Optimal Estimation when the Parameter Space is of Infinite Dimension

Junnan He and Werner Ploberger

Department of Economics, Washington University in St. Louis

Abstract

Many classical nonparametric estimation problems (density estimation, nonparametric estimation of a regression function, estimation of the spectrum of a stationary process) can be reduced to the estimation of an infinite dimensional parameter vector. Typical examples are the representation of a function by a series (Fourier,..), or the estimation of a spectrum by a very long, (possibly infinite) autoregressive process. We adapt concepts of the classical Hajek-Blackwell-LeCam theory to develop a theory of asymptotically optimal estimation of the parameter in our case. Maximum likelihood estimators do not exist in most of the cases, so we have no "canonical" candidate for a "best" estimator. We define suitable loss functions for the estimation error, which allow us to uniquely characterize some estimators. When estimating functions, it is quite common to assume higher order differentiability or some other smoothness conditions. We construct some simple prior distributions for the parameter which force the parameter to obey said smoothness condition. Our results show that certain estimators are asymptotically efficient. We analyze estimators constructed as a function of "sieve estimators". A sieve estimator is constructed by choosing a number - say m (which increases to infinity, just as the sample size increases) - and then simply estimate the first m parameters by setting all the other parameters to zero. We show that estimators of the following type are asymptotically optimal :1. "Shrunken sieve estimators": The estimator is constructed by "shrinking" the sieve estimator (i.e. multiplying with matrix which makes it smaller - analogous to Ridge estimators or Bayesian estimators in linear models. 2. The sieve estimator, whose length is determined by AIC, BIC: The variance of the prior distribution of the components of the parameter converge to zero exponentially. 3. If the

decay in the variance of the components is only of polynomial order, then the optimal estimator cannot be represented as a sieve estimator.

1. Introduction

In some joint papers, Peter C.B. Phillips and one of the authors investigated optimal estimation and inference under nonstandard conditions (Phillips and Ploberger(2012), Ploberger and Phillips(2003)). In this paper, we want to apply the methodology similar to these papers to derive admissible estimators in general nonparametric settings.

There is a huge literature on nonparametric estimators, but relatively little is known about their optimality properties. In one of his landmark papers, (Andrews(1991)), Don Andrews investigated a lot of estimators for the asymptotic variance of an estimator, which essentially boils down to estimate the long-term variance of the score process. Right at the beginning of his paper, he states that "Currently the consistency of these estimators has been established, but their relative merits are unknown". Obviously a lot of progress has been made. An overview over recent developements can be found in Gine and Nickl(2016) or Armstrong and Kolesar(2018).

Under standard regularity conditions, a theory of "optimal" estimation for finite-dimensional parameters developed by Le Cam, Blackwell and Hayek (cf. van der Vaardt(2000), Strasser(1996)) is now a well-established part of statistical science. This theory allows a characterization of the maximum-likelihood estimator as "best estimator" compared to a large class of competing estimators. Here, in this paper, we will investigate the case of a parameter space of infinite dimensions. In particular, we assume that our parameter space Θ is a subset of the $\boldsymbol{R}^{\boldsymbol{N}}$, where $\boldsymbol{N} = \{1, 2, \ldots\}$ is the set of all natural numbers. So our parameters θ are sequences,

$$\theta = (\theta_1, \theta_2, \ldots) \ .$$

We assume that we have given a squence of data - at time n, our information is contained in the σ-algebra $\mathfrak{F}_n$ and for each θ a measure P_θ on the σ-algebra $\mathfrak{F} \supseteq \sigma$-algebra $\mathfrak{F}_n$. Although this formulation seems quite different to many problems in nonparametric estimations, any classical nonparametric models can be formulated within this framework. Interpretation of the θ_i as Fourier coefficients allows us to consider any problem parametrized by reasonable functions as part of our class. As examples, I would like to consider three

traditional nonparametric problems: Nonparametric modeling of Gaussian processes, regression models with Gaussian errors, and density estimation. Theses examples are for illustration only: they should show that typical nonparametric estimation problems can be cast in our framework.

- General models for stationary Gaussian process: Parameter of interest here is in most cases the spectral density f. Assume that the data y_t are generated by an infinite autoregressive process

$$y_t = \sum_{k \geq 1} \gamma_k y_{t-k} + u_t \ ,$$

  where u_t is Gaussian white noise uncorrelated with y_{t-i}. Then

$$f(\exp(i\lambda)) = \frac{\sigma_u^2}{|1 - \sum_{k \geq 1} \gamma_k \exp(ki\lambda) \ |^2}$$

  So our parameter $\theta = (\theta_1, \theta_2, ..) = (\sigma_u^2, \gamma_1, \gamma_2, \ldots)$, and it is quite an easy task to set up the likelihood as a function of θ .

- Nonparametric regression with e.g. Gaussian errors. Assume - for the sake of simplicity - that we only have a scalar regressor x_t, which we assume to take values in a fixed interval. Without limitation of generality, we can assume that this interval equals $[0, \pi]$. Let y_t be the dependent variable . Then our model is of the form

$$y_t = f(x_t) + u_t$$

  where the u_t are i.i.d. $G(0, \sigma^2)$. Then the function f can be written as a Fourier series

$$f(x) = \sum_{n=0}^{\infty} \gamma_n \cos(nx) \ .$$

  Again, our parameter $\theta = (\theta_1, \theta_2, ..) = (\sigma_u^2, \gamma_0, \gamma_1, \ldots)$, and we can then write down the likelihood.

- Density estimation can be analyzed in our context, too. Assume for simplicity that we have a sample of i.i.d random variables X_i , taking values in an interval $[a, b]$. We assume that we want to estimate the density f of the random variables' distribution. For simplicity, assume

that $\ln f$ is a square integrable function. Then we can choose a complete orthonormal set of functions φ_n (e.g. trigoniometric functions), and write

$$\ln f(.) = C(\gamma_1, \ldots) + \sum \gamma_n \varphi_n(.) \ .$$

$C(., \ldots)$ has to be chosen in such a way that.

$$\int \exp(\ln f) = 1 \ ,$$

and thus is a function of the γ_n. Then our $\theta = (\theta_1, \theta_2, ..) = (\gamma_1, \ldots)$, and we can define a likelihood.

One of the problems, however, is that the maximum-likelihood estimator does not work in infinite-dimensional settings. First of all, we will have problems defining the maximum-likelihood estimator, since there is no guarantee that the likelihood function will have a minimum. Moreover, it is quite easy to construct examples where the maximum likelihood estimator is inconsistent.

In all of the examples given above, the parameter vector θ, is related to a function - be it a density, a regression function or a transfer function. In nonparametric situations, it is often assumed that the function is smooth - differentiable up to a certain order, or even more. If the parameters are Fourier coefficients, then differentiability of the underlying function is determined by the decay of the coefficients: If (θ_1, θ_2) represent the Fourier coefficients of the function

$$f(\omega) = \sum \theta_k exp(ik\omega) \ .$$

then the Fourier coefficients of the m-th derivative $f^{(m)}$ equal $i^m m^k \theta_k$. Hence for $f^{(m)}$ to be square integrable, we have

$$\sum \theta_k^2 m^{2k} < \infty \ . \tag{1}$$

So imposing growth conditions on the coefficients is essentially equivalent to the requirements of varying degrees of smoothness of the underlying function. We will assume that the prior distributions on the set of parameters are essentially independent Gaussian distributions with expectations zero and variances c_k^2, where c_k^2 converges to zero. Consequently, parameters with larger index will be very small.

2. Main Theorems

Our primary goal is the estimation of θ, and define criteria to compare estima-tors, and especially finding the "best" estimator. We will use an adaption of a technique used quite often in the finite dimensional context. In order to find the asymptotically optimal estimator, one first establishes that the posterior distribution is asymptotically $G(\hat{\theta}, \hat{\Sigma})$, and thereby simplifying the problem. Then it is quite plausible that $\hat{\theta}$ is the "optimal" estimator. The importance of "conditional Gaussianity" was first recognized by Kim(1998). Ploberger and Phillips (2012) utilized this property to characterize estimator in cases of stochastic information matrices.

Our first task is to establish that the posterior distribution is asymptotically normal: To do so, we need only a few more than the standard assumptions.

First of all let us assume that the conditional log-likelihoods

$$\ell_t(\theta) = \ln p_\theta(x_k / x_{k-1}, \ldots)$$

are 3 times di¤erentiable, all have uniformly bounded second moments, and the requirements for all the CLTs for scores and information-matrix-type theorems are fulfilled. Furthermore, we assume that the eigenvalues of the expected in-formation matrix are bounded from above and below. I think this is quite a plausible requirement, since it guarantees that all unidimensional restrictions of the parameter (i.e. curves of parameters) allow for standard ML estimation. Furthermore we assume that the "long-run" variances for all the derivatives of the log-likelihood are uniformly bounded.

Heuristically, for large n the average of n such expression differs from its expectation by $O(1/\sqrt{n})$. We also assume that for some bounded "neighborhood" O of the

$$E_\theta \left(\sup_{\theta \in O} \left| \frac{\partial^3 \ell_t(\theta)}{\partial \theta_i \partial \theta_j \partial \theta_k} \right| \right) \leq M \tag{2}$$

We do not want to assume that the parameter space Θ is the whole R^N This may be inconvenient. Consider e.g. the case of z-transforms of autoregressive parameters. The parameters have to be such that there are no zeroes in or on the unit circle. It would be very inconvenient to describe this set of these parameters directly. We will later on describe some of the assumptions of the parameter set. We assume, however, that our parameterspace is an open set in a topology where consistent estimation of the parameter θ is possible.

We construct our prior distribution by starting with a product of Gaussian distributions with zero mean and variances c_i^2. The support of theses measures is, however, the whole space R^N. So we have to restrict our prior distribution to our parameter space Θ. Let $C^{-1} = diag(c_i^2)$. To exclude trivial cases, we assume that

$$G(0, C^{-1})(\Theta) > 0.$$

Let us suppose that

$$k^8 c_k = o(1). \tag{3}$$

This condition is a bit stringent: Essentially we assume that the function de-scribing the "parameter θ is differentiable eight times - a bit extreme even for non-parameteric statistics. I am convinced that it is possible to reduce the smoothness requirements to a reasonable form. We will be rather wasteful when we compute bounds for matrices with increasing dimensions. Let us now define

$$(A_n)_{i,j} = \left(\sum_{t=1}^{n} \frac{\partial^2 \ell_t(\hat{\theta})}{\partial \theta_i \partial \theta_j} \right)$$

Then we have the following theorem:

Theorem 1. *Assume the above conditions are met. Let Π_n be the posterior distribution of the parameter θ (which is a random probability measure on R^N). Then the total variation of the difference between Π_n and $G((A_n + C)^{-1} A_n \tilde{\theta}_n, (A_n + C)^{-1})$ converges to zero in probability with respect to P_n, where $\widetilde{\theta}_n$ is the (standard) ML estimator.*

The "mean" of the posterior distribution is not the ML-estimator, but some kind of "shrinkage estimator" We can think of the estimator as a linear combination of ML-estimator and prior mean, which we have set to zero. Moreover, it can easily be seen that $(A_n + C)^{-1} A_n \hat{\theta}_n$ is asymptotically equivalent to the penalized ML estimator which maximizes

$$\sum \ell_t(\theta) - \theta' C \theta / 2. \tag{4}$$

The proof is very technical, so I have put it in an appendix. Any Bayesian estimator - and especially the conditional mean of a parameter is, by construction, admissible. The conditional mean of a parameter is evidently the estimator with minimum variance. Here, however, we have the problem that

our estimator is not exactly the Bayesian estimator, but only so asymptotically. Hence we do not know if even the expectation of the estimator exists. Nevertheless, minimizing quadratic distances seems to be a worthwhile goal. As starting point, we could take an arbitrary estimator $\tilde{\theta}$ and some matrices B_n (which select the components we are interested in) and consider the quadratic distance

$$Q(\tilde{\theta}) = (\theta - \tilde{\theta})\prime B_n B_n\prime(\theta - \tilde{\theta}).$$

typical examples would be:

- B_n projects on finitely many components of θ: finitely many parameters;
- $B_n\prime$ = (1,1,...):Sum of parameters, "long-term" parameters. It now would be natural to try to minimize "E"$Q(\tilde{\theta})$ over all estimators. There are, however, some problems:

1. We have to define the expectation: This can be accomplished easily, just define a prior distribution on parameter.

2. A more serious problem is the fact that for many estimators, $Q(\tilde{\theta})$ may very nice asymptotic properties, but the expectation does not exist, or would be hard to compute. Typical examples are the usual ML estimators for finite-dimensional parameters. The estimation errors are asymptotically normal, but the exact moments are often unknown. Therefore it might be a good idea to classify estimators according to

$$Ef(Q(\tilde{\theta})),$$

where E is the expectation wrt to prior(s) and f is from a class of "squashing function": They should be bounded, but the class should be large enough to approximate the identity function.

A "Classical Example" is Le Cam theory. A function $\phi(.)$ defined on a finite-dimensional vector space is called "bowl-shaped" if it is bounded and its level-sets are symmetric and convex In this context, Anderson's lemma (cf. Strasser(1995)) guarantees that the expectations of all bowl-shaped functions of estimation errors are minimized by the mean of the asymptotic normal distribution.

In fact, Anderson's lemma allows us our first asymptotic result: Let $b = (b_1, ...)$ be a sequence so that

$$\sum b_k^2/c_k < \infty. \tag{5}$$

Then the difference of the posterior distribution of $b'\theta$ and $G(b'(A_n + C)^{-1}A_n\widetilde{\theta}_n,\ b'(A_n + C)^{-1}b)$ converges to zero. Let

$$\sigma_n = \sqrt{b'(A_n + C)^{-1}b.} \tag{6}$$

Then the posterior distribution of

$$(b'\theta - b'(A_n + C)^{-1}A_n\widetilde{\theta}_n) \tag{7}$$

converges to a standard normal. Now let $f(.)$ be a "bowl shaped" function, defined on the real line.Then we have the following theorem:

Theorem 2. *Let $\widehat{\mu}_n = b'(A_n + C)^{-1}A_n\widetilde{\theta}_n$, b be a vector satisfying (5), and σ_n defined by (6). Then for any "bowl shaped" function f and any other estimator $\overbrace{\mu_n}$ we have*

$$\liminf_{n\to\infty} Ef((\widehat{\mu}_n - b'\theta)/\sigma_n) - Ef((\overbrace{\mu}^{\frown}_{n} - b'\theta)/\sigma_n) < 0.$$

So the "best" estimator for $b'\theta$ is $b'(A_n+C)^{-1}A_n\widetilde{\theta}_n$. For another approach we use a smaller class of squashing functions, but allow for more general "scaling matrices".

We assume that

$$f = Const - g,$$

where $Const$ is a constant and g is "completely monotone": differentiable infinitely often, and

$$(-1)^k f^{(k)}$$

is negative.

Typical examples: $\exp(-sx)$, $\frac{a}{b+cx^a}$, Consequence: For arbitrary $a > 0$,

$$f(x) = a\frac{x}{x+a} = x(\frac{1}{1+x/a})$$

falls in our class. Let

$$A_n = \sum \left(\frac{\partial \ell}{\partial \theta}\right)\left(\frac{\partial \ell}{\partial \theta}\right)\prime \approx -\sum \frac{\partial^2 \theta}{\partial \theta^2}$$

Then we call B_n "reasonably normed" if and only if

$$tr(B_n\prime(A_n + C)^{-1}B_n) = O(1).$$

Furthermore, observe that

$$B_n\prime(A_n + C)^{-1}B_n$$

$$= B_n\prime\sqrt{C}^{-1}(\sqrt{C}^{-1}A_n\sqrt{C}^{-1} + I)^{-1}\sqrt{C}^{-1}B_n$$

Theorem 3. *Let us assume that total variation of the difference the posterior distribution for the parameter θ and $G(\hat{\theta}, (A_n + C)^{-1})$ converges to zero for some estimator $\hat{\theta}$. Let $\tilde{\theta}$ be an arbitray esimator. Then the following propositions are equivalent:*

1. *For " reasonably normed" B_n, $(\tilde{\theta} - \hat{\theta})\prime B_n B_n\prime(\tilde{\theta} - \hat{\theta})$ does not converge to 0 stochastically wrt P_n*
2. *For one (nontrivial) of our loss functions f*

$$\lim \sum (Ef(Q_n(\tilde{\theta})) - Ef(Q_n(\hat{\theta}))) > 0.$$

3. *For all of our loss functions f,*

$$\lim \sum (Ef(Q_n(\tilde{\theta})) - Ef(Q_n(\hat{\theta}))) > 0.$$

Proof: The structure of completely monotonic functions is well known. Bernstein's Theorem (cf. Bernstein (1928)) guarentees that every completely monotonic function f can be written as

$$g(x) = \int_0^\infty \exp(-sx) d\mu(x).$$

Hence it suffices to analyse

$$E_n \exp(-sQ_n(\tilde{\theta})) - E_n \exp(-sQ_n(\hat{\theta}))$$

Observe that

$$\tilde{\theta} - \theta = (\tilde{\theta} - \hat{\theta}) + (\hat{\theta} - \theta)$$

and therefore

$$\begin{aligned} \exp(-sQ_n(\tilde{\theta})) = & \exp(-(\tilde{\theta} - \hat{\theta})\prime s B_n B_n\prime(\tilde{\theta} - \hat{\theta})) \\ & \exp(-(\tilde{\theta} - \hat{\theta})\prime 2s B_n B_n\prime(\tilde{\theta} - \hat{\theta})) \\ & \exp(-(\tilde{\theta} - \theta)\prime s B_n B_n\prime(\tilde{\theta} - \theta)) \end{aligned} \tag{8}$$

Now compute

$$E_n(\exp(-sQ_n(\tilde{\theta}))/X_1, \ldots X_n)$$

. The first factor of $\exp(-sQ_n(\tilde{\theta}))$ is $X_1, \ldots X_n$ measurable and therefore can be taken outside of the conditional expectation.

Moreover, the conditional expectation of a function of θ is exactly the expectation with respect to the posterior distribution.

We did assume that the difference of the posterior and normal with expectation $\hat{\theta}$ and variance $(A_n + C)^{-1}$ converges to zero. Since all the function involved are bounded, we can asymptotically replace the posterior with the normal. For easier manipulation of the infinite-dimensional matrices, observe that $(A_n + C)^{-1} = \sqrt{C}^{-1}(\sqrt{C}^{-1} A_n \sqrt{C}^{-1} + I)^{-1}\sqrt{C}^{-1}$.

Then the integral of the wrt a Gaussian can be calculated in closed form, quite analogous to the finite dimensional case.

The integral for the two factors

$$\exp((\widetilde{\theta} - \widehat{\theta})'(sB_nB_n')(A_n + 2sB_nB_n' + C)^{-1}(sB_nB_n')(\widetilde{\theta} - \widehat{\theta})/2)$$

is easily seen to be equal to

$$\sqrt{\det(\sqrt{C}^{-1} A_n \sqrt{C}^{-1} + I)}/ \tag{9}$$

$$\sqrt{\det(\sqrt{C}^{-1} A_n \sqrt{C}^{-1} + \sqrt{C}^{-1} sB_nB_n' \sqrt{C}^{-1} + I)} \tag{10}$$

We did assume that the diagonal elements of C^{-1} decrease rapidly. So even if we would let A_n, B_n be infinite matrices, the corresponding determinants would be well defined, because the operators would be "trace class" (cf. Lang(1993)).

It is easy to see that

$$E_n \exp(-sQ_n(\tilde{\theta}))$$

is asymptotically equal to the product of

$$\exp(-(\tilde{\theta} - \hat{\theta})\prime sB_nB_n\prime(\tilde{\theta} - \hat{\theta}) + \tag{11}$$

$$(\widetilde{\theta} - \widehat{\theta})'(sB_nB_n')(A_n + 2sB_nB_n' + C)^{-1}(sB_nB_n')(\widetilde{\theta} - \widehat{\theta})/2) \tag{12}$$

and

$$\frac{\sqrt{\det(\sqrt{C}^{-1} A_n \sqrt{C}^{-1} + I)}}{\sqrt{\det(\sqrt{C}^{-1} A_n \sqrt{C}^{-1} + I + \sqrt{C}^{-1} B_n \sqrt{C}^{-1})}} \tag{13}$$

The second factor does not depend on $(\tilde{\theta}-\hat{\theta})$ and hence represents $E_n \exp(-sQ_n(\hat{\theta}))$. The first one is equal to

$$\exp(-(\tilde{\theta}-\hat{\theta})\prime H_n(\tilde{\theta}-\hat{\theta})),$$

where

$$H_n = sB_nB_n\prime - (sB_nB_n\prime)(A_n + 2sB_nB_n\prime + C)^{-1}(sB_nB_n\prime)/2,$$

Hence the expectation is smaller than 1 if $\tilde{\theta}-\hat{\theta}$ are different. The first factor depends on B_n, our norming of B_n guarentees that H_n does not vanish asymptotically.

3. Order Estimation

In a previous paper (He and Ploberger (2016)) we discussed the relation of the optimal estimator with sieve estimators based on Akaike's AIC. (A *sieve estimator* simply estimates a number of parameters by e.g. ML and sets the rest of them to 0. In our context here this length is determined by the AIC criterion).

For the special case of a regression model with constant regressors and matrix C being equal to

$$C = diag(Const.\lambda^i), \tag{15}$$

we did show that these two estimators are asymptotically the same. We discussed the precise meaning of this statement. One of the properties we were able to show is that the norm of the difference between sieve estimator and optimal estimator is asymptotically neglible compared to the expected estimation error of the optimal one. It seems natural to try to generalize this result to our case.

There are various ways to do so. The first is "brute force". In the simplest of all cases, where the matrices A_n are asymptotically diagonal, and reasonably regular, like e.g

$$0 < \lambda_1 < \min_{i \leq n^\alpha} (A_n)_{i,i}/n \leq \max_{i \leq n^\alpha} (A_n)_{i.i}/n < \lambda_2, \tag{16}$$

it is easy to calculate

$$(A_n + D_n)^{-1}A_n \tag{17}$$

and show that it is approximately a projection on the first

$$\log n / \log \lambda \tag{18}$$

components.

For general A_n, this approach becomes more complicated. So we propose to make us of one of the most fruitful concepts of modern statistics, namely asymptotic equivalence in a decision theoretic sense. This concept was originally developed by Le Cam. A full exposition can be found in most modern statistical textbooks, so I we will only give a basic description here. A more detailed discussion can be found in the textbooks of Strasser(1995) and Van der Vaardt(2000).

Suppose you have given two sequences of statistical experiments, and we want to make decisions (e.g. tests, estimations,..) based on the experiments. Each decision has attached a loss function (for statistical tests the power function, for estimators some expected distance of the estimated from the true value,..), as well as an ordering of loss functions.(E.g. in the cases of hypotheses testing, one test is better than another if the power function is bigger on the alternative and smaller on the null/) The important assumption we make is that the parameter spaces for the two sequences are the same. Hence we can compare loss functions.

The first sequence of the experiments is said to $\varepsilon-$dominate, $\varepsilon > 0$, the second one, if for large enough n the loss function of any decision from a second experiment one can find a decision from the first one so that the corresponding loss function of the first experiment dominates the other one "up to ε"".

The sequences are equivalent if they $\varepsilon-$dominate each other.

Typical examples are standard experiments with increasing sample sizes on the one hand, and experiments based on a sufficient statistic. One can immediately see that these two sequences of experiments are equivalent (as an example, consider samples of i.i.d normals with identical, but unknown mean and variance on the one hand. The sample mean and variance are known to be sufficient statistics. So a sample of size on a random vector of dimension 2 - the first component distributed as a normal, the second one a χ^2 with the appropriate number of degrees of freedom - will contain the same information.

In the case of finite dimensional models, there is a well developed theory of asymptotic equivalence. In most cases, where the properly normed

second order derivative of the likelihood function converges to a nontrivial constant, the experiments are asymptotically equivalent to an experiment, of sample size one, of a Gaussian random vector with unknown mean and known variance.

The theory in the infinite dimensional case is not that far developed, yet. They are, however, a lot of results available, which show asymptotic equivalence of experiments to a simple Gaussian model. Nonparametric regression models were first considered by Brown and Low(1996) and density estimation in Nussbaum(1996). The general, multivariate regression case was first solved in Reiß(2008). Spectral density estimation was analyzed in Golubev, Nussbaum and Zhou(2010). Autoregressive models and regressiobs were analyzed in Grama and Neumann(2006). More references can be found in the comprehensive textbook of Gine-Nickl(2016).

We do not want to discuss the specific requirements for asympttotic equivalence to a Gaussian problem here. Apart from a central limit theorem for the scores, many assumptions are specific for certain problems.

We think that these references show that in many interesting and practically important cases asymptotic equivalence can be established, and that therefore our result about the equivalence between Bayesian estimator for exponentially declining prior and sieve estimator based on AIC carries over to many more cases. Unfortunately, right now, however, a general theory is not available yet.

4. Concluding Remarks

The theorems here establish some asymptotic optimality properties of a "shrunken" ML-estimator. The restrictions on the parameter are quite severe: We have to assume that, if the parameter is interpreted as a function, this function has to be di□erentiable 8 times. I think that future research will make it possible to relax this rather stringent requirement.

Another promising line of research are empirical Bayesian methods. Theoretically, with an enormous amount of data, one should be able to estimate a large number of coefficients θ_k very precisely. Assuming the c_k e.g. to be of the form $An^{-\gamma}$. theoretically, with an astronomical amount of data, one should be able to estimate A and γ consistently.

Obviously this is not possible for realistic sample sizes. Nevertheless, this raises the questions what kind of inference on the hyperparameters is possible.

5. References

Andrews, D.W.K. (1991): Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation, Econometrica, 59, 817-858.

Bernstein, S.N. (1928): Sur les fonctions absolument monotones, Acta Mathematica 52, 1-66

Armstrong, T.B. and M. Kolesar (2018): Optimal inference in a class of regression models, forthcoming in Econometrica.

Brown, L.D. and M. Low (1996): Asymptotic equivalence of nonparametric regression and white noise, Annals of Statistics 24, p. 2384–98.

Gine, E, and R. Nickl (2016): Mathematical Foundations of Infinite-Dimensional Statistical Models, Cambridge University Press.

Golubev, G.K., M. Nussbaum and H.H. Zhou (2010): Asymptotic equivalence of spectral density estimation and Gaussian white noise. Annals of Statisitics 38,p. 181–214.

Grama, I. and M. Neumann (2006): Asymptotic equivalence of nonparametric autoregression and nonparametric regression, Annals of Statistics 34, p. 1701-1732.

He, J. and W. Ploberger (2016): Prior Free Bayesian forecasting through frequentist's procedures, manuscript, Washington University of St. Louis

Kim, J.Y. (1998): Large Sample Properties of Posterior Densities, Bayesian Information Criterion and the Likelihood Principle in Nonstationary Time Series Models, Econometrica, 66, 359-380

Lang, S.(1993): Real and Functional Analysis, Springer

Nussbaum, M.(1996): Asymptotic equivalence of density estimation and Gaussian white noise. Annals of Statistics 24,p. 2399–430.

Ploberger, W. and Peter C.B. Phillips (2003): An Introduction to Best Empirical Models when the Parameter Space is Infinite Dimensional. Oxford Bulletin of Economics and Statistics 65 (s1), p. 877-890

Ploberger, W and P.C.B. Phillips(2012): Optimal Estimation under Nonstandard Conditions. Journal of Econometrics 169, p. 258-265

Reiß, M. (2008): Asymptotic equivalence for nonparametric regression with multivariate and random design. Annals of Statistics 36, p. 1957–82.

Strasser, H. (1985): *Mathematical Theory of Statistics,* de Gruyter

Van der Vaardt, A.W. (2000): *Asymptotic Statistics*, Cambridge University Press.

6. Appendix: Proof of Theorem

Proof. Let us define $k(n) = n^{1/7}$, and let us for $\theta \in R^N$ define by $\theta^{[k]}$ the vector

$$\theta^{[k]} = (\theta_1, \dots \theta_k, 0, \dots, 0).$$

Let us first observe that by integrating successively, we can conclude from (2) that, perhaps for a different M,

$$E_\theta \left(\sup_{\theta \in O} \left| \frac{\partial \ell_t(\theta)}{\partial \theta_i \partial} \right| \right) \leq M \tag{19}$$

Then we have

$$\sup_{\theta \in O} \left| \sum_{t=1}^{n} \ell_t(\theta) - \sum_{t=1}^{n} \ell_t \left(\theta^{[k(n)]} \right) \right| \leq \sum_{m > k(n), 1 \leq t \leq n} \theta_m \sup_{\theta \in O} \left| \frac{\partial \ell_t(\theta)}{\partial \theta_i} \right| \tag{20}$$

and

$$\begin{aligned} E \sum_{m > k(n), 1 \leq t \leq n} |\theta_m| \sup_{\theta \in O} \left| \frac{\partial \ell_t(\theta)}{\partial \theta_i} \right| &\leq const \sum_{m > k(n), 1 \leq t \leq n} c_m M \\ &\leq const \sum_{m > n^{1/7}, 1 \leq t \leq n} o(1) m^{-8} M = o(1), \end{aligned} \tag{21}$$

Hence the probability measures of θ and $\theta^{[k(n)]}$ are asymptotically equivalent: The difference of the logarithms converges to zero in probability uniformly on O; hence the ratio converges to 1. So let us analyze $\ell_t \left(\theta^{[k(n)]} \right)$ as a function of θ.

Let us first analyze the ML-estimator. Since $\theta^{[k(n)]}$ only contains finitely many parameters, we can use the classical approach for linearization of the first order condition:

$$0 = \sum \ell_t^{(1)}(\theta^{[k(n)]}) + \sum \ell_t^{(2)}(\theta^{[k(n)]})(\hat{\theta} - \theta^{[k(n)]}) + R_n(\hat{\theta} - \theta^{[k(n)]}),$$

where R_n is the remainder term of the Taylor series expansion. With some tedious, but elementary calculations, we can show that (with $||.||$ denoting the usual matrix norm)

$$E||\frac{1}{n} \sum \ell_t^{(2)}(\theta^{[k(n)]}) - \frac{1}{n} E \sum \ell_t^{(2)}(\theta^{[k(n)]})||, \tag{22}$$

$$P[||R_n|| \leq const \cdot k^3(\hat{\theta} - \theta^{[k(n)]})] \to 1,$$

and for all $\epsilon > 0$, we can find a $C(\epsilon)$ so that

$$P[|| \sum \ell_t^{(1)}(\theta^{[k(n)]})|| < \sqrt{k} \cdot \sqrt{n} \cdot C(\epsilon)] > 1 - \epsilon.$$

Since we did assume that the information matrix $\frac{1}{n} E \sum \ell_t^{(2)}(\theta^{[k(n)]})$ is well conditioned, we can conclude that

$$(\sqrt{n}/\sqrt{k})(\hat{\theta} - \theta^{[k(n)]} - \left(\sum \ell_t^{(2)}(\theta^{[k(n)]})\right)^{-1} \sum \ell_t^{(1)}(\theta^{[k(n)]}) \to 0 \tag{23}$$

(where the convergence is to be understood to be in probability) and when utilizing (22)

$$(\sqrt{n}/\sqrt{k})(\hat{\theta} - \theta^{[k(n)]}) \text{ remains } O_P(1) \tag{24}$$

A_n was defined as $\sum \ell_t^{(2)}(\hat{\theta})$. Using a third-order Taylor series expansion, and again using the fact that we assumed the information matrix to be well-conditioned, we may conclude that for all $\eta > 0$

$$P\left[(1-\eta)E \sum \ell_t^{(2)}(\theta^{[k(n)]}) < \sum \ell_t^{(2)}(\hat{\theta}) < (1+\eta)E \sum \ell_t^{(2)}(\theta^{[k(n)]})\right] \to 1 \tag{25}$$

We now have all the tools to compute posterior distribution. The posterior distribution is a random probability measure on Θ, and the density of this measure is proportional to the likelihood function. Let us denote this measure by Π_n. Then Π_n is measurable with respect to $\mathfrak{F}_n$ (the information available at time n). and is a measure defined on the σ-algebra $\mathfrak{I}$ of the measurable subsets of Θ. Let $D_n n$ be events from $\mathfrak{F}_n \times \mathfrak{I}$. Then define the random variables Δ_n by:

Let

$$\Delta_n(\omega) = \pi_n(\{\theta : (\omega, \theta) \in D_n\})$$

It is quite easy to show that Δ_n are random variables: It is trivial if D_n is a product set itself, and then apply a monotone-class argument. Then, trivially

$$P(D_n) = E(\Delta_n).$$

Si uf $P(D_n) \to 1$, $E(\Delta_n) \to 1$, too. As $0 \leq \Delta_n \leq 1, \Delta_n \to 1$ stochastically, too. Hence for all $\epsilon > 0$ we can find $F_n \in \mathfrak{F}_n$ with $P(F_n) > 1 - \epsilon$ so that for $\omega \in F_n$

$$\Delta_n(\omega) = \Pi_n(\{\theta : (\omega, \theta) \in D_n\}) \geq 1 - \epsilon.$$

So if we have given a sequence of events with probability converging to one. Then - automatically - the projections of this set will have - except for events from $\mathfrak{F}_n$ with arbitrary small probabilities - conditional probabilities arbitrarily near to one.

Let us now come back to our original problem, namely analyzing the poste- rior distribution. First of all let us observe that we postulated that our parame- ter can be estimated consistently. So we have estimators -$\mathfrak{F}_n$ DBC0 DC00 valued functions -$\tilde{\theta}_n$ which converge to the true parameter. So $P[\tilde{\theta}_n \in O(\theta)] \to 1$, where O is the bounded neighbourhood we used in (2) and (19). Hence we can conclude that

$$\Pi_n[O(\theta)] \to 1 \tag{26}$$

. in probability.

In principle, $\Pi_n(.)$ should be easy to construct. We know that the density is proportional to the likelihood, and we did derive some simplifying approximations to the likelihood. Our first problem is the normalizing factor. We would have to integrate the likelihood over the whole parameter space, which is inconvenient. When we use relations like (26), we can limit our averaging to e.g. $O(\theta)$. For all $\epsilon > 0$, for n large enough there exist sets $F_n \in \mathfrak{F}_n$ with $P(F_n) \geq 1 - \epsilon$ and for $\omega \in F_n$, $\Pi_n[O(\theta)](\omega) \geq 1 - \epsilon$. So with Π denoting the prior on Θ we have

$$\frac{d\Pi_n}{d\Pi} = \frac{\exp(\sum \ell_t(\theta))}{\int \exp(\sum \ell_t(\theta)) d\Pi(\theta)}.$$

For $\omega \in F_n$, however, $\Pi_n[O(\theta)](\omega) \geq 1 - \epsilon$. Hence

$$\int_{O(\theta)} \frac{d\Pi_n}{d\Pi} d\Pi \geq 1 - \epsilon,$$

and therefore

$$\frac{\int_{O(\theta)} \exp(\sum \ell_t(\theta)) d\Pi}{\int \exp(\sum \ell_t(\theta)) d\Pi} \geq 1 - \epsilon.$$

Since the ration on the LHS is bounded by 1, we may conclude that

$$\lim \frac{\int_{O(\theta)} \exp(\sum \ell_t(\theta)) d\Pi}{\int \exp(\sum \ell_t(\theta)) d\Pi} = 1$$

which implies that

$$\lim \frac{\int_{O(\theta)^c} \exp(\sum \ell_t(\theta)) d\Pi}{\int \exp(\sum \ell_t(\theta)) d\Pi} = 0$$

Therefore we can conclude that total variation of the difference between Π_n and the random measure $\Pi_n^{(1)}$ defined by its with density

$$\frac{I_{O(\theta)} \exp(\sum \ell_t(\theta)) d\Pi}{\int I_{O(\theta)} \exp(\sum \ell_t(\theta)) d\Pi} n$$

converges to zero.

Let us now define the random measure $\Pi_n^{(2)}$ to have the density

$$\frac{I_{O(\theta)} \exp(\sum \ell_t(\theta^{[k]})) d\Pi}{\int I_{O(\theta)} \exp(\sum \ell_t(\theta^{[k]})) d\Pi} n$$

Then (20),(21) imply that the total variation of the difference converges to zero.

Now lets construc events of probability converging to one based on equation (23),(24) and (25). What we want to do is to approximate

$$\sum \ell_t(\theta^{[k]})$$

by its second order Taylor approximation around $\hat{\theta}_n$. First observe that in (24) implies that

$$P\left([\hat{\theta}_n \in O(\theta)]\right) \to 1.$$

We can apply our technique again to this sequence of events and construct measure $\Pi_n^{(3)}$ with the corresponding densities: this way we guarentee that $\hat{\theta}_n$ is in $O(\theta)$, so all our functions are differentiable. Next we analyze (10):

$$\left(\sqrt{n}/\sqrt{k}\right) (\hat{\theta} - \theta^{[k(n)]}) \text{ remains } O_p(1).$$

An equivalent formulation of this statement is: for any sequence $B_n \uparrow \infty$ define the events

$$S_n = \left[|| \left(\sqrt{n}/\sqrt{k}\right) (\hat{\theta} - \theta^{[k(n)]}) || \leq B_n \right]$$

Then $P(S_n) = 1$.

Then we have, again, we can construct measures $\Pi_n^{(4)}$ for which our density equals

$$\frac{I_S \exp(\sum \ell_t(\theta^{[k]}))d\Pi}{\int I_S \exp(\sum \ell_t(\theta^{[k]}))d\Pi} n.$$

For this density, however, we can use a Taylor series expansion with $\hat{\theta}$ as base value

$$\sum \ell_t(\theta^{[k]}) = \sum \ell_t(\hat{\theta} - (\theta^{[k]} - \hat{\theta}))\prime \left(\sum \ell_t^{(2)}(\hat{\theta})\right) (\theta^{[k]} - \hat{\theta}))/2 + r_n$$

where

$$r_n \leq \sum \sup_{\theta \in O} |\frac{\partial^3 \ell_t(\theta)}{\partial \theta_i \theta_j \theta_k}| \max |(\theta^{[k]} - \hat{\theta}))_i| \max |(\theta^{[k]} - \hat{\theta}))_j| \max |(\theta^{[k]} - \hat{\theta}))_k|$$

. Then for $\theta^{[k]} \in S_n$,

$$E|r_n| \leq \left(E \sup_{\theta \in O} |\frac{\partial^3 \ell_t(\theta)}{\partial \theta_i \theta_j \theta_k}|\right) nB^3 \frac{k^{3/2}}{n^{3/2}} = B_n^3 \frac{n^{3/(2\times 7)}}{n^{1/2}} = B_n^3 n^{-4/14}.$$

So choosing

$$B_n = o(n^{1/14})$$

guarentees that $E|r_n| \to 0$.

Hence we can again construct $\Pi_n^{(5)}$, being asymptotically equivalent to Π_n, with density

$$\frac{I_{S_n} \exp(\sum \ell_t(\hat{\theta}) + (\theta^{[k]} - \hat{\theta})\prime P_k\prime \left(\sum \ell_t^{(2)}(\hat{\theta})\right) P_k(\theta^{[k]} - \hat{\theta})/2)}{\int I_{S_n} \exp(\sum \ell_t(\hat{\theta}) + (\theta^{[k]} - \hat{\theta})\prime P_k\prime \left(\sum \ell_t^{(2)}(\hat{\theta})\right) P_k(\theta^{[k]} - \hat{\theta})/2) d\Pi}$$

$\sum \ell_t(\hat{\theta})$ does not depend on θ so the corresponding term $\exp(\sum \ell_t(\hat{\theta}))$ cancels out. Furthermore, observe that

$$\theta^{[k]} = P_k \theta.$$

where P_k is the matrix describing projection to the first k components of a vector. As $\hat{\theta}$ only contains k components, we have

$$\hat{\theta} = P_k \hat{\theta}$$

. Hence we can write our density as

$$\frac{I_{S_n} \exp((\theta-\hat{\theta})\prime P_k\prime \left(\sum \ell_t^{(2)}(\hat{\theta})\right) P_k(\theta-\hat{\theta})/2)}{\int I_{S_n} \exp((\theta-\hat{\theta})\prime P_k\prime \left(\sum \ell_t^{(2)}(\hat{\theta})\right) P_k(\theta-\hat{\theta})/2) d\Pi}$$

which looks very much like a Gaussian density. The only problem is the factor I_{S_n}. We know that this factor converges to 1 so it is sufficient to establish that the measure with density

$$\frac{\exp((\theta-\hat{\theta})\prime P_k\prime \left(\sum \ell_t^{(2)}(\hat{\theta})\right) P_k(\theta-\hat{\theta})/2)}{\int \exp((\theta-\hat{\theta})\prime P_k\prime \left(\sum \ell_t^{(2)}(\hat{\theta})\right) P_k(\theta-\hat{\theta})/2) d\Pi}$$

is

$$G\left(\left(C-\sum \ell_t^{(2)}(\hat{\theta})\right)^{-1}\left(\sum \ell_t^{(2)}(\hat{\theta})\right)\hat{\theta}, \left(C-\sum \ell_t^{(2)}(\hat{\theta})\right)^{-1}\right)$$

and

$$G\left(\left(C-\sum \ell_t^{(2)}(\hat{\theta})\right)^{-1}\left(\sum \ell_t^{(2)}(\hat{\theta})\right)\hat{\theta}, \left(C-\sum \ell_t^{(2)}(\hat{\theta})\right)^{-1}\right)(S_n^C) \to 0.$$

Where S_n^C is the complememt of S_n. Both are tedious but elementary calculations with normal random variables. so we will omit these proofs. ■