks with only one hidden unit has been compared with the other prediction procedures. Also these networks prove to be inferior.

Keywords: Wind power, prediction, neural networks.

1 Introduction

In Denmark wind energy is becoming of increasing importance and hence it is important to be able to perform short term predictions of the wind

[1]Department of Mathematical Modelling, Technical University of Denmark, DK-2800 Lyngby, Denmark

power production. Adaptive prediction procedures based on autoregressive models with external signals (ARX models) have been developed and implemented for on-line wind power prediction in the western part of Denmark, see (ELSAM 1995). In this paper predictors based on neural networks are compared with ARX-based predictors. Furthermore the naive predictor, which corresponds to predicting the future value as the most resent observed value, has been used as a reference of prediction performance. The data used is half-hourly averages of wind speed and wind power production. Prediction horizons from 30 minutes to 3 hours are considered.

The paper contains a brief description of the ARX models. These models use the wind speed and a diurnal profile (representing a time-varying mean) as inputs. The parameters of the models are estimated using the adaptive least squares method with exponential forgetting (Ljung 1987). The estimation method is modified in (ELSAM 1995) to handle multi-step predictions.

A feed-forward neural network with one hidden layer has been used. This kind of network is described in (Ripley 1993) and a software package for S-Plus is available, see (Ripley 1994). The networks considered all use the same variables as the ARX models. In this context the potential advantage of neural networks over ARX models is a more adequate description of any non-linear relationships between the variables. The disadvantages are a larger number of parameters and the non-adaptive estimation.

2 Predictors based on ARX models

In (ELSAM 1995) a careful investigation of the problem of wind power prediction for the ELSAM (power distributor for the western part Denmark) area are described. Based on this investigation the following models has been implemented and used for k-step wind power predictions:

$$\begin{aligned}\sqrt{p_{t+k}} &= \mu^k + a_1^k\sqrt{p_t} + b_1^k\sqrt{w_t} + b_2^k w_t + \\ &\quad c_1^k \sin\tfrac{2\pi h_{t+k}}{24} + c_2^k \cos\tfrac{2\pi h_{t+k}}{24} + e_{t+k}^k,\end{aligned} \tag{1}$$

where μ^k, a_1^k, $b_\cdot^k$, and $c_\cdot^k$ are constants. The symbols p_t and w_t represent average wind power production and wind speed in the interval $[t-1, t[$. h_t is the 24-hour clock at time t. $\{e_t^k\}$ is a sequence of independent identically distributed random variables with zero mean and variance σ_k^2. One time step corresponds to 30 minutes.

The parameters of the models (1) are estimated adaptively using recursive least squares with exponential forgetting, see (Ljung 1987). The algorithm has, however, been modified in order to handle multi-step predictions. This modification consists of updating only the most resent parameter estimate, say $\hat{\theta}_{t-1}$. In order to make this feasible a pseudo prediction of $\sqrt{p_t}$ is used in the update of parameters; this prediction is constructed from p_{t-k}, w_{t-k}, h_t, and $\hat{\theta}_{t-1}$. See (ELSAM 1995) for further details. A forgetting factor of 0.999, as suggested in (ELSAM 1995), is used in this paper.

3 Neural networks

3.1 Type of neural network

A feed-forward neural network with one hidden layer and without connections directly from input to output is used, see e.g. the documentation on the software (Ripley 1994). Suppose that observations (indexed by i) of the independent variables (indexed by j) x_{ij} and the dependent variable y_i are present. The dependence of y on x can then be modelled by a neural network of the above type as:

$$y_i = \phi_o\left(\alpha_o + \sum_{h=1}^{n_h} w_{ho}\ \phi_h\left(\alpha_h + \sum_{j=1}^{n_j} w_{jh} x_{ij}\right)\right) + e_i, \qquad (2)$$

where $w_{\cdot o}$ are the weights on the connections from the hidden layer to the output layer, $w_{\cdot h}$ are the weights on the connections from the input layer to unit h in the hidden layer, α_o is the bias on the output unit, and α_h is the bias on the hidden units. n_h and n_j are the No. of hidden units and inputs, respectively. It is seen that the weights and the biases are just parameters of the model. Considering (2) as a statistical model one would assume that the residuals (e_i) are independent identical distributed.

The functions $\phi_h(\cdot)$ and $\phi_o(\cdot)$ are predefined functions associated with the units in the hidden and output layer, respectively. Most frequently these functions are sigmoid (also called logistic), i.e. the output of the network is restricted to the interval $]0,1[$. This is, however, not desirable in this application since the future output of the network is then restricted to the range of observations in the data set used for estimating the parameters. For this reason the output unit has been chosen to be linear, i.e. $\phi_o(z) = z$.

3.2 Estimation of parameters

The parameters (weights and biases) of the model (2) are estimated by non-linear least-squares. The initial values of the estimates are rather important since the minimization problem may contain local minima due to the fact that the model is non-linear in the parameters. For this reason each model should be estimated several times using different initial parameter estimates.

Since it is rather difficult to suggest appropriate values it seems reasonable to select these values at random. In this case the data has been scaled to the interval $[0,1]$ (see Section 4) and according to the documentation on the software used (see Section 3.4 and (Ripley 1994)) it should be sufficient to sample from the $U(-1,1)$ distribution. In spite of this it was decided to sample from the $U(-5,5)$ distribution in order to cover a wider interval of initial parameter estimates. The parameters of each network were estimated 20 times with initial parameter estimates chosen at random.

3.3 Selection of network size

To use a neural network it remains to decide upon the independent variables to include in the model and on the No. of hidden units. This may be done by using some kind of information criteria. Here the Bayes Information Criterion (BIC) has been used, see (Schwarz 1978). With L^*, n_p, and N being the value of the likelihood in the optimum, the No. of parameters, and the No. of observations used in the estimation, respectively, the criteria corresponds to chose the model so that

$$\log L^* - \frac{n_p}{2} \log N, \tag{3}$$

is maximized. For a large class of linear time series models and other linear models with the residuals being normally distributed the criteria is equivalent to minimizing

$$BIC = N \log \left(\frac{1}{N} \sum_{i=1}^{N} \hat{e}_i^2 \right) + n_p \log N, \tag{4}$$

where $\hat{e}_i$ is the prediction errors from the model (2), with the unknown parameters replaced by the estimates. Note that $\hat{e}_i$ must be based on maximum likelihood estimates.

In this case the procedure used for estimation of the parameters (see Section 3.2) is not a maximum likelihood procedure. Hence the above criteria must be considered as an approximation.

It was decided to use the criteria (4) to select the appropriate No. of hidden units only. The independent variables have been considered fixed, see Section 4.

3.4 Software

The neural network software used is written by Professor of Applied Statistics, B.D. Ripley, University of Oxford. The software can be obtained from StatLib by anonymous ftp from `lib.stat.cmu.edu`. The software is written for S-Plus and briefly described in (Ripley 1994).

The adaptive predictions have been calculated using the software Off-line Wind Power Prediction Tool, version 1.0, see (Nielsen & Madsen 1995).

4 Variables in the models

Based on the investigation described in (ELSAM 1995) (see also Section 2) it was decided to use the following independent variables: The present wind power production (p_t) scaled to approximately $[0, 1]$, the present wind speed (w_t) scaled to approximately $[0, 1]$, $\frac{1}{2}\sin(2\pi h_{t+k}/24)+\frac{1}{2}$, and $\frac{1}{2}\cos(2\pi h_{t+k}/24)+\frac{1}{2}$. The wind power production k step ahead (p_{t+k}) scaled to approximately $[0, 1]$ was used as the dependent variable.

5 Validation

The models have been validated using a different data set than the one on which the selection of the No. of hidden units and the estimation of parameters is based. This data set is called the validation set.

The neural network model selected for each prediction horizon k has been compared with the naive k-step predictor and with the adaptive predictor described in Section 2.

The estimation and validation set are just two parts of one time series. Therefore it was possible to allow the adaptive predictions to settle before the validation was initiated. This method was chosen since this corresponds to the real application.

Based on the validation set the k-step residuals (or prediction errors) were calculated on the original scale and based on these the Root Mean Square (RMS) were calculated. For the residuals $(r_1, r_2, \ldots, r_N)$ the RMS of the residuals is defined as $\sqrt{\frac{1}{N}\sum_i r_i^2}$.

6 Data

The data used in this investigation has been collected in the Vedersø Kær wind farm in the ELSAM area during the period July 2, 1993, 5.30 p.m. until October 11, 1993, 7 a.m. The original sampling time was 5 minutes. Based on these values half-hourly averages were calculated. The data until September 6 at 7 a.m. (3148 averages) is used for estimation purposes whereas the remaining data (1680 averages) has been used for validation.

The maximum wind speed observed is 15.9 m/s and 75% of the time it did not exceed 8.5 m/s. The corresponding values for the wind power production are 5789 and 1783 kW.

7 Results

7.1 Estimation

In Figure 1 plots of the resulting values of BIC are shown. It is seen that for prediction horizons $k = 1, 2, 3$ the lowest value of BIC is observed for a network with three hidden units. For $k = 4, 6$ a network with four hidden units results in the lowest observed BIC, and for $k = 5$ a network with five hidden units results in a marginally lower BIC, than for $k = 4$.

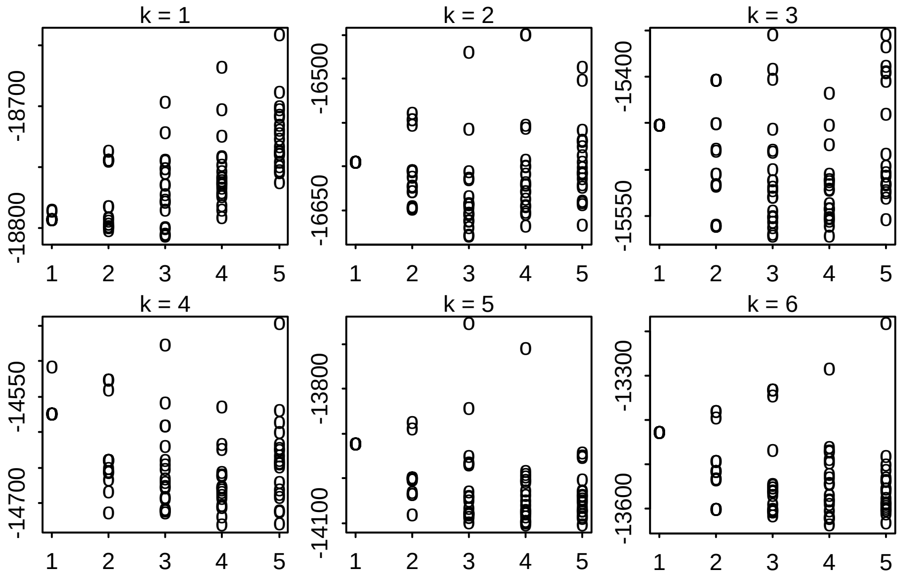


Figure 1: Bayes Information Criterion versus No. of hidden units. A few extreme (high) values are not shown.

7.2 Validation

For all prediction horizons the neural network with the lowest BIC was validated as described in Section 5. The results are shown in Table 1.

k	RMS_{nn} (kW)	RMS_{naive} (kW)	RMS_{adap} (kW)
1	297.6	261.7	262.8
2	448.6	375.7	377.0
3	534.9	440.9	442.0
4	621.2	507.6	507.5
5	678.4	569.2	565.1
6	705.8	623.1	614.1

Table 1: Validation set; RMS of prediction errors of the best neural network, naive, and adaptive predictor.

It is seen that the neural network models investigated are all inferior to both the naive and the adaptive predictors. It is noted that the naive predictor is slightly better than the adaptive for $k = 1, 2, 3$. For $k = 5, 4, 6$ the adaptive predictor is better than the naive. However working out the ratios between RMS_{adap} and RMS_{naive} will reveal that the difference is minor.

7.3 Networks with one hidden unit

From the validation of the network of optimal size it is seen that the naive predictor performs well compared to the other methods investigated. It is therefore peculiar that the selection procedure does not lead to a selection of the most simple network; a network with one hidden unit only. It was therefore decided to compare this kind of network with the network selected according to BIC. Results are indexed by *opt* and 1 for the optimal and the simple network, respectively. The validation results are displayed in Table 2.

From the table it is seen that for $k = 1, 2, 3$ the neural network with one hidden unit actually performs better than the network selected according to BIC. However comparing Table 2 with Table 1 it is seen that the neural network with one hidden unit is inferior to the naive and the adaptive predictor.

k	RMS_{opt} (kW)	RMS_1 (kW)
1	297.6	268.8
2	448.6	411.6
3	534.9	523.3
4	621.2	622.8
5	678.4	699.9
6	705.8	758.3

Table 2: Validation set; comparison of optimal network with a network with one hidden unit.

8 Conclusion

The type of neural networks investigated is inferior in prediction performance to both the adaptive predictor and the simple naive predictor for the prediction horizons investigated (1/2 to 3 hours).

For the prediction horizons investigated the naive predictor performs better than the adaptive predictor for the short prediction horizons (up to $1\frac{1}{2}$ hour). However, the difference between the two predictors is minor. For horizons larger than 2 hours the adaptive predictor is better than the naive.

9 Discussion

Performance of predictors: Apart from the non-linear response of the hidden units, a neural network predictor includes the naive predictor. The reason why the neural network predictor performs considerably worse than the naive and the adaptive predictors is probably that: (i) The estimation of the parameters in the neural network is not adaptive, (ii) the No. of parameters which must be estimated in the neural network is large (seven or larger), and/or (iii) the non-linear response of the hidden units is inappropriate for wind power predictions.

For the low horizons investigated the naive predictor performs slightly better than the adaptive predictor based on the autoregressive model.

For the larger horizons the adaptive predictor is slightly superior. For very large horizons a simple profile (probably containing harmonics corresponding to daily and yearly periods) will probably be the best predictor. The adaptive predictor processes the characteristic of being able to interpolate between these extremes. For this reason the adaptive predictor is attractive.

Maximum size of neural network investigated: The largest No. of hidden units investigated is five. In most cases the optimal network size is found to be less than five. Since the sum of the squared prediction errors for the estimation data set is a non-increasing function of the No. of hidden units BIC will have one minima only. Therefore the maximum size of the networks investigated is sufficient.

Estimation of parameters in neural networks: For one particular neural network model the estimation with random initial values of the parameters results in different values of the mean square of the residuals. This is seen from the random scatter of the values of Bayes Information Criterion (BIC). This clearly reveals that the surface on which the minimization is performed in order to obtain the parameter estimates contains local minima. If this was not true the final estimates and hence BIC should be independent of the initial values of the estimates.

References

ELSAM (1995), *Final Report, Wind Power Prediction Tool in Central Dispatch Centres, JOU2-CT92-0083*, ELSAM.

Ljung, L. (1987), *System Identification, Theory for the User*, Prentice-Hall, Englewood Cliffs, New Jersey.

Nielsen, H. & Madsen, H. (1995), *Off-line Wind Power Prediction Tool - Users Manual*, Institute of Mathematical Modelling, Technical University of Denmark, Lyngby, Denmark.

Ripley, B. (1993), Statistical aspects of neural networks. In *Chaos and Networks - Statistical and Probabilistic Aspects, pp. 40-123*, Chapman and Hall, London.

Ripley, B. (1994), 'Software for neural networks', Statlib Index (`lib.stat.cmu.edu`), S library, `nnet` package. Updated 24 January 1993, 19 September 1993, 13 February 1994.

Schwarz, G. (1978), 'Estimating the dimension of a model', *The Annals of Statistics* **6**(2), 461–464.

Paper I

Approximating building components using stochastic differential equations

Originally published as Chapter 14 in

> J. J. Bloem, editor, *System Identification Competition.* Joint Research Centre, European Commission, 1996. EUR 16359 EN.

I

This book reports the background and results of a competition set up to compare alternative techniques and to clarify particular problems of system identification applied to the thermal performance of buildings. The competition consisted of five cases of simulated data. In the paper included here results for cases 3 and 4 are presented.

For case 3 the data are generated from a second order linear thermal network with three conductances and two capacitances. The data are corrupted by white noise and 20 data sets covering 30 days of hourly observations are provided. The estimates obtained are compared with the true values.

For case 4 the data are generated from partial differential equations describing a two-layer wall with on the inside an insulation layer and on the outside a brick layer. The data are corrupted by white noise. Two data sets are provided, each consisting of 70 days of hourly observations. In one of the data sets the heat flow rate is not provided. Instead this must be predicted using the estimates obtained for the other data set.

In the introduction of the paper there is a reference to Chapter 13 in the book mentioned above.

Approximating building components using stochastic differential equations

Lars Henrik Hansen[1], Judith Lone Jacobsen[1], Henrik Aalborg Nielsen[1] and Torben Skov Nielsen[1]

Abstract

This chapter reports the results for case 3 and 4. In both cases the systems have been modelled by stochastic differential equations with a observation equation. This model may be viewed as lumped parameter systems influenced by noise. In both cases the maximum likelihood method has been used for parameter estimation.

In case 3 the true system is of the lumped parameter type. Hence the system may be modelled without approximations. This case has been used to investigate the estimation method used.

In case 4 the true system is approximated by modelling it as a lumped parameter system. Here the focus is on producing reliable predictions. This was sought obtained by requiring the estimates of the physical parameters to be reasonable with respect to the building materials and the dimension of the wall. In general a lumped parameter system does not have a unique formulation as stochastic differential equations with a observation equation. It is shown that the different formulations may lead to very different estimates. It is also shown that better parameter estimates may be obtained by restricting the number of free parameters in the models by incorporating physical knowledge into the model.

1 Introduction

The cases 3 and 4 are considered. In both cases the results are obtained by modelling the systems as lumped parameter systems. Thus, in case 3 the deterministic part of the system may be modelled without any approximations, therefore the focus is on evaluating the estimation method used. The method requires the user to make some choices before estimation. These choices may influence the appropriateness of the estimates

[1]Department of Mathematical Modelling, Technical University of Denmark, DK-2800 Lyngby, Denmark

and of the standard errors of the estimates obtained. Two approaches are investigated. Furthermore the noise model are considered. In case 4 the system is of the distributed parameter type. In this case the model class used may be regarded as an approximation and the focus is on how to obtain a good approximative model. A "good model" is taken to be a model which incorporates some physical knowledge about the system and for which the validation results (see Sec. 3.3) are acceptable. It is a basic assumption of the work presented in this chapter that if the model is good it will produce reliable predictions, i.e. predictions for which the prediction errors are independent of each other and of external influences.

A robust maximum likelihood method is used for parameter estimation. A brief overview of the method may be found in chapter 13. The method has been implemented as the program CTLSM (Continuous Time Linear Stochastic Modelling), which uses the maximum likelihood method when 1-step prediction errors are requested as the estimation criteria. In this case version 2.6 has been used. The program and manual may be obtained by anonymous ftp from `ftp.imm.dtu.dk`, in the directory `pub/unix/ctlsm`. See the file `README.1ST` for further instructions.

For a more thorough account of the method used as well as for previous work specific to buildings and building components please see Melgaard (1994) and Madsen & Holst (1995).

2 Case 3

In this case the true system is known and corresponds to a lumped parameter system, also called a thermal network. The true system may be represented by stochastic differential equations, which can be handled by CTLSM. Model approximation is therefore not an issue. The case is used to investigate the estimation procedure built into CTLSM.

The connections between the thermal resistances (H_1, H_2, H_3) [$^\circ$Cm^2/W] and the thermal capacitances (G_1, G_2) [$Wh/^\circ$Cm^2] are shown in Figure 2 of the competition rules. In the following T_e [$^\circ$C] and T_i [$^\circ$C] represent the external and internal surface temperature, respectively. The internal heat flow into the wall is denoted q_i [W/m^2]. The temperature between

H_1 and H_2 is called T_1 [°C] and the temperature between H_2 and H_3 is called T_2 [°C].

2.1 Model formulation

When using the software tool CTLSM, the model must be formulated in terms of a set of linear first order ordinary differential equations called the system equation and a set of algebraic equations called the observation equation. The system and the observation equation may each be of multivariate nature. Hence CTLSM is useful for estimation of parameters in e.g. thermal network models.

In this case, as well as in the case where a true wall component is considered, there is more than one possible formulation of the model. The following system equation was used

$$\begin{aligned}\begin{bmatrix} dT_1 \\ dT_2 \end{bmatrix} &= \begin{bmatrix} -(\frac{1}{H_1G_1}+\frac{1}{H_2G_1}) & \frac{1}{H_2G_1} \\ \frac{1}{H_2G_2} & -(\frac{1}{H_2G_2}+\frac{1}{H_3G_2}) \end{bmatrix} \begin{bmatrix} T_1(t) \\ T_2(t) \end{bmatrix} dt \\ &+ \begin{bmatrix} \frac{1}{H_1G_1} & 0 \\ 0 & \frac{1}{H_3G_2} \end{bmatrix} \begin{bmatrix} T_e(t) \\ T_i(t) \end{bmatrix} dt + dW(t),\end{aligned} \quad (1)$$

where $dW(t)$ is a process with independent increments, which in this case represents the measurement noise on the internal and external surface temperatures. The corresponding observation equation is

$$q_i(t) = \begin{bmatrix} 0 & -\frac{1}{H_3} \end{bmatrix} \begin{bmatrix} T_1(t) \\ T_2(t) \end{bmatrix} + \begin{bmatrix} 0 & \frac{1}{H_3} \end{bmatrix} \begin{bmatrix} T_e(t) \\ T_i(t) \end{bmatrix} + e(t), \quad (2)$$

where $e(t)$ is assumed to be white noise. According to the competition rules T_e is associated with a high level of noise, q_i with a low level of noise, and T_i with no noise.

With $\Delta W = [\ \Delta W_1 \ \ \Delta W_2\ ]^T$ denoting the increment in $W(t)$ from time t until time $t+\Delta t_s$, where Δt_s is the sampling interval, it may be deduced that

$$[\ \Delta W_1 \ \ \Delta W_2 \ \ e\ ]^T = [\ \frac{\epsilon_e}{H_1G_1} \ \ 0 \ \ \epsilon_q\ ]^T. \quad (3)$$

where noise on $T_e(t)$ and $q_i(t)$ is denoted $\epsilon_e(t)$ and $\epsilon_q(t)$, respectively. In the following it will be assumed that $\epsilon_e(t)$ iid $N(0, \sigma_e^2)$ and

$\epsilon_q(t)$ iid $N(0, \sigma_q^2)$. Hence the increase in $W(t)$ over the sampling period will be normally distributed with zero mean and constant variance, i.e. $W(t)$ may be regarded as a Wiener process. Furthermore it is seen that $e(t)$ is iid $N(0, \sigma_q^2)$.

From (3) it is seen that since $\epsilon_e(t)$ and $\epsilon_q(t)$ are independent so are also $dW(t)$ and $e(t)$. This comply with the assumptions of the estimation method used. For this reason the above formulation is attractive.

Remark:
If the model is reformulated so that T_e and q_i are used as inputs (corresponding to T_e and T_i above) the covariance between ΔW_2 and e will be $\frac{H_3}{G_2}\sigma_q^2$. The remaining covariances will be zero.

2.2 Estimation of parameters

Using the formulation (1) - (2) Maximum Likelihood estimates of the parameters H_1, H_2, H_3, G_1, and G_2 were found. Furthermore the initial values of T_1 and T_2 were estimated together with the system and observation noise variances.

In order to perform the estimation it is necessary to supply some prior knowledge about the value of the internal and external surface temperature between samples. In this case a linear interpolation was used. Furthermore lower and upper bounds on the estimates have to be supplied. It is very important to select these bounds so that the calculation of the exponential of the matrix

$$\begin{bmatrix} -(\frac{1}{H_1G_1} + \frac{1}{H_2G_1}) & \frac{1}{H_2G_1} \\ \frac{1}{H_2G_2} & -(\frac{1}{H_2G_2} + \frac{1}{H_3G_2}) \end{bmatrix} \tag{4}$$

is feasible i.e. the estimation procedure should be prevented from investigating sets of parameters where the eigenvalues of the matrix (4) has very large and/or very small real parts. The expected true values should of course be well within the bounds.

In order to find appropriate starting values for the twenty data sets the first few data sets were investigated more closely. Two different approaches were used; (i) the system and observation variances were estimated directly, and (ii) the Kalman gain was estimated instead of the noise variances. This was done in order to investigate if the approximation of the Kalman gain results in better convergence of the estimates, than the direct method. The initial values and the bounds are shown in Table 1. For method (i) the variance of ΔW_2 was fixed at 10^{-9}. The initial value of $V[\Delta W_1]$ was 10^{-6}, and the initial value of $V[e]$ was 10^{-2}. For both estimates the bounds were 10^{-9} and 10^{-1}. For method (ii) the initial value of both elements of the Kalman gain vector were 0.1, with the bounds 0 and 100.

		H_1	H_2	H_3	G_1	G_2	$T_1(t=0)$	$T_2(t=0)$
	LB	10^{-1}	10^{-1}	10^{-3}	10^{-1}	10^{-1}	5.0	5.0
(i)	init.	1.0	1.0	1.0	1.0	1.0	22.0	12.0
	UB	10.0	10^{2}	10.0	5×10^{2}	2×10^{2}	40.0	20.0
	LB	10^{-4}	10^{-4}	10^{-4}	10^{-4}	10^{-4}	-10.0	-10.0
(ii)	init.	2.0	20.0	2.0	10.0	10.0	20.0	10.0
	UB	10^{2}	10^{3}	10^{2}	10^{4}	10^{4}	10^{2}	10^{2}

Table 1: Initial values (init.), lower bound (LB), and upper bound (UB), for methods (i) and (ii).

Method (i)

Using method (i) the first four data sets were investigated. The result of this investigation indicated that

$$\begin{aligned} &H_1 \approx 1, \quad H_2 \approx 10, \quad H_3 \approx 0.1 \\ &G_1 \approx 100, \quad G_2 \approx 50, \\ &T_1(t=0) \approx 12, \quad T_2(t=0) \approx 22 \end{aligned}$$

The initial values of temperatures and variances, as well as the bounds were not changed. The suggested values were used as initial estimates for a second run of all 20 data sets, and these were the results reported.

Method (ii)

When using method (ii) only the first data set was investigated. The resulting estimates were not used directly as initial values for all 20 data sets. Instead the values were altered approximately 20%, in a way that kept the sum of the resistances and the sum of the capacitances approximately constant. This was done in order to ensure that the algorithm approximated the covariance matrix of the estimates well, see the description of the BFGS-update of the Hessian in Dennis & Schnabel (1983). The following *initial* estimates were used

$$\begin{aligned}
&H_1 = 1.8, \quad H_2 = 8.0, \quad H_3 = 0.12, \\
&G_1 = 77.0, \quad G_2 = 40.0, \\
&T_1(t=0) = 15.0, \quad T_2(t=0) = 23.0
\end{aligned}$$

In the second run the variances were estimated, except $V[\Delta W_2]$ which was set to zero as indicated by (3). The initial values for the remaining variances were set to 1.0 with a lower bound of 10^{-5} for $V[\Delta W_1]$ and 10^{-6} for $V[e]$. For both variances an upper bound of 100.0 was used. For method (ii) the results of these 20 runs are the ones reported.

2.3 Results

2.3.1 Individual estimates

The results of the estimations are shown on Figure 1. It is seen that the two methods lead to approximately the same results. Furthermore it is seen that the true values, in most cases, are within the approximate 95% confidence interval calculated as ± two times the standard error. This indicates that the calculation of the estimates as well as the standard errors are appropriate. For both methods the estimates of the initial value of T_1 are all approximately 25°C and 13-14°C for T_2. The estimates of the variances corresponds to $\sigma_q \approx 0.01 W/m^2$ and $\sigma_e \approx 1.6$°C. This comply well with the information in the competition rules regarding the noise levels, although σ_q seems to be low.

2.3.2 Mean of estimates

Since each estimate is asymptotically normally distributed with constant mean and variance (all data sets are of equal length) it is possible to estimate the mean and the standard error of the estimates.

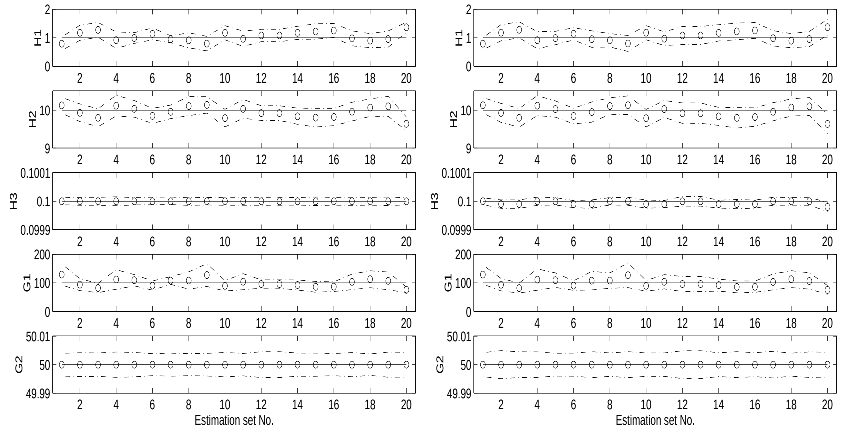


Figure 1: Estimation results, left method (i), right method (ii), "- · -"marks ± two times the standard error, "–" indicates the true values.

In Table 2 the means of the 20 estimates are displayed together with a 95% confidence interval of the mean (based on the estimates) and the true value of the estimates. From the table it is seen that the two methods give very similar results. In all cases the true value is included in the 95% confidence interval and hence the mean does not differ from the true values at the 5% level of significance. However the confidence intervals are quite wide for H_1 (-2% - 13% of the true value) and G_1 (-6% - 7%), indicating a possibility of these estimates to be biased. More than 20 sets of simulations will be needed to narrow these confidence intervals.

	True	Method (i)			Method (ii)		
		Mean	Lower	Upper	Mean	Lower	Upper
H_1	1.0	1.053	0.977	1.130	1.055	0.979	1.132
H_2	10.0	9.948	9.882	10.014	9.946	9.880	10.012
H_3	0.1	0.100	0.100	0.100	0.100	0.100	0.100
G_1	100.0	100.365	93.737	106.994	100.197	93.588	106.806
G_2	50.0	50.000	49.999	50.001	50.000	49.999	50.001

Table 2: Means of estimates and 95% confidence interval of the mean.

2.3.3 Variance of estimates

It is noted that for method (i) the No. of iterations ranges from 45 to 61, for method (ii) the corresponding range is 49 to 62. Hence, for both methods the No. of iterations should be large enough to ensure an appropiate approximation of the 2nd order derivatives.

In Figure 2 the standard errors calculated by CTLSM (SEc) are shown together with the mean of these. Furthermore the standard error (SEe) based on the set of 20 parameter estimates, together with a 95% confidence interval of the true standard error of the parameter estimates, are shown in Figure 2.

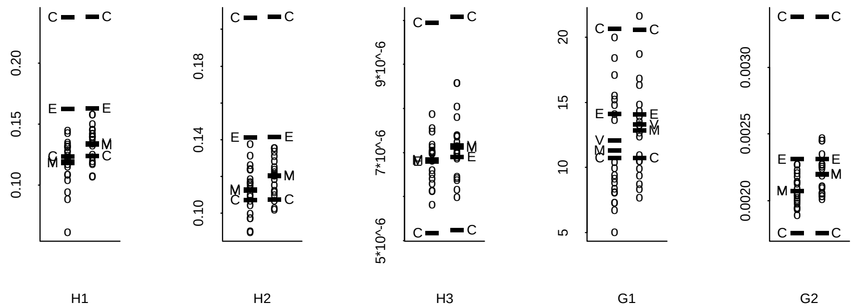


Figure 2: Calculated standard error (SEc) of estimate for each data set "o", mean of SEc "M", mean of SEc based on variances "V", estimated standard error (SEe) based on parameter estimates "E", and 95% confidence interval of the true standard error "C". Method (i); left.
Method (ii); right. In case a "V" is not visible it indicates that in the particular case $\sum \mathrm{SEc}_i/20 \approx \sqrt{\sum \mathrm{SEc}_i^2/20}$.

From the figure it is seen that the methods give similar results, although method (i) tends to result in slightly lower values of SEc than method (ii). The SEc's are in most cases in the low range of the 95% confidence interval.

The confidence intervals are wide, indicating that the SEe's are quite uncertain. It is therefore difficult to judge the appropriateness of the SEc's. However the spread of the SEc's is in some cases in the range of the 95% confidence interval (e.g. for parameter G_1), i.e. for one particular data set the SE-values calculated by CTLSM are uncertain.

Another issue is whether CTLSM, on average, calculates the SE's correctly, i.e. if the expected SEc is equal to the true standard error of the parameter estimates. If the expected SEc is approximated with the square root of the average variance (i.e. $\sqrt{\frac{1}{20}\sum \mathrm{SEc}_i^2}$), the test result on the 5% level of significance can be seen on Figure 2. Only for parameter H_1, method (i) the "V" is slightly outside the 95% confidence interval indicating that in this case the SEc is on average different from the true standard error of the parameter estimate (the p-value of the test equals 0.028). However, as mentioned above the confidence intervals is wide, indicating that the test just mentioned has low power and that a large average difference may be present. This may be investigated performing more than 20 simulations, the No. of which can be determined by use of power calculations.

2.4 Temperatures as input, heat flow as output

The system and observation equations (1) - (2) may be reformulated so that q_i and T_e are used in the new system equation and T_i is the variable on the left hand side of the new observation equation. As mentioned in Section 2.1 this violates the assumption built into CTLSM regarding independent system and observation noise. It was observed that (i) the estimate of H_3 is systematically larger than the true value, and (ii) in many cases the true value is not included in the approximate 95% confidence interval. Hence, it is concluded that the assumption regarding the independence of the system and observation noise is significant with respect to the estimation results.

3 Case 4

The subject in case 4 is prediction of heat flow density through a two-layer wall. Two types of approximative models – a general and a homogeneous – are investigated in this case.

3.1 Model formulation

The general approximative model
The first approximation type is a R-C network (same structure as the network in case 3) where a thermal capacitance is placed between two thermal resistances etc.

The homogeneous approximative model
The second approximation type is also a R-C network, but more restricted. The basic idea is to place the thermal capacitance between two thermal resistances of the same size, as explained below. Figure 3 shows the structure of the homogeneous model using a 3rd order approximation. As described in Section 2.1 the model must be formulated in terms

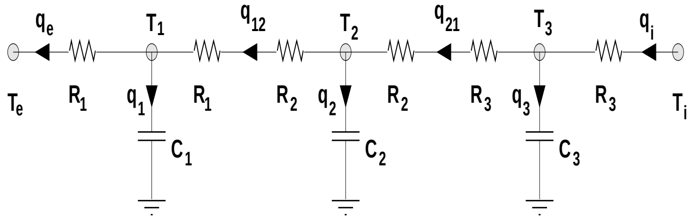


Figure 3: 3rd order approximation model

of a set of first order ordinary differential equations. If q_i e.g. is chosen to be the model output, the CTLSM model for the 3rd order approximation shown in Figure 3 becomes

$$\begin{bmatrix} dT_1 \\ dT_2 \\ dT_3 \end{bmatrix} = \begin{bmatrix} -\frac{1}{C_1}(\frac{1}{R_1}+\frac{1}{R_1+R_2}) & \frac{1}{C_1}\frac{1}{R_1+R_2} & 0 \\ \frac{1}{C_2}\frac{1}{R_1+R_2} & -\frac{1}{C_2}(\frac{1}{R_1+R_2}+\frac{1}{R_2+R_3}) & \frac{1}{C_2}\frac{1}{R_2+R_3} \\ 0 & \frac{1}{C_3}\frac{1}{R_2+R_3} & -\frac{1}{C_3}(\frac{1}{R_2+R_3}+\frac{1}{R_3}) \end{bmatrix} \begin{bmatrix} T_1 \\ T_2 \\ T_3 \end{bmatrix} dt$$
$$+ \begin{bmatrix} \frac{1}{R_1 C_1} & 0 \\ 0 & 0 \\ 0 & \frac{1}{C_3 R_3} \end{bmatrix} \begin{bmatrix} T_e \\ T_i \end{bmatrix} dt + dW(t) \quad (5)$$

where $dW(t)$ is a process with independent increments, which in this case represents the measurement noise on the internal and external surface temperatures. The corresponding observation equation is

$$q_i(t) = \begin{bmatrix} 0 & 0 & -\frac{1}{R_3} \end{bmatrix} \begin{bmatrix} T_1 \\ T_2 \\ T_3 \end{bmatrix} + \begin{bmatrix} 0 & 0 & \frac{1}{R_3} \end{bmatrix} \begin{bmatrix} T_e \\ T_i \end{bmatrix} + e(t) \quad (6)$$

where $e(t)$ is assumed to be white noise. As in case 3 it is also possible to use T_i as output. This will change the last equation in the set of system equations and of course the observation equation.

For a homogeneous wall of a specific material, capacitance and resistance are directly proportional with thickness. Increasing the thickness of the wall by a factor 2 will result in a capacitance and resistance twice as big. Therefore, setting the ratio between the thermal capacitance C_i and thermal resistance R_i for all layers to a constant, is an interpretation of the physical property *homogeneous*. Applying this concept to the R-C network shown in Figure 3 will result in the homogeneous approximation model. Thus, the homogeneous approximation model is a special case of the general approximation model. Comparing the two models it can be stated that

- the homogeneous approximative model has a more physical interpretation than the general model.

- the general approximative model has $2n + 1$ parameters ($n + 1$ resistances and n capacitances), whereas the homogeneous model has $n + 1$ parameters (n resistances and one RC-ratio), where n is the model order.

3.2 Notation

To help the reader, the following notation is introduced. Let "G" denote the general approximation model and "H" the homogeneous approximative model. The model type will be followed by the term "q" or "T" to distinguish if the internal heat flow or the internal temperature was used as output. Then a "/" follows. The notation term ends with the order of the used model. E.g. the term "Hq/2" can be decoded to: the homogeneous approximative model in the case of q_i as output and with a model order of 2.

3.3 Model validation

The model validation has been based on two criterias; (i) a prediction error analysis from the estimation package (CTLSM Melgaard & Madsen (1993)) and (ii) a physical interpretation of the estimated parameters.

Prediction error validation (i)
When performing an estimation based on the prediction error method, the 1. step prediction errors for an appropriate model must be white noise and uncorrelated with the inputs. CTLSM provides the cross correlation functions (CCF) of all the inputs with the PE as well as auto correlation function (ACF) and cumulative periodogram (CP) for PE to verify that these requirements are fullfilled.

Physical validation (ii)
The physical validation of the estimated parameters is based on the assumption that estimates indicating wall components magnitudes larger / smaller than 0.1 m is caused by a local minimum in the optimization criterion or an inappropriate model and therefore should be rejected.

3.4 The system equations

At first a simple second order model was formulated for the whole wall. The model is of the general type (GT/2 and Gq/2) as described in Section 3.1.

The other model is formulated using the assumption that each wall component is fully homogeneous, i.e. each of the components are of the type HT/2 and Hq/2, thus becoming a 4th order model. Because of the homogeneity assumption, the ratio between the resistance and the capacities in each layer are set to remain constant. The relationships are:

$$K_w = \frac{R_{1w}}{C_{1w}} = \frac{R_{2w}}{C_{2w}} \quad \text{and} \quad K_i = \frac{R_{1i}}{C_{1i}} = \frac{R_{2i}}{C_{2i}} \tag{7}$$

The system and observation equations, for the system with T_i as output, becomes:

$$
\begin{aligned}
\frac{dT_1}{dt} &= -\frac{K_w}{R_{1w}}\left(\frac{1}{R_{2w}+R_{1w}}+\frac{1}{R_{1w}}\right)T_1 \;+\; \frac{K_w}{R_{1w}(R_{2w}+R_{1w})}T_2 \;+\; \frac{K_w}{R_{1w}^2}T_e \\
\frac{dT_2}{dt} &= \frac{K_w}{R_{2w}(R_{2w}+R_{1w})}T_1 - \frac{K_w}{R_{2w}}\left(\frac{1}{R_{1i}+R_{2w}}+\frac{1}{R_{2w}+R_{1w}}\right)T_2 \;+\; \frac{K_w}{R_{2w}(R_{1i}+R_{2w})}T_3 \\
\frac{dT_3}{dt} &= \frac{K_i}{R_{1i}(R_{1i}+R_{2w})}T_2 - \frac{K_i}{R_{1i}}\left(\frac{1}{R_{2i}+R_{1i}}+\frac{1}{R_{1i}+R_{2w}}\right)T_3 \;+\; \frac{K_i}{R_{1i}(R_{2i}+R_{1w})}T_4 \\
\frac{dT_4}{dt} &= \frac{K_i}{R_{2i}(R_{2i}+R_{1i})}T_3 - \frac{K_i}{R_{2i}(R_{2i}+R_{1i})}T_4 \;+\; \frac{K_i}{R_{2i}}T_i \qquad (8)
\end{aligned}
$$

$$
T_i = T_4 \;+\; R_{2i}\,T_i \qquad (9)
$$

3.5 Initial parameter estimates

In order to ensure global optimality of the optimization criteria the initial estimates were chosen by using the physical knowledge of the system. Case 4 is described as a two-layer wall with a brick layer and an insulation layer on the internal side of the wall. Key material constants, such as the ones that goes into the thermal capacity, C and thermal resistance, R for brick and insulation can be found in Incropera & Witt (1985). Using danish standard wall components, it has been assumed that the wall consists of 0.11 m brick and 0.1 m isolation.

Resistance may be calculated by dividing the thickness, d of the building component by the thermal conductivity, λ i.e.: $R = d/\lambda$. The thermal capacity is calculated as the density, ρ multiplied by the specific heat capacity, c and the thickness of the wall i.e.: $C = c_p\ \rho\ d$. In Table 3, the densities and specific heat capacities for some expected materials are listed together with the resulting values of R and C.

Constant	d	λ	ρ	c_p	R	C
Unit	m	$W/m\ K$	kg/m^3	$W\,h/kg\,K$	$^oC\,m^2/W$	$W\ h/^oC\ m^2$
Brick	0.11	0.720	1920	0.250	0.153	52.800
Glass fiber	0.10	0.036	105	0.210	2.778	0.058
Wall + isol.					2.931	52.858

Table 3: Physical parameters for thermal characteristics of two types of building components

3.6 Estimation results

The estimates of the resulting thermal capacities and resistances for the two model types are listed in Table 4. Standard deviations are only shown for the simple second order model. To calculate the standard deviations for the more complex 4th order model, the linear approximation formula should be used along with the standard formulas for correlated stochastic variables.

Model	C_i	C_w	R_i	R_w	C	R
Gq/2	—	—	—	—	8775	3.202
	—	—	—	—	(6754)	(0.038)
GT/2	—	—	—	—	23.159	2.707
	—	—	—	—	(2.623)	(0.060)
$2 \cdot$ Hq/2	1.368	345.987	3.101	0.036	347.346	3.137
$2 \cdot$ HT/2	1.016	28.158	2.590	0.543	29.174	3.132

Table 4: The resulting thermal capacities and resistances for the two model types. Non-applicable cells are marked by —. Standard deviations for the 2nd order model is shown in brackets.

3.7 Validation of results

The results from the prediction error analysis of the different models tried for case 4 are shown in Table 5. These checks suggests that a 2nd order model may not adequately describe this system. However, the model validation tools indicate that the parameters found in the 2×HT/2 case, are reasonable. Also, it is clear that there are several unresolved problems, when estimating a system that defines q_i as output. While the 4th order model was excellent with T_i as output, this was not the case with q_i as output. The analysis of the residuals were not that far apart, but the parameter estimates were quite unrealistic in the last case. A thermal capacitance of 346 $W\ h/^oC\ m^2$ for the brick wall means that this wall should be 0.72 m thick! The estimate of the insulation is not as bad, it corresponds to 0.06 m of glassfiber.

Both types of models gave unrealistic estimates, when q_i was used as output. For the 4th order model it is seen that the estimates reflects

Model	CP	ACF	CCF
Gq/2	Nearly straight line. 6/60 values outside the CI	Not quite adequate. 8/20 values outside the CI. Highest value = 0.106	Inadequate.
GT/2	Nearly straight line. 2/60 values outside the CI	Not quite adequate. 3/20 values outside the CI.	Inadequate.
2 × Hq/2	Nice straight line. No values outside the CI	O.K. 1/20 value outside the CI.	Inadequate, largest value = 0.132.
2 × HT/2	Nice straight line. No values outside the CI	Adequate	Adequate

Table 5: Description of the residual tests. CI: 95 % confidence interval ± 0.048. 6/60 means 6 out of 60, etc.

the physical system, i.e. that the capacity of the insulation layer was much smaller than for the brick wall, while the opposite was true for the resistance. When T_i was used as output, the estimates were smaller than expected from the comparison with physical values. However, the real building components were not known; perhaps the layer of insulation was only 0.05 m, or a different brick-type was used. Both of these possibilities could explain the estimated values.

3.8 Prediction

Because of the problems with estimating the parameters, based on the system described with q_i as output, it was decided to perform the prediction by using the parameters estimated with T_i as output in a model with q_i defined as output. The coefficient of variance (CV) for the general case, which was found to be inadequate, was calculated by the competition committee to 92.6 %. Figure 4 shows the predictions for both the general and the specialized system. It can be observed that the homogeneous model follows the internal and external temperatures in a way that is plausible. This is not the case for the general model. In the predictions from the homogeneous model the fast variations in q_i follow the ones in T_i quite accurately, while the slow variations in q_i can be explained by the slow variations in T_e.

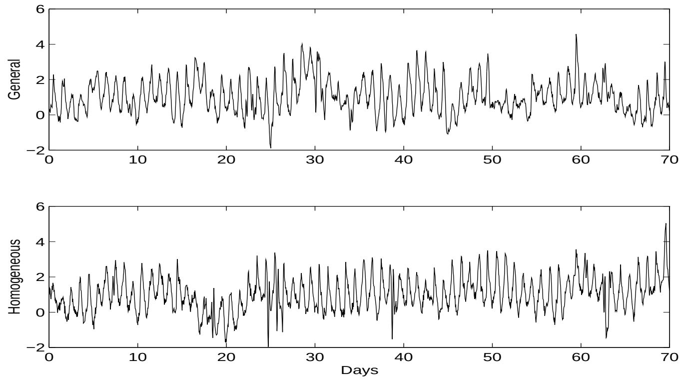


Figure 4: Prediction of heat flow for data42.dat in both the general and the homogeneous case.

4 Discussion and conclusion

In this chapter two different aspects of modelling building components were addressed. In case 4 real, but simple, building components were considered. The true equations describing these components are known in principle. However, software for stochastic partial differential equations were not available and these were therefore approximated by stochastic ordinary differential equations. This approximative model may be viewed as a lumped parameter model like the one in case 3. This case therefore gives a way of investigating the estimation procedure when the true model is known, while case 4 focus on the problem where a distributed system is approximated by a lumped parameter model.

The estimation procedure (case 3)
In this case CTLSM was able to estimate the true parameter values and to provide reasonable estimates of the uncertainty in general.

The investigation of the estimation procedure indicates, that when the assumption of independence of the system and observation noise is violated, some of the estimates will be biased. When the assumption holds, the estimates do not seem to be biased, although for two parameters (G_1 and H_1) the confidence interval of the mean indicates the possibility of a bias; or maybe the input series are not the most optimal for estimating these parameters. The standard errors of the estimates calculated by the estimation procedure are for three of the five parameters somewhat low, compared to the variation between the 20 sets of parameter estimates. On average the standard errors calculated by the estimation procedure seems to be appropriate (insignificant difference), but this statement is uncertain due to wide 95% confidence intervals of the true variation between sets of parameter estimates. When analyzing a single data set slightly uncertain values of the standard errors are obtained. Furthermore it is concluded that starting values seem to be of minor importance with respect to the convergence of the estimates.

Model approximation (case 4)
From this investigation it is concluded that when approximating wall components with lumped parameter systems it is important to (i) consider different model formulations of the same lumped parameter system, i.e. the choice of output variable (ii) consider whether the lumped parameter system should be restricted to represent one or more homogeneous wall components, and (iii) investigate different approximation orders.

Re. (i): The necessity of this investigation may be due to the violation of the independence assumption build into the estimation procedure. However, the fact that the model is an approximation may also make this kind of investigation necessary.

Re. (ii): The results obtained tend to imply that all physical knowledge available, i.e. information regarding homogeneity, should be used when modelling the system. This will limit the No. of parameters and to some extent remove correlations between estimates. Physical knowledge regarding the magnitude of the parameters may be used to obtain appropriate starting values, thus facilitating a more successful estimation.

Re. (iii): The order of the system of ordinary differential equations is important. On one hand it should be low in order to keep the number of parameters low while on the other hand it must be high enough to model all frequencies of the system. Therefore the order necessarily is dependent on the frequencies of the input signal.

Evaluation of models
As the most objective criteria the models should be evaluated by inspecting the prediction errors (cumulative periodogram, autocorrelation, cross correlation with input). Furthermore it is important to evaluate the final

parameter estimates of thermal resistance and capacitance, although this criteria is somewhat subjective.

References

Dennis, J. & Schnabel, R. (1983), *Numerical Methods for Unconstained Optimization*, Prentice-Hall.

Incropera, F. & Witt, D. D. (1985), *Fundamentals of Heat and Mass Transfer*, 2nd edn, John Wiley & Sons, New York.

Madsen, H. & Holst, J. (1995), 'Estimation of continuous-time models for the heat dynamics of a building', *Energy and Buildings* **22**, 67–79.

Melgaard, H. (1994), Identification of Physical Models, PhD thesis, Institute of Mathematical Modelling, Technical University of Denmark, Lyngby, Denmark.

Melgaard, H. & Madsen, H. (1993), *CTLSM Continuous Time Linear Stochastic Modelling*, Commission of the European Communities, Institute for Systems Engineering and Informatics, Joint Research Centre. In: Workshop on Application of System Identification in Energy Savings in Buildings.

Paper J

Goodness of fit of stochastic differential equations

Originally published as

> Jakob Bak, Henrik Aalborg Nielsen, and Henrik Madsen. Goodness of fit of stochastic differential equations. In Peter Linde & Anders Holm, editors, *21st Symposium of applied statistics*, Copenhagen Business School, Copenhagen, January 1999.

J

Goodness of fit of stochastic differential equations

Jakob Bak[1], Henrik Aalborg Nielsen[1], and Henrik Madsen[1]

Abstract

We propose a method to test for lack-of-fit of an estimated stochastic differential equation. The method is based on Monte Carlo simulation of trajectories between neighbour observations and, thus, it does not rely on the availability of explicit expressions of the conditional densities. Consequently, both non-linear models and models with state-dependent drift and diffusion can be handled. The method is illustrated by an example.

1 Introduction

After parametric or non-parametric estimation of the drift and diffusion of a stochastic differential equation it is often of interest to access the validity of the model obtained by testing for lack-of-fit. In this paper we consider models of the class

$$dX(t) = \mu(X(t))dt + \sigma(X(t))dW(t), \tag{1}$$

in which the drift $\mu()$ and diffusion $\sigma()$ depend on the univariate continuous-time process $\{X(t),\ t \geq 0\}$ and where $\{W(t),\ t > 0\}$ is a standard Wiener process. The drift and diffusion may be parametrized so that a vector θ of real numbers completely characterizes the functions, but the method presented is not dependent on this assumption.

Having obtained estimates of the drift $\hat{\mu}()$ and diffusion $\hat{\sigma}()$ we use

$$dX(t) = \hat{\mu}(X(t))dt + \hat{\sigma}(X(t))dW(t), \tag{2}$$

as the null hypothesis (H_0), which is tested against the alternative (H_a) that $\{X(t),\ t \geq 0\}$ can not be described by (2). The conditional densities are known only for some parametrizations of (1) as for instance linear models, see also (Maybeck 1982). For parametrized drift and diffusion

[1]Department of Mathematical Modelling, Technical University of Denmark, DK-2800 Lyngby, Denmark

Aït-Sahalia (1998) uses transformations followed by Hermite polynomial expansions to obtain approximations of the conditional densities. However, to our knowledge, approximations of conditional densities are currently not available when the drift and diffusion are not parametrized. Hence, for general use the test can not be based on expressions of conditional densities. However, under H_0 the process can be simulated, e.g. between sample points. And these simulations can be used to test H_0 against H_a, which is the approach considered in this paper. The approach were originally suggested by Bak (1998).

2 Monte Carlo simulation

The tests described in this paper are based on Monte Carlo simulations of trajectories between observation points as illustrated in Figure 1.

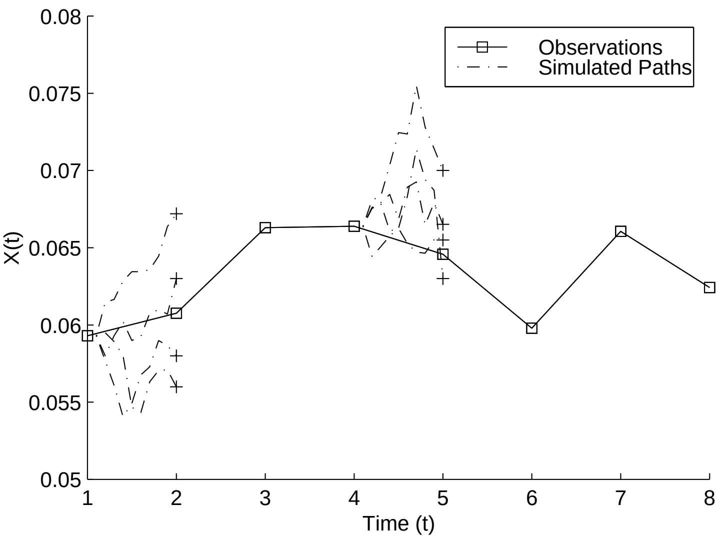


Figure 1: Simulation of trajectories for $t \in [1, 2]$ and $t \in [4, 5]$ when the trajectories start at the observed value of $X(1)$ and $X(4)$, respectively.

Provided the time step used for simulation is small enough the Euler scheme (Madsen, Nielsen & Baadsgaard 1998) can be used to obtain an adequate precision. However, for small time steps the possibility of numerical difficulties can not be ruled out. In these situations simulations can be performed using, e.g., the Milstein scheme (Madsen et al. 1998). The sampling interval is divided into a number of smaller intervals of length Δ. Hereafter the Euler scheme is just a simple discretization

$$Y_{n+1} = Y_n + \hat{\mu}(Y_n)\Delta + \hat{\sigma}(Y_n)\Delta W, \tag{3}$$

where $\Delta W \sim N(0, \Delta)$ is the increment of the Wiener process over the interval Δ. The Milstein scheme includes an additional term from the Taylor approximation:

$$Y_{n+1} = Y_n + \hat{\mu}(Y_n)\Delta + \hat{\sigma}(Y_n)\Delta W + \frac{1}{2}\hat{\sigma}(Y_n)\widehat{\frac{d\sigma}{dX}}(Y_n)\left[(\Delta W)^2 - \Delta\right]. \tag{4}$$

For parametric models the derivative of the diffusion with respect to the state is readily available when the parameters are estimated, when using non-parametric methods it will de necessary to estimate this derivative.

If the drift and diffusion are estimated by non-parametric methods (Bak 1998), estimates will only exist for an interval $[a, b]$ as spanned by the observations of $X(1), \ldots, X(N)$. However, almost certainly, some of the simulated trajectories will take values outside the afore mentioned interval, as for instance for $t \in [4, 5]$ in Figure 1. As a consequence, it is necessary that the estimates are accomplished by a decision about how extrapolation should be performed. In practice it is further recommend that the maximum deviation from $[a, b]$ are investigated to access the validity of the test. This is also relevant when parametrizations of the drift and diffusion are used; for instance polynomial parametrizations may be completely misleading outside $[a, b]$. These considerations are actually just a consequence of the fact that in order to test the null hypothesis the drift and diffusion need to be defined for all possible values of $\{X(t),\ t > 0\}$.

3 A positional lack-of-fit test

Assume that the process $\{X(t), t > 0\}$ is observed at time points $t = 1, \ldots, N$, i.e. for simplicity time is recorded in a unit such that the time step between observations one. Let $x_1, \ldots, x_N$ denote the observations. For each $t = 2, \ldots, N$ simulations are performed to obtain M trajectories from time $t-1$ until t starting at x_{t-1}, cf. Section 2. Hereafter the rank r_t of x_t as compared to the endpoints of the M simulated trajectories is calculated, for example $r_2 = 3$ on Figure 1. With R_t being the stochastic variable corresponding to the observation r_t it holds under H_0, i.e. (2), that

$$P\{R_t = q\} = p_{tq} = \frac{1}{M+1}; \quad q = 1, \dots, M+1; \quad t = 2, \dots, N. \tag{5}$$

In general the dependence of p_{tq} on t and q is of interest. However, for every $t = 2, \dots, N$ only one observation of R_t is available and therefore we must assume that the probability is independent of time, i.e. $p_{tq} = p_q$. Under this assumption

$$\hat{p}_q = \frac{\Omega_{N-1}(q)}{N-1}; \quad q = 1, \dots, M+1, \tag{6}$$

with

$$\Omega_{N-1}(q) = \sum_{t=2}^{N} I(R_t = q); \quad q = 1, \dots, M+1, \tag{7}$$

where $I(R_t = q) = 1$ if $R_t = q$ and 0 otherwise. Under H_0 it is clear that

$$E[\hat{p}_q] = E\left[\frac{\Omega_{N-1}(q)}{N-1}\right] = \frac{1}{M+1}. \tag{8}$$

The Pearson test statistic (Kendall & Stuart 1961) for the hypothesis that $p_q = 1/(M+1)$; $q = 1, \dots, M+1$ is therefore

$$\mathrm{X}^2 = \sum_{q=1}^{M+1} \frac{\left(\Omega_{N-1}(q) - \frac{N-1}{M+1}\right)^2}{\frac{N-1}{M+1}}, \tag{9}$$

which, under H_0, asymptoticly is distributed as $\chi^2(M)$. According to Kendall & Stuart (1961, p. 440) the approximation fails when the frequencies expected under H_0 are small. Many researchers have used the rule that no expected frequencies should be less than 5. Therefore, in this case $(N-1)/(M+1) \geq 5$, yielding an upper bound of $(N-6)/5$ on M. In practice, the number of simulated trajectories will be well below that bound. If for instance $N = 500$ the rule requires that no more than 98 inter-observation trajectories are simulated for each $t = 2, \dots, N$.

Remark: Following Kendall & Stuart (1961) the ranks R_t; $t = 2, \dots, N$ should be independently identical distributed under both H_0 and H_1.

From the assumption $p_{tq} = p_q$ the identical distribution of the ranks follow. All simulations are performed so that the trajectories are independent given the initial value, therefore, if H_0 is true, the ranks will also be independent. However, if H_0 is not true the dependence between observations of the process will possibly result in some dependence of ranks close in time. In effect this means that the number of observations of the rank $N - 1$ is not an adequate measure of the effective number of observations N_{eff}, but if N_{eff} tends to infinity as N tends to infinity the asymptotic result should still hold. However, the upper bound on M is possibly too large. Further research is needed to clarify these aspects.

4 Example

In this section an Ornstein-Uhlenbeck process with a square-root state-dependent diffusion is considered. In finance this is known as a CIR model and is used frequently for spot interest rates. The model contains three parameters α, β, and σ and is written

$$dX(t) = \beta(\alpha - X(t))dt + \sigma\sqrt{X(t)}dW(t). \tag{10}$$

For $\alpha = 0.0593$, $\beta = 0.3294$, and $\sigma = 0.05$ a time series consisting of 500 equidistant observations (the sampling interval is 0.1) from (10) is generated using the Milstein scheme and dividing each sampling interval into 10 subintervals. Hereafter, for $\sigma = 0.04, 0.05, 0.06$ and for each sampling interval $M = 10$ inter-observation trajectories are simulated as described in Section 3 with α and β fixed at the true values. The test statistic (9) and the corresponding p-value are shown in Table 1. It is seen that only for $\sigma = 0.5$ (the true value) the hypothesis that the observations come from (10) can not be rejected.

σ	X^2	p-value
0.04	28.1	0.002
0.05	3.5	0.967
0.06	24.0	0.008

Table 1: Test statistic X^2 and p-value for each value of σ used in the inter-observation simulations.

The 25 first endpoint results for $\sigma = 0.06$ is depicted in Figure 2. Since the variance is too large in the inter-observational trajectories, the real observations are too often located in the middle of the M simulated endpoints.

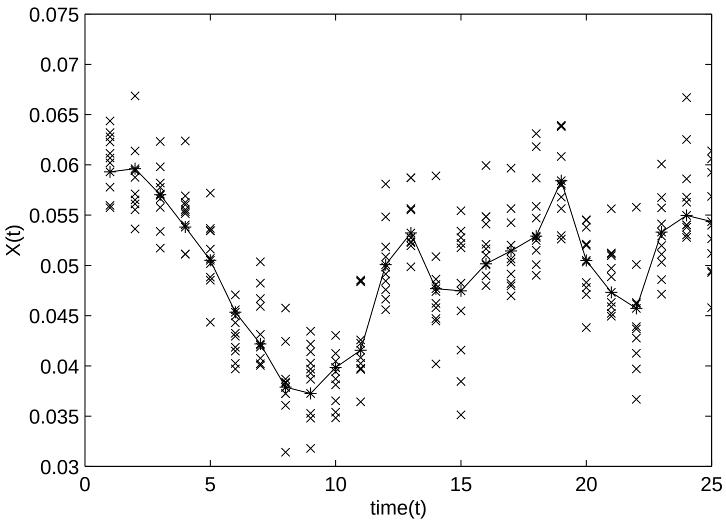


Figure 2: 25 observations and the corresponding endpoints for inter-observational trajectories.

5 Conclusion

A general method for testing for lack-of-fit of an estimated univariate stochastic differential equation model is described. Using a simple example the method has been demostrated and the method seems to perform well. However, further studies should be performed in order to understand the properties of the method. Furthermore simulations could be used to verify that the type one error level is correct and to study the influence on the power of the test of the number of inter-observation trajectories and the simulation method used.

References

Aït-Sahalia, Y. (1998), 'Maximum likelihood estimation of discretely sampled diffusions: A closed-form approach', National Bureau of Economic Research, Inc., NBER Technical Working Paper number 222. http://netec.wustl.edu/WoPEc/data/Papers/nbrnberte0222.html.

Bak, J. (1998), Nonparametric methods in finance, Master's thesis, Department of Mathematical Modelling, Technical University of Denmark, Lyngby. IMM-EKS-1998-34.

Kendall, M. & Stuart, A. (1961), *Inference and Relationship*, Vol. 2 of *The Advanced Theory of Statistics*, Charles Griffin, London.

Madsen, H., Nielsen, J. N. & Baadsgaard, M. (1998), *Statistics in Finance*, Department of Mathematical Modelling, Technical University of Denmark, Lyngby.

Maybeck, P. S. (1982), *Stochastic models, Estimation and Control*, Academic Press, London.

Ph.D. theses from IMM

1. **Larsen, Rasmus.** (1994). *Estimation of visual motion in image sequences. xiv* + 143 pp.

2. **Rygard, Jens Moberg.** (1994). *Design and optimization of flexible manufacturing systems. xiii* + 232 pp.

3. **Lassen, Niels Christian Krieger.** (1994). *Automated determination of crystal orientations from electron backscattering patterns. xv* + 136 pp.

4. **Melgaard, Henrik.** (1994). *Identification of physical models. xvii* + 246 pp.

5. **Wang, Chunyan.** (1994). *Stochastic differential equations and a biological system. xxii* + 153 pp.

6. **Nielsen, Allan Aasbjerg.** (1994). *Analysis of regularly and irregularly sampled spatial, multivariate, and multi-temporal data. xxiv* + 213 pp.

7. **Ersbøll, Annette Kjær.** (1994). *On the spatial and temporal correlations in experimentation with agricultural applications. xviii* + 345 pp.

8. **Møller, Dorte.** (1994). *Methods for analysis and design of heterogeneous telecommunication networks.* Volume 1-2, *xxxviii* + 282 pp., 283-569 pp.

9. **Jensen, Jens Christian.** (1995). *Teoretiske og eksperimentelle dynamiske undersøgelser af jernbanekøretøjer. viii* + 174 pp.

10. **Kuhlmann, Lionel.** (1995). *On automatic visual inspection of reflective surfaces.* Volume 1, *xviii* + 220 pp., (Volume 2, *vi* + 54 pp., fortrolig).

11. **Lazarides, Nikolaos.** (1995). *Nonlinearity in superconductivity and Josephson Junctions.* *iv* + 154 pp.

12. **Rostgaard, Morten.** (1995). *Modelling, estimation and control of fast sampled dynamical systems.* *xiv* + 348 pp.

13. **Schultz, Nette.** (1995). *Segmentation and classification of biological objects.* *xiv* + 194 pp.

14. **Jørgensen, Michael Finn.** (1995). *Nonlinear Hamiltonian systems.* *xiv* + 120 pp.

15. **Balle, Susanne M.** (1995). *Distributed-memory matrix computations.* *iii* + 101 pp.

16. **Kohl, Niklas.** (1995). *Exact methods for time constrained routing and related scheduling problems.* *xviii* + 234 pp.

17. **Rogon, Thomas.** (1995). *Porous media: Analysis, reconstruction and percolation.* *xiv* + 165 pp.

18. **Andersen, Allan Theodor.** (1995). *Modelling of packet traffic with matrix analytic methods.* *xvi* + 242 pp.

19. **Hesthaven, Jan.** (1995). *Numerical studies of unsteady coherent structures and transport in two-dimensional flows.* Risø-R-835(EN) 203 pp.

20. **Slivsgaard, Eva Charlotte.** (l995). *On the interaction between wheels and rails in railway dynamics.* *viii* + 196 pp.

21. **Hartelius, Karsten.** (1996). *Analysis of irregularly distributed points.* *xvi* + 260 pp.

22. **Hansen, Anca Daniela.** (1996). *Predictive control and identification - Applications to steering dynamics.* *xviii* + 307 pp.

23. **Sadegh, Payman.** (1996). *Experiment design and optimization in complex systems.* *xiv* + 162 pp.

24. **Skands, Ulrik.** (1996). *Quantitative methods for the analysis of electron microscope images. xvi* + 198 pp.

25. **Bro-Nielsen, Morten.** (1996). *Medical image registration and surgery simulation. xxvii* + 274 pp.

26. **Bendtsen, Claus.** (1996). *Parallel numerical algorithms for the solution of systems of ordinary differential equations. viii* + 79 pp.

27. **Lauritsen, Morten Bach.** (1997). *Delta-domain predictive control and identification for control. xxii* + 292 pp.

28. **Bischoff, Svend.** (1997). *Modelling colliding-pulse mode-locked semiconductor lasers. xxii* + 217 pp.

29. **Arnbjerg-Nielsen, Karsten.** (1997). *Statistical analysis of urban hydrology with special emphasis on rainfall modelling.* Institut for Miljøteknik, DTU. *xiv* + 161 pp.

30. **Jacobsen, Judith L.** (1997). *Dynamic modelling of processes in rivers affected by precipitation runoff. xix* + 213 pp.

31. **Sommer, Helle Mølgaard.** (1997). *Variability in microbiological degradation experiments - Analysis and case study. xiv* + 211 pp.

32. **Ma, Xin.** (1997). *Adaptive extremum control and wind turbine control. xix* + 293 pp.

33. **Rasmussen, Kim Ørskov.** (1997). *Nonlinear and stochastic dynamics of coherent structures. x* + 215 pp.

34. **Hansen, Lars Henrik.** (1997). *Stochastic modelling of central heating systems. xxii* + 301 pp.

35. **Jørgensen, Claus.** (1997). *Driftsoptimering på kraftvarmesystemer.* 290 pp.

36. **Stauning, Ole.** (1997). *Automatic validation of numerical solutions. viii* + 116 pp.

37. **Pedersen, Morten With.** (1997). *Optimization of recurrent neural networks for time series modeling. x* + 322 pp.

38. **Thorsen, Rune.** (1997). *Restoration of hand function in tetraplegics using myoelectrically controlled functional electrical stimulation of the controlling muscle.* x + 154 pp. + Appendix.

39. **Rosholm, Anders.** (1997). *Statistical methods for segmentation and classification of images.* *xvi* + 183 pp.

40. **Petersen, Kim Tilgaard.** (1997). *Estimation of speech quality in telecommunication systems.* x + 259 pp.

41. **Jensen, Carsten Nordstrøm.** (1997). *Nonlinear systems with discrete and continuous elements.* 195 pp.

42. **Hansen, Peter S.K.** (1997). *Signal subspace methods for speech enhancement.* x + 226 pp.

43. **Nielsen, Ole Møller.** (1998). *Wavelets in scientific computing.* *xiv* + 232 pp.

44. **Kjems, Ulrik.** (1998). *Bayesian signal processing and interpretation of brain scans.* *iv* + 129 pp.

45. **Hansen, Michael Pilegaard.** (1998). *Metaheuristics for multiple objective combinatorial optimization.* x + 163 pp.

46. **Riis, Søren Kamaric.** (1998). *Hidden markov models and neural networks for speech recognition.* x + 223 pp.

47. **Mørch, Niels Jacob Sand.** (1998). *A multivariate approach to functional neuro modeling.* *xvi* + 147 pp.

48. **Frydendal, Ib.** (1998.) *Quality inspection of sugar beets using vision.* *iv* + 97 pp. + app.

49. **Lundin, Lars Kristian.** (1998). *Parallel computation of rotating flows.* *viii* + 106 pp.

50. **Borges, Pedro.** (1998). *Multicriteria planning and optimization. - Heuristic approaches.* *xiv* + 219 pp.

51. **Nielsen, Jakob Birkedal.** (1998). *New developments in the theory of wheel/rail contact mechanics.* *xviii* + 223 pp.

52. **Fog, Torben.** (1998). *Condition monitoring and fault diagnosis in marine diesel engines.* *xii* + 178 pp.

53. **Knudsen, Ole.** (1998). *Industrial vision. xii* + 129 pp.

54. **Andersen, Jens Strodl.** (1998). *Statistical analysis of biotests. - Applied to complex polluted samples. xx* + 207 pp.

55. **Philipsen, Peter Alshede.** (1998). *Reconstruction and restoration of PET images. vi* + 132 pp.

56. **Thygesen, Uffe Høgsbro.** (1998). *Robust performance and dissipation of stochastic control systems.* 185 pp.

57. **Hintz-Madsen, Mads.** (1998). *A probabilistic framework for classification of dermatoscopic images. xi* + 153 pp.

58. **Schramm-Nielsen, Karina.** (1998). *Environmental reference materials methods and case studies. xxvi* + 261 pp.

59. **Skyggebjerg, Ole.** (1999). *Acquisition and analysis of complex dynamic intra- and intercellular signaling events.* 83 pp.

60. **Jensen, Kåre Jean.** (1999). *Signal processing for distribution network monitoring. x* + 140 pp.

61. **Folm-Hansen, Jørgen.** (1999). *On chromatic and geometrical calibration. xiv* + 241 pp.

62. **Larsen, Jesper.** (1999). *Parallelization of the vehicle routing problem with time windows.xx* + 266 pp.

63. **Clausen, Carl Balslev.** (1999). *Spatial solitons in quasi-phase matched struktures. vi* + (flere pag.)

64. **Kvist, Trine.** (1999). *Statistical modelling of fish stocks. xiv* + 173 pp.

65. **Andresen, Per Rønsholt.** (1999). *Surface-bounded growth modeling applied to human mandibles. xxii* + 116 pp.

66. **Sørensen, Per Settergren.** (1999). *Spatial distribution maps for benthic communities.*

67. **Andersen, Helle.** (1999). *Statistical models for standardized preclinical studies. viii* + (flere pag.)

68. **Andersen, Lars Nonboe.** (1999). *Signal processing in the dolphin sonar system.* *xii* + 214 pp.

69. **Bechmann, Henrik.** (1999). *Modelling of wastewater systems.* *xviii* + 161 pp.

70. **Nielsen, Henrik Aalborg.** (1999). *Parametric and non-parametric system modelling.* *xviii* + 209 pp.