MATHEMATISCHES FORSCHUNGSINSTITUT OBERWOLFACH

T a g u n g s b e r i c h t 54/1993

Model Selection

5. - 11.12.1993

Die Tagung fand unter der Leitung von R. Dahlhaus (Heidelberg) und C. Z. Wei (Taipei) statt.

Ziel der Tagung war es, die verschiedenen Ansätze in der Modellwahl sichtbar zu machen, und den Meinungsaustausch zwischen den einzelnen Wissenschaftlern zu fördern. An der Tagung nahmen 23 Wissenschaftler teil. Es wurden 20 Vorträge gehalten.

Gegenstand der Tagung waren Modellwahlverfahren, wie AIC, BIC, C_p und minimum description length, Bootstrap-Methoden bei der Modellwahl, die Beziehungen zwischen Modellwahl und Parameterschätzung, konkrete Anwendungen in der Regressions- und Zeitreihenanalyse, sowie alternative Methoden der Modellbildung.

Die Tagung war für alle Teilnehmer sehr anregend, und die Vielfalt der Vorträge führte zu einem intensiven Gedankenaustausch über die Probleme des Fachgebietes.

Vortragsauszüge

Modelling of Nonstationary Time Series

Rainer Dahlhaus

Institut für Angewandte Mathematik, Universität Heidelberg

A general procedure is presented for modelling nonstationary time series by parametric models. The parameters are estimated by a minimum distance procedure. More precisely we consider the Kullback-Leibler distance between two stochastic processes that have a time dependent spectral representation. It is shown that this theoretical distance can be written in terms of the time varying spectral densities of the processes. To get an empirical distance function we replace the spectral density of the observed process by the local periodogram. Minimizing the resulting distance leads to the estimator. We prove asymptotic normality of the estimator. Of particular interest is the case where a stationary model is fitted to a nonstationary process. The result gives the estimated value, the theoretically optimal fit and the asymptotic distribution in this case. As a model selection criterion we suggest an AIC-like criterion. The results are demonstrated in a simulation study.

Data Features

Laurie Davies

Fachbereich für Mathematik und Informatik, Universität Essen

Let $\mathcal{X}$ be a sample space and $\mathcal{P}$ a class of distributions on the Borel sets of $\mathcal{X}$. A data feature $\mathcal{F}$ is a mapping of $\mathcal{X}\times\mathcal{P}$ into the two point set $\{0,1\}$. If $\mathcal{F}(x,P) = 1$ with $x\in\mathcal{X}$ and $P\in\mathcal{P}$ then the data x exhibit the feature $\mathcal{F}$ with respect to the model $P\in\mathcal{P}$. Given a number α, $0 < \alpha < 1$, a feature $\mathcal{F}$ is called an α-feature if $P(\mathcal{F}(X(P),P) = 1) \geq \alpha$ for all $P\in\mathcal{P}$. $X(P)$ denotes a random sample in $\mathcal{X}$ with distribution P and it is assumed that the random variables are defined on a common probability space $(\Omega,\mathcal{G},P)$. Given a feature $\mathcal{F}$ the adequacy region based on the data x is defined by $\mathcal{A}(x,\mathcal{F},\mathcal{P}) = \{P: P\in\mathcal{P},\mathcal{F}(x,P) = 1\}$. If $H: \mathcal{P}\rightarrow \mathbf{R}^k$ is some functional then the adequacy region for the values of $H(P)$, $P\in\mathcal{P}$, is given by $H(\mathcal{A}(x,\mathcal{F},\mathcal{P})) = \{H(P): P\in\mathcal{P}, \mathcal{F}(x,P) = 1\}$. Examples of features based on weaker metrics, order statistics, runs, location and scale functionals and deviations from spectral density functions were given.

Assessment and Propagation of Model Uncertainty

David Draper

School of Mathematical Sciences, University of Bath

In most examples of inference and prediction, the expression of uncertainty about unknown quantities y on the basis of known quantities x is based on a model M that formalizes assumptions about how x and y are related. M will typically have two parts: *structural* assumptions S, such as the form of the link function and choice of error distribution in a generalized linear model, and *parameters* θ whose meaning is specific to a given choice of S. It is common in statistical theory and practise to acknowledge parametric uncertainty about θ given a particular assumed structure S; it is less common to acknowledge structural uncertainty about S itself. A widely used approach, in fact, involves enlisting the aid of x to specify a single "best" choice S^* of S, and then proceeding as if S^* were known to be correct. In general this approach fails to fully assess and propagate structural uncertainty, and may lead to miscalibrated uncertainty assessments about y given x. When miscalibration occurs it is often in the direction of understatement of uncertainty about y, leading to inaccurate scientific summaries and overconfident decisions that do not incorporate sufficient hedging against uncertainty. In this talk I discuss a Bayesian approach to solving this problem, based on integrating over structural uncertainty as in the expression
$\mu(y \mid x) = \int \mu(y \mid x,S)\, \mu(S \mid x) dS$, which has long been available in principle but is only now becoming routinely feasible by virtue of recent computational advances, and illustrate its application in several examples.

On Finite-Sample Properties of Model Selection Procedures

Bernd Droge

Institut für Stochastik, Humboldt-Universität zu Berlin

We derive finite-sample properties of procedures for the selection of linear regression models under the assumption of normally distributed errors. The selection of an "appropriate" model may be regarded as a smoothing problem, and will usually be done in a data-dependent way. A straightforward application of a result by P.J. Kempthorne provides that, under a normalized squared error loss, all selection procedures (data-dependent or not) are admissible. Furthermore, the minimax approach provides an interpolating estimator of the regression function, which is often

impractical. Therefore, within a certain class of selection procedures an optimal one is determined using the minimax regret principle. This optimal procedure depends on the degree of freedom for estimating the unknown error variance, and behaves similar to the minimum-C_p-procedure.

Graphical Methods for Applications - Dependent Model Selection

David Findley

U.S. Bureau of the Census, Washington

In practice, statistical models, at their best, only capture some of the features of the data to which they are fit. This limitation means that different uses of the same data, such as long-term forecasting versus short-term forecasting, often require different models. If an application - appropriate criterion can be found, such as mean square forecast error, which can be shown to converge appropriately as the data set is expanded, then by observing the relative movements of this statistic for two competing models fit to a subset of the data which is expanded one point at a time to the full data set, one can often demonstrate the persistent superiority of one model. In the forecasting situations we discussed, both empirically and theoretically, this superiority is demonstrated by basically linear movement, with non-zero shape, in the graph of the accumulating sums of differences of squared forecast errors. We presented a uniform almost sure convergence result for sample-size normalized sums of squared errors over compact families of invertible ARMA models to provide theoretical support for the method. The data were only required to have second order moments which can be estimated a.s., so they can be far removed from ARMA data.

On the Identification of a Bilinear System

Jürgen Franke

Fachbereich Mathematik, Universität Kaiserslautern

We discuss the problem of identifying "large" bilinear systems (i.e. high dimension, very large samples) with, e.g., state space representation:

$$x(t+1) = Fx(t) + Gu(t) + Nu(t)\otimes x(t) + \varepsilon(t),$$
$$y(t) = Hx(t) + \tilde{\varepsilon}(t).$$

We describe a simple procedure inspired by Guidorzi's algorithm for linear systems which basically is a stepwise regression (backwards) procedure for a "linearized" input-output representation of the system. We illustrate the ability of the procedure to cope with large systems by applying it to data from a test rig for cars.

Predictive Stochastic Complexity

László Gerencsér

Computer and Automation Institute, Hungarian Academy of Science, Budapest

Predictive stochastic complexity is defined for Gaussian ARMA-processes as $\sum_{n=1}^{N} \varepsilon_n^2(\hat{\theta}_{n-1})$, where $\varepsilon_n(\theta)$ is the reconstructed innovation process, assuming that the true ARMA parameter vector is θ, and $\hat{\theta}_n$ is the maximum-likelihood estimator of the true parameter vector, say θ^*. One of the main results that was presented is that $\lim_{N\to\infty} \sum_{n=1}^{N} (\varepsilon_n^2(\hat{\theta}_{n-1}) - e_n^2) / \sigma^2(e) \log N = p + q$ almost surely, where (e_n) is the innovation process, and p,q are the ARMA orders.

This result can be extended to the case when estimation with forgetting is applied. Let the forgetting rate be λ, $0 < \lambda < 1$. A typical result is then that

$$E(\varepsilon_n^2(\hat{\theta}_{n-1}) - e_n^2) = \sigma^2(e) \frac{\lambda}{2}(p + q)\,(1 + O(\lambda^c)), \quad c > 0.$$

How Many Terms Should be Added into an Additive Model

Wolfgang Härdle* and A. B. Tsybakov

Fachbereich Wirtschaftswissenschaften, Humboldt-Universität Berlin

Institute for Problems of Information Transmission, Academy of Sciences, Moskow

Smoothing in high dimensions faces the problem of data sparseness. Additive regression models alleviate this problem by fitting a sum of one-dimensional smooth functions. Given a set of predictor variables, some of these functions could actually be zero, so that a further simplification of high dimensional data analysis occurs. A three-stage procedure is proposed here to decide how many and which components should be added into such an additive model. In a first step the predictor variables are made orthogonal by a principal component transformation. After the second step, determining

the number and sequence of components, the model is fit by the kernel method. The asymptotic distribution of this regression estimate is given. The practical performance is investigated via a simulation study.

On the Extended Information Criterion BIC and its Use

Makio Ishiguro*, Y. Sakamoto and G. Kitagawa

The Institute of Statistical Mathematics, Tokyo

Akaike proposed AIC as an estimate of the expected log likelihood to evaluate the goodness of models fitted to a given set of data. The introduction of AIC has greatly widened the range of applications of statistical methods. However, its limit lies in the point that it can be applied only to the cases where the parameter estimation is performed by the maximum likelihood method.

AIC was derived performing necessary integration utilizing the asymptotic normality of M.L.E..

Using bootstrap technique to perform the integration, we can extend AIC to EIC which can be used to evaluate the goodness of models whose parameters are not necessarily estimated with maximum likelihood procedures.

Bootstrap Autoregressive Order Selection

Jens-Peter Kreiss

Institut für Mathematische Stochastik, Technische Universität Braunschweig

It is dealt with the problem of fitting an autoregression of finite order p to given data $X_1, \dots, X_n$ generated by a stationary autoregressive process $(X_t: t \in Z)$ with infinite order, i.e.

$$X_t = \sum_{\nu=1}^{\infty} a_\nu X_{t-\nu} + \varepsilon_t \qquad \text{for all } t \in Z = \{0, \pm 1, \pm 2, \dots\}.$$

The white noise process $(\varepsilon_t: t \in Z)$ consists of i.i.d. random variables with zero mean and finite variance σ^2.

On the basis of the final prediction error (FPE) an order selection procedure is proposed in which one relevant term can be approximated using the bootstrap. The advantage of the proposal is, that the bootstrap order selection allows for different kinds of parameter estimates and is not restricted to Yule-Walker parameter estimates, as, for

example, the Akaike information criterion (AIC) is.

The proposed procedure has the wanted property, that the more precise one can estimate the unknown parameters the higher one probably wants to choose the order p of the fit.

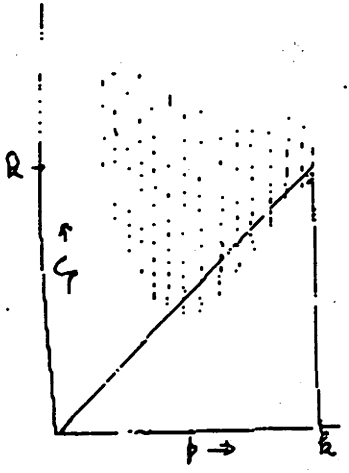
Some asymptotic results are proved and also simulations were shown, which demonstrate the behaviour of the proposal. The research is jointly with Jürgen Franke, Kaiserslautern.

More Comments on C_p

Colin Mallows

AT&T Bell Labs, Murray Hill

A very common configuration on a C_p-plot (when the number of independent variables is large) is as shown:



with no single subset being strongly indicated as "best". A simple mechanism that generates such a configuration is the following: assume the true regression coefficients are spread out approximately uniformly near the origin, with k-local density λ. In the case of a standardized orthogonal design, the minimum-C_p-values will then be approximately

$$\min C_p = k + \frac{q^3}{12\lambda^2} - 2q \quad (q = k - p)$$

and the average predictive mean square error of the "min C_p least-squares" rule will be

$$k + \frac{q^3}{12\lambda^2}$$

i.e. > k for all q > 0. Simulations show that similar results hold for correlated regressors, and demonstrate that when such a configuration is seen, prediction using a subset-LS estimate is not a good idea. Some kind of shrinkage is esssential. A pseudo-Bayesian method based on a "spike + slab" prior is suggested.

Optimal Smoothing in Adaptive Location Estimation

Enno Mammen

Institut für Stochastik, Humboldt-Universität zu Berlin

The problem of estimating the location parameter of an i.i.d. sample is considered for the case that the location density is symmetric and unknown. This is a classical problem going back to the fifties and some estimation procedures have been proposed over the years which attain the same asymptotic efficiency as for the location model with known density f. Typically, the procedures are based on two steps. In the first step the efficient score function f'/f is estimated. Then a linear estimate using this estimate of f'/f and a preliminary estimate of θ. In this talk some higher order analysis is presented for two step procedures based on kernel estimation of f'/f. We show that higher order optimality for estimating the location parameter will *not* be reached by optimal estimation of f'/f. Suboptimal estimates of f'/f will give smaller asymptotic variances. The influence of the preliminary estimate on the operation characteristics of the adaptive estimate is also discussed.

"Random-Sets"-Procedures

D.W. Müller

Institut für Angewandte Mathematik, Universität Heidelberg

This talk presents the excess mass approach in statistics. The basic idea of this approach is to measure the amount of probability mass not fitting a given statistical

model. It came up in the context of modality testing (Müller & Sawitzki 1987, 1991), was later applied to density estimation (Polonik 1992) and even extended to regression (Müller 1992). Müller & Sawitzki proposed the excess mass difference statistic for testing unimodality (mentioned already in an Oberwolfach conference 1981). The asymptotics of this statistic for regularly unimodal and uniform distributions is described. Polonik (1992) has extended these results (using different methods) to the higher dimensional case. He applied empirical process methods. According to Polonik the density estimator of Müller & Sawitzki ("the silhouette") is a generalization of the Grenander estimator. The ideas can be carried over per analogiam to quantile regression. Here they lead to minimax correlation procedures. The talk presents simulation results and asymptotic statements explaining the remarkably weak dependence of these procedures on the underlying probability structure.

Consistency Properties of Model Selection Criteria in Multiple Linear Regression

Marlene Müller

Fachbereich Wirtschaftswissenschaften, Humboldt-Universität zu Berlin

The talk concerns consistency properties for a class of model selection criteria in multiple linear regression. This class covers wellknown criteria as e.g. Mallows' C_p, CV (cross-validation), GCV (generalized cross-validation), Akaike's AIC and FPE as well as Schwarz' BIC. All these criteria are shown to be consistent in the sense that the probability of selecting the true or larger models converges to 1 if the sample size is growing to infinity. The proofs assume i.i.d. errors with $Ee_i = 0$ and $De_i = \sigma^2$ and allow a possible inadequacy of the linear model. For a subclass of these criteria (BIC-type criteria, suitable corrected C_p-criteria) a stronger consistency property is proved (asymptotic choice of the true model in probability). Upper bounds for the speed of convergence are obtained for some more assumptions on the error distribution (existence of higher order moments or normal errors). Convergence results for loss and risk of the regression fit using the selected model complete these investigations.

A Generalized Final Prediction Error for Robust Regression

Deborah Nolan

Department of Statistics, University of California, Berkeley

Akaike's Final Prediction Error (FPE) minimizes an estimate of the expected squared error in predicting new, independent observations. This criterion was designed for models fit by least squares. A modification of FPE is required for models fit by other loss functions, such as least absolute deviation regression. It is shown here one can mimic the form of FPE for a general ρ function, provided: ρ is convex, has a unique minimum at 0, $E\rho(\varepsilon + t)$ is twice differentiable in t, and $E\rho_1(\varepsilon) = 0$ where ρ_1 is an approximative derivative for ρ and ε is the error in the true model. The modified FPE for ρ is

$$\sum_{i=1}^{n} \rho(y_i - x_i'\hat{\beta}) + k\hat{\sigma}^2/\hat{R} \quad \text{where}$$

$\hat{\sigma}^2 = \frac{1}{n-k}\sum_{i=1}^{n} \rho_1(y_i - x_i'\hat{\beta})^2$, $\hat{R} = \frac{1}{n-k}\sum_{i=1}^{n} R(y_i - x_i'\hat{\beta})$, R is the second derivative of $E\rho(\varepsilon + t)$, $\hat{\beta}$ is the minimizer of $\sum \rho(y_i - x_i'\beta)$.

The proposed generalization is the result of joint work with Prabir Burman at the University of California at Davis.

Asymptotic Equivalence of Discrete And Continuous Models

Michael Nussbaum

Institut für Angewandte Analysis und Stochastik, Berlin

We investigate the relationships of discrete and continuous versions of some statistical models, in terms of their deficiency distance Δ (Le Cam's distance between experiments). The first example is the asymptotic equivalence between the signal-in-white noise problem and Gaussian nonparametric regression, which was found by Brown and Low (1992). The second example is the problem of density estimation from an i.i.d. sample, which is asymptotically equivalent to a white noise model with a signal which is the root of the density. The third example is a nonparametric stochastic differential equation model, i. e. a model of small diffusion type where the function which governs the drift term varies in a nonparametric set. It is shown that an Euler difference scheme as a discrete version of the stochastic differential equation is asymptotically equivalent in the sense of Le Cam's deficiency distance, when the

discretization step decreases with the noise intensity. We thus obtain a nonparametric version of diffusion limit results for autoregression. It follows that in the continuous diffusion model, discrete sampling on a uniform grid is asymptotically sufficient. The key technical step utilizes the notion of Hellinger process from semimartingale theory.

Model Selection and Inference

Benedikt M. Pötscher

Institut für Statistik und Informatik, Universität Wien

Frequently, when fitting a model to data, the choice of the model itself is based on the same data set. As a consequence, the standard distributional results for the parameter estimators do no longer apply. If inference is nevertheless based on these distributional results, then this inference becomes invalid. For example, confidence intervals will typically be too short. In this talk we derive the correct asymptotic distribution of M-estimators when the model is selected from a set of nested models by a multiple testing procedure. A numerical example illustrates that the difference between the correct asymptotic distribution and the "naive" asymptotic distribution (i.e., ignoring the model selection process) can be substantial. Some results concerning the construction of asymptotically correct confidence sets from the asymptotic distribution are also given.

Rounding Probabilities

Friedrich Pukelsheim

Institut für Mathematik, Universität Augsburg

Let the weights $W_1, \ldots, W_c$ be uniformly distributed on the probability simplex of $\mathbf{R}^c$. Given a multiplier $v \geq 0$, the q-stationary roundings are defined by

$$n_i = R_q(vW_i) \Leftrightarrow vW_i \in (n_i - 1 + q,\ n_i + q).$$

We show that the multiplier $v_n = n + (q - \frac{1}{2})c$ is unbiased for n up to an error term of order n^{-1}, $E\left[\sum_{i=1}^{c} R_q(v\, W_i)\right] = n + O(n^{-1})$.

Minimum Description Length and Model Selection

Bin Yu

Department of Statistics, University of California, Berkeley

Rissanen's description length of data based on a model class has three standard forms: predictive, two-stage, and mixture form. These three forms can be used in regression variable selection. In the normal regression context when the true model is assumed finite-dim, these three forms are compared with other criteria in terms of overfitting and underfitting probabilities and in terms of two prediction errors. (cf. Speed and Yu, JISM, 1993). These three forms are seen to be asymptotically equivalent in this model and other nice parametric models. They are optimal in term of prediction as well.

The same three forms of MDL can be studied in the context of nonparametric density estimation. The model class is smooth bounded densities on [0,1] with bounded derivatives and the approximating classes are histograms with equal binwidths. Yu and Speed (PTRF, 1993) gives a predictive coding scheme which achieves the lower bound given in the same paper. The work in progress with Barron shows that a smart two-stage coding has the same optimality and gives optimal histogram density estimator.

Choice of Penalty Term in FPE Criterion

Ping Zhang

Department of Statistics, University of Pennsylvania

The talk summarizes some recent work of mine in an effort of trying to unify the practice of model selection using the generalized final prediction error criterion. We establish a natural connection between FPE and random walk. Using the classical results of Frank Spitzer on random walk, we suggest the use of a penalty term between 3 and 4 for the FPE criterion. Applications of our results to multifold cross validation method is also discussed. The talk concludes with brief review of progresses we have made in decision theoretic approaches of model selection.

Berichterstatter: R. Dahlhaus

Tagungsteilnehmer

Prof.Dr. Rainer Dahlhaus
Institut für Angewandte Mathematik
Universität Heidelberg
Im Neuenheimer Feld 294

D-69120 Heidelberg

Prof.Dr. Jürgen Franke
Fachbereich Mathematik
Universität Kaiserslautern

D-67653 Kaiserslautern

Prof.Dr. P. Laurie Davies
FB 6 - Mathematik und Informatik
Universität-GH Essen

D-45117 Essen

Prof.Dr. Laszlo Gerencser
Computer and Automation Institute
Hungarian Academy of Science
MTA Sztaki
P.O. Box 63

H-1518 Budapest

Prof.Dr. David Draper
Statistics Group
School of Mathematical Sciences
University of Bath
Claverton Down

GB-Bath, Avon BA2 7AY

Günter Hainz
Institut für Angewandte Mathematik
Universität Heidelberg
Im Neuenheimer Feld 294

D-69120 Heidelberg

Dr. Bernd Droge
Institut für Stochastik
Humboldt-Universtität Berlin

D-10099 Berlin

Prof.Dr. Wolfgang Härdle
Inst.für Statistik und ökonometrie
Humboldt Universität zu Berlin
FB Wirtschaftswissenschaften

D-10099 Berlin

Prof.Dr. David F. Findley
Stat. Res. Div.
US Bureau of the Census

Washington , DC 20233
USA

Prof.Dr. Makio Ishiguro
Institute of Statistical
Mathematics
4-6-7 Minami Azabu, Minato-ku

Tokyo 106
JAPAN

Prof.Dr. Jens-Peter Kreiss
Institut für Mathematische
Stochastik der TU Braunschweig
Pockelsstr. 14

D-38106 Braunschweig

Prof.Dr. Deborah Nolan
Department of Statistics
University of California
367 Evans Hall

Berkeley , CA 94720
USA

Prof.Dr. Colin L. Mallows
AT & T Bell Laboratories
P.O. Box 636
600 Mountain Avenue

Murray Hill , NJ 07974-2070
USA

Dr. Michael Nussbaum
Institut für Angewandte Mathema[redacted]
und Stochastik
Hausvogteiplatz 5 - 7

D-10117 Berlin

Prof.Dr. Enno Mammen
Institut für Stochastik
Fachbereich Mathematik
Humboldt-Universität Berlin

D-10099 Berlin

Prof.Dr. Benedikt M. Pötscher
Institut für Statistik und
Informatik
Universität Wien
Universitätsstr. 5/3

A-1010 Wien

Prof.Dr. Dietrich Werner Müller
Institut für Angewandte Mathematik
Universität Heidelberg
Im Neuenheimer Feld 294

D-69120 Heidelberg

Prof.Dr. Friedrich Pukelsheim
Lehrstuhl für Stochastik und ihre
Anwendungen
Institut für Mathematik
Universität Augsburg

D-86135 Augsburg

Marlene Müller
Inst.für Statistik und ökonometrie
Humboldt Universität Berlin
FB Wirtschaftswissenschaften
Spandauer Str. 1

D-10178 Berlin

Prof.Dr. Ludger Rüschendorf
Institut für Mathematische
Stochastik
Universität Freiburg
Hebelstr. 27

D-79104 Freiburg

Prof. Dr. Ching-Zong Wei
Institute of Statistical Science
Academia Sinica

Taipei , 11529
TAIWAN

Prof.Dr. Bin Yu
Department of Statistics
University of California
367 Evans Hall

Berkeley , CA 94720
USA

Prof.Dr. Ping Zhang
Department of Statistics
The Wharton School
University of Pennsylvania

Philadelphia , PA 19104-6302
USA