Mathematisches Forschungsinstitut Oberwolfach

Report No. 47/2005

Statistische und Probabilistische Methoden der Modellwahl

Organised by
James O. Berger (Durham)
Holger Dette (Bochum)
Gabor Lugosi (Barcelona)
Axel Munk (Göttingen)

October 16th – October 22nd, 2005

Abstract. Aim of this conference with more than 50 participants, was to bring together leading researchers from roughly three different scientific communities who work on the same issue, *data based model selection*. Their different methodological approaches can be roughly classified into

(1) Frequentist model selection and testing
(2) Statistical learning theory and machine learning
(3) Bayesian model selection

The key task in model selection is to select a *proper* mathematical model based on information generated by data and/or by prior knowledge. Proper might mean a model with minimal prediction error, a model which describes the main qualitative data features, such as bumps and modes, or a model of low computational complexity. Mathematical techniques and concepts encountered with this workshop are wide spread, ranging from concentration and oracle inequalities, asymptotic analysis and distribution theory to testing theory, information measures and nonconvex optimization.

Mathematics Subject Classification (2000): 62G05, 62G10, 62G20, 62F15, 62H30, 62G99, 62C12, 60F15, 62N03, 62B10, 62F40.

Introduction by the Organisers

In order to achieve our goal to enhance discussion between these communities, every day the conference was opened by a survey talk. Friday afternoon the conference has been closed by a discussion session.

1. Frequentist model selection and testing

Nils Hjort introduced in his talk the fundamental concept of a focused information criterion for model selction, which does not propagate a model per se, rather it reflects the more realistic situation, that specific aspects of a model should drive the model selction process. He adressed various questions related to this, e.g. robustness issues, or how do classical information criteria such as AIC or BIC behave from this perspectice. He gives strong evidenve by various examples that different models may result when focussing on different parameters of primary interest.

The issue of testing a model was adressed by various talks, N. Neumeyer used bootstrap techniques applied to residual processes whereas L. Györfy's criterion is based on the L^1 distance between densities. J. M. Loubes and N. Bissantz were concerned with model selection in inverse problems, i.e. for noisy integral operator equations. J.M. Loubes considers nonlinear operators which are locally linear and investigates convergence rates of penalized M-estimators. N. Bissantz focuses on L^2 distance based model testing and selection methods and discusses various applications in astrophysics. To this end, a general analyisis of numerical and statistical regularisation methods is given. Finally, he constructed uniform confidence bands in deconvolution problems which allow graphically to select a proper model. The problem of deconvolution was also highlighted by J. P Kreiss in the context of time series analysis. Conceptually related to N. Hjorts talk, J.K. Ghosh discussed different roles of different penalties in penalized likelihood model selection rules, making the case that the penalty used should depend on the goal (typically either prediction or selection of the best model) and that it is important to incorporate practical features such as growing model dimension in choosing penalties. L. Dümbgen was concerned with prediction regions in gaussian shift models. He suggested a solution but also pointed out that adaptive construction of prediction regions via a sequence of nested models is limited in various ways. This is in contrast to adaptive estimation. He discussed a 'no go' result on the asymptotic diameter of the confidence ball in the spirit of Li (1989).
Other talks included topics on *Empirical process techniques for locally stationary processes* by Rainer Dahlhaus and *Universal principles, approximation and model choice* by Patrick Laurie Davies and *Local Parametric Methods in Nonparametric Regression* by Vladimir Spokoiny.

2. Statistical learning theory and machine learning

Research on statistical learning theory and nonparametric classification has also been strongly represented by several attendants who partly or completely focus their research on these topics. Several talks have been given in these fields, offering a nice overview on some of the most active areas of investigation, such as oracle inequalities for penalized model selection, margin-based performance bounds, empirically calibrated penalties, model selection focusing on sparse solutions of corresponding optimization problems, convex aggregation of estimators, as well as

some closely related issues emerging in density estimation, microarray analysis, etc.

Peter Bartlett (UC Berkeley) gave a survey talk on nonparametric classification based on empirical minimization of convex cost functionals, a subject that offers a theoretical framework for many successful classification algorithms, including boosting and support vector machines. Marten Wegkamp's talk (Florida State University) discussed a closely related problem of classification with a reject option. Another survey talk on a closely related subject was delivered by Sara van de Geer (ETH Zürich) who showed why empirical process theory and concentration inequalities play a crucial role in model selection problems for classification and nonparametric regression. Similarly to Prof. van de Geer, Vladimir Koltchinskii (Georgia Tech) also considered L1-type penalties that lead to sparse models and derived sharp oracle inequalities.

Both Alexandre Tsybakov (University of Paris 7) and Florentina Bunea (Florida State University) considered methods for convex aggregation of certain estimates for regression, and proved close-to-optimal performance bounds. László Györfi (Technical University of Budapest) presented a model selection method and a corresponding L1 performance bound for density estimation when the unknown density is assumed to be in one of an infinite sequence of "parametric" classes of densities.

Andrew Nobel (University of North Carolina) discussed algorithmic and probabilistic problems arising in some problems of data mining that can be modeled as searching for large homogeneous blocks in random matrices.

3. Bayesian model selection

In *Bayesian model selection and BART*, E. George and R. McCulloch gave a survey of the Bayesian approach to model selection, while giving an illustration (BART) that seems to have remarkable predictive properties in function estimation and variable selection. This was followed by Merlise Clyde, giving a talk on *Bayesian nonparametric function estimation using overcomplete representations and Lévy random field priors.* This focused on the novel notion in Bayesian analysis that simultaneous use of multiple bases for functions (leasing to overcompleteness) can be quite valuable in practice, because it can allow for extremely sparse representations of functions. The final Bayesian talk on Monday was by Christian Robert, on *Prior choice and model selection.* This highlighted the key issue faced by Bayesians in model choice, namely the choice of the prior distribution. Modern approaches to this issue were reviewed, and a new approach (based on a criterion of 'matching' between models) was introduced.

Later talks included *A synthesis and unification of Bayes factors for model selection and hypothesis testing*, by Luis Pericchi. This talk discussed the prominent role of training samples (or bootstrapping), in many modern model selection scenarios. Valen Johnson, in *A note on the consistency and interpretation of Bayes factors based on test statistics* considered the problem of developing easy to use

Bayesian procedures as replacements for standard statistical procedures, such as chi-squared tests, t-tests, etc. He demonstrated how many Bayesian testing problems can be reduced to situations with only a one-dimensional unknown, which lend themselves to graphical description.

On the final day, the issue of multiple testing was addressed. This is one of the currently hottest areas of statistical and scientific research, and two talks were presented. M.-J. Bayarri gave a survey talk entitled *Multiple testing: the problem and some solutions*, which reviewed the connections between 'false discovery rate,' Bayesian posterior probabilities, and utility functions common in multiple testing scenarios. P. Müller followed with elaborations on the utility side, involving applications to significant problems in bioinformatics and clinical trials.

The final session in the workshop consisted of very short talks to give other participants (especially newer researchers) a chance to discuss their interests, and several Bayesian talks were presented. M. Bogdan presented *Model selection approach to the problem of locating genes influencing quantitative traits*, presenting a very nice generalization of BIC for a genetics problem. Katja Ickstadt presented *Comparing classification procedures using misclassification rates*, with an interesting application to determining genetic 'snips.' Angelika van der Linde spoke on *Posterior predictive model choice*, discussing a new asymptotic Bayesian approach to model choice, requiring a careful decomposition of entropy.

Closing Discussion Session: The workshop ended with a discussion session designed to identify key problems remaining to be addressed, and to identify key ways to bridge the gaps between the communities present at the workshop. The questions – together with short descriptions of the results of the discussion – are below.

- Do we all mean the same thing by the phrase model selection? Is it selection of a statistical model for the data, selection of a prediction function, or some averaged version of either?
  - *Conclusion:* If prediction is the identified goal, then the various communities have the same view of model selection. Otherwise, interesting differences exist.
- Are fundamental problems of statistics and machine learning different? If they are the same, why are the commonly used techniques so different?
  - *Conclusion:* Machine learning is concerned primarily with action and associated risk, and is less focused on inference, which is often viewed as the primarily goal of statistics.
- Discuss the parametric aspects of nonparametric models.
  - *Conclusion:* Any nonparametric procedure is only good in certain finite dimensional regions of the nonparametric space.

- Is model selection fundamentally different when the true model is outside the class of models being considered?
  - *Conclusion:* This is primarily an issue in Bayesian statistics, because the other viewpoints formulate the model class so that it is supposedly assured to contain the true model; there was, however dissension as to whether the latter was actually possible.
- How does information theory contribute to statistics?
  - *Conclusion:* Notions such as 'minimum description length' are difficult to encode, and are arguably as difficult to implement as the more usual model/prior paradigm.
- Given that regularization is very related to Bayesian analysis,
  - Do oracle or risk inequalities tell us about performance of Bayesian procedures? In practice? For (growing) finite sample size? Asymptotically?
  - Can regularization results help Bayesians in choosing priors? Do oracle based convergence rates relate to optimal objective priors?
  - How do oracle inequalities relate to AIC, BIC, ...?
  - *Conclusion:* AIC and BIC are not derivable as oracle inequalities. Indeed, only if the constants in oracle inequalities are essentially one (i.e., the inequalities are exact in some regions), can there be a hope that oracle inequalities and Bayesian analysis will coincide. The other questions are fundamentally unknown issues for future study.

Workshop: Statistische und Probabilistische Methoden der Modellwahl

Table of Contents

Abstracts

Regression methods for pattern classification: Statistical properties of large margin classifiers

Peter L. Bartlett

(joint work with Ambuj Tewari, Michael Jordan, and Jon McAuliffe)

In the pattern classification problem, we have independent and identically distributed random pairs $(X,Y),(X_1,Y_1),\ldots,(X_n,Y_n)$ with distribution P on $\mathcal{X}\times\mathcal{Y}$, where the label space $\mathcal{Y}$ is finite. The aim is to use the data $(X_1,Y_1),\ldots,(X_n,Y_n)$ to choose a function $f_n : \mathcal{X} \to \mathcal{Y}$ with small risk, $R(f) = \Pr(f(X) \neq Y) = \mathbb{E}\ell(Y,f(X))$, where we have defined the 0-1 loss ℓ in the obvious way. In a wide variety of cases, the natural approach of minimizing the empirical risk, $\hat{R}(f) = \hat{\mathbb{E}}_n\ell(Y,f(X)) = n^{-1}\sum_{i=1}^n \ell(Y_i,f(X_i))$, is computationally intractable (see (8; 1) for a review). For this reason, many of the pattern classification algorithms developed in the machine learning literature, including the support vector machine (6) and AdaBoost (9), replace the 0-1 loss ℓ with a convex surrogate ϕ, and minimize the sample average of this surrogate loss function. Thus, these methods can be viewed as minimum contrast methods. The convexity makes these algorithms computationally efficient. The use of a surrogate, however, has statistical consequences that must be balanced against the computational virtues of convexity. This talk surveyed some recent results in this area.

Most of the talk focused on two-class classification, where $\mathcal{Y} = \{\pm 1\}$ and $f_n : \mathcal{X} \to \mathbb{R}$. In this case, we can write the risk as $R(f_n) = \Pr(\mathrm{sign}(f_n(X)) \neq Y) = \mathbb{E}\ell(Y,f_n(X))$, where we have redefined the 0-1 loss in the obvious way. To replace ℓ with a convex surrogate, we define the ϕ-risk of a function $f : \mathcal{X} \to \mathbb{R}$ as $R_\phi(f) = \mathbb{E}\phi(Yf(X))$. The methods that we study choose f_n from some sequence of classes $\mathcal{F}_n$, so as to minimize the empirical ϕ-risk, $\hat{R}_\phi(f) = \hat{\mathbb{E}}_n\phi(Yf(X)) = n^{-1}\sum_{i=1}^n \phi(Y_if(X_i))$, or a regularized version. For example, AdaBoost chooses f_n from $\mathrm{span}(\mathcal{G})$, for a VC-class $\mathcal{G}$, to minimize $\hat{R}_\phi(f)$ using greedy basis selection, with $\phi(\alpha) = \exp(-\alpha)$. Support vector machines (SVMs) choose f_n from a ball in a reproducing kernel Hilbert space $\mathcal{H}$ to minimize $\hat{R}_\phi(f) + \lambda_n\|f\|_{\mathcal{H}}^2$, where $\phi(\alpha) = \max(0,1-\alpha)$ and $\|\cdot\|_{\mathcal{H}}$ is the norm in the RKHS $\mathcal{H}$.

There has been a considerable body of work on the statistical consequences of using a convex surrogate in place of the 0-1 loss. For instance, it is known (19; 22; 15; 10) that AdaBoost (suitably regularized) and SVMs are universally consistent methods, that is, for suitable choice of the regularization schedule, the risk of f_n converges in probability to the Bayes risk, $R^* = \inf_f R(f)$, where the infimum is over all measurable functions.

There is a simple characterization of surrogate loss functions that lead to a universally consistent method, that is, for which minimization of the risk, as assessed using ϕ, leads to minimal risk, as assessed using the 0-1 loss. To state the result, we need some more definitions. The first two are based on two simple

observations. First, the Bayes risk R^* is obtained by $f^*(x) = \text{sign}(2\eta(x) - 1)$, where $\eta(x) = \Pr(Y = 1|X = x)$. Second, we can write the ϕ-risk as $R_\phi(f) = \mathbb{E}\left(\eta(X)\phi(f(X)) + (1 - \eta(X))\phi(-f(X))\right)$. Define the optimal conditional ϕ-risk for a conditional probability $\eta \in [0, 1]$ as $H(\eta) = \inf_{\alpha\in\mathbb{R}}\left(\eta\phi(\alpha) + (1 - \eta)\phi(-\alpha)\right)$, and the corresponding quantity when the argument α is restricted to have a sign that disagrees with f^*, $H^-(\eta) = \inf_{\alpha:\alpha(2\eta-1)\le 0}\left(\eta\phi(\alpha) + (1 - \eta)\phi(-\alpha)\right)$. Define the optimal ϕ-risk as $R^*_\phi = \inf_f R_\phi(f)$, where the infimum is over all measurable functions. Finally, for convex ϕ, define $\psi(\theta) = \phi(0) - H\left((1 + \theta)/2\right)$. The following result (2) gives a characterization of the loss functions that lead to a universally consistent method, and it also shows how the excess risk is related to the excess ϕ-risk. (There is a straightforward generalization to non-convex ϕ.)

Theorem 1. *Consider a convex $\phi : \mathbb{R} \to \mathbb{R}_+$.*
a. *For any probability distribution P and any f,*

$$\psi(R(f) - R^*) \le R_\phi(f) - R^*_\phi.$$

b. *For $|\mathcal{X}| \ge 2$, $\epsilon > 0$ and $\theta \in [0, 1]$, there is a distribution P and a function f with $R(f) - R^* = \theta$ and $\psi(\theta) \le R_\phi(f) - R^*_\phi \le \psi(\theta) + \epsilon$.*
c. *The following conditions are equivalent:*

(1) *ϕ satisfies $H(\eta) < H^-(\eta)$ for all $\eta \ne 1/2$.*
(2) *ϕ is differentiable at zero and its derivative is negative at zero.*
(3) *$\psi(\theta_i) \to 0$ iff $\theta_i \to 0$.*
(4) *$R_\phi(f_i) \to R^*_\phi$ implies $R(f_i) \to R^*$.*

Consider a method of sieves approach, in which we choose $f_n \in \mathcal{F}_n$ to minimize the empirical ϕ-risk. As an aside, note that a regularization approach, in which we choose $f_n \in \mathcal{F}$ to minimize a regularized empirical ϕ-risk functional $\hat{R}_\phi(f) + \lambda_n\Omega(f)$, for some regularization functional Ω, can be viewed as a method of sieves, with $\mathcal{F}_n = \{f \in \mathcal{F} : \lambda_n\Omega(f) \le B_n\}$, where B_n satisfies $\hat{R}_\phi(0) + \lambda_n\Omega(0) \le B_n$ and 0 denotes the constant zero function. We can decompose the excess risk estimate as

$$\begin{aligned} R(f_n) - R^* &\le \psi^{-1}\left(R_\phi(f_n) - R^*_\phi\right) \\ &= \psi^{-1}\left(R_\phi(f_n) - \inf_{f\in\mathcal{F}_n} R_\phi(f) + \inf_{f\in\mathcal{F}_n} R_\phi(f) - R^*_\phi\right). \end{aligned}$$

The first term inside the functional inverse of ψ is the estimation error, and the second term is the approximation error. Notice that this decomposition is in terms of the ϕ-risk, rather than the risk. If the class is suitably rich (so that $\inf_{f\in\mathcal{F}} R_\phi(f) = R^*_\phi$), and the classes $\mathcal{F}_n$ get large suitably slowly, $R_\phi(f_n)$ converges to R^*_ϕ in probability. Then universal consistency (convergence of $R(f_n)$ to R^*) follows iff ϕ satisfies the derivative condition of the theorem above.

The question of rates of convergence is also important. Consider the following complexity penalized approach. Define $\hat{f}_k = \arg\min_{f\in\mathcal{F}_k} \hat{R}_\phi(f)$ and $f_n = \hat{f}_{\hat{k}}$ with $\hat{k} = \arg\min_k \left(\hat{R}_\phi(\hat{f}_k) + p_k\right)$, for some penalty p_k (that depends on n).

We are interested in *oracle inequalities* of the form

$$R_\phi(f_n) - R_\phi^* \leq \inf_k \left(\inf_{f \in \mathcal{F}_k} R_\phi(f) - R_\phi^* + cp_k \right).$$

In such an inequality, if the cp_k is of the same order of magnitude as the estimation error $R_\phi(\hat{f}_k) - \inf_{f \in \mathcal{F}_k} R_\phi(f)$, then the inequality shows that our choice f_n has excess risk not much worse than would be obtained if we had the advice of an oracle who tells us the correct complexity class $\mathcal{F}_k$ to choose. Such inequalities follow easily from uniform convergence results. For example, it is straightforward to show that $\sup_k \left(\sup_{f \in \mathcal{F}_k} \left| R_\phi(f) - \hat{R}_\phi(f) \right| - p_k \right) \leq 0$ implies $R_\phi(f_n) \leq \inf_k \inf_{f \in \mathcal{F}_k} \left(R_\phi(f) + 2p_k \right)$. So it suffices to choose the penalty p_k as a high-probability upper bound on the maximal deviation between empirical ϕ-risks and ϕ-risks.

It is sometimes possible to obtain faster rates of convergence (smaller values of p_k as a function of n) as follows.

Theorem 2. *Suppose $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \mathcal{F}_3 \subseteq \cdots$. If*

$$\sup_k \sup_{f \in \mathcal{F}_k} \left(R_\phi(f) - R_\phi(f_k^*) - 2 \left(\hat{R}_\phi(f) - \hat{R}_\phi(f_k^*) \right) - \epsilon_k \right) \leq 0,$$

$$\sup_k \sup_{f \in \mathcal{F}_k} \left(\hat{R}_\phi(f) - \hat{R}_\phi(f_k^*) - 2 \left(R_\phi(f) - R_\phi(f_k^*) \right) - \epsilon_k \right) \leq 0,$$

then with $p_k = 7\epsilon_k/2$, we have

$$R_\phi(f_n) \leq \inf_k \left(R_\phi(f_k^*) + 9\epsilon_k \right).$$

This approach is useful, for example, if the loss function is strictly convex and uniformly Lipschitz, or, in classification, if the conditional probability is unlikely to be close to 1/2. See, for example, (13; 17; 16; 21; 18; 12).

Logistic regression can be viewed as a method that minimizes the sample average of a convex loss function ϕ. It can be shown that, for any differentiable loss ϕ, minimization of empirical ϕ-risk corresponds to estimation of the conditional probability of Y given X (22). Further, the points of non-differentiability correspond to subsets of the interval $[0, 1]$ where the value of the conditional probability cannot be estimated asymptotically (3).

Finally, the talk discussed the multiclass problem, where $\mathcal{Y}$ is a set of cardinality greater than two. For a family of related methods, which choose a vector-valued function to minimize a convex criterion Ψ, there is a characterization of the universal consistency property in terms of geometric properties of the function Ψ (20). It follows from these results that many loss function that have been proposed in the literature cannot lead to universally consistent methods.

References

[1] Martin Anthony and Peter L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press (1999).

[2] Peter L. Bartlett, Michael I. Jordan and Jon D. McAuliffe, *Convexity, classification, and risk bounds*, Journal of the American Statistical Association (2005), to appear. (Was Department of Statistics, U.C. Berkeley Technical Report number 638, 2003).

[3] Peter L. Bartlett and Ambuj Tewari, *Sparseness vs estimating conditional probabilities: Some asymptotic results*, Proceedings of the 17th Annual Conference on Learning Theory, Springer, volume **3120** (2004), 564–578.

[4] P. J. Bickel and Y. Ritov, *The golden chain*, Annals of Statistics (2004) (to appear). http://pluto-mscc.huji.ac.il/∼yaacov/adaGoldenChain.pdf.

[5] G. Blanchard, G. Lugosi and N. Vayatis, *On the rate of convergence of regularized boosting classifiers*, Journal of Machine Learning Research **4** (2003), 861–894.

[6] B. E. Boser, I. Guyon and V. Vapnik, *A training algorithm for optimal margin classifiers*, ACM Workshop on Computational Learning Theory (1992), 144–152.

[7] Peter Buhlmann and Bin Yu, *Boosting with the L2 loss: Regression and classification*, Journal of the American Statistical Association **98** (2003), 324–339. http://www.stat.berkeley.edu/∼binyu/ps/boostingl2.ps.

[8] Luc Devroye, Lázsló Györfi and Gábor Lugosi, *A probabilistic theory of pattern recognition*, Applications of Mathematics: Stochastic Modelling and Applied Probability, Springer, **31** (1996).

[9] Yoav Freund and Robert E. Schapire, *A decision-theoretic generalization of on-line learning and an application to boosting*, Journal of Computer and System Sciences **55(1)** (1997), 119–139.

[10] Wenxin Jiang,*Process consistency for AdaBoost*, Annals of Statistics **32** (2004) (to appear).
http://neuman.stats.nwu.edu/jiang/boost/boost.process.ps.

[11] V. I. Koltchinskii and D. Panchenko, *Rademacher processes and bounding the risk of function learning*, In Evarist Giné, David M. Mason, and Jon A. Wellner, editors, High Dimensional Probability II, Birkhäuser, volume **47** (2000), 443–459.

[12] Vladimir Koltchinskii, *Local Rademacher complexities and oracle inequalities in risk minimization*, Technical report (2003).

[13] W. S. Lee, P. L. Bartlett and R. C. Williamson, *Efficient agnostic learning of neural networks with bounded fan-in*, IEEE Transactions on Information Theory **42(6)** (1996), 2118–2132.

[14] Y. Lin, *A note on margin-based loss functions in classification*, Statistics and Probability Letters (2004), to appear.

[15] G. Lugosi and N. Vayatis, *On the Bayes-risk consistency of regularized boosting methods*, Annals of Statistics **32** (2004), to appear.

[16] Enno Mammen and Alexandre B. Tsybakov, *Smooth discrimination analysis*, Annals of Statistics **27(6)** (1999), 1808–1829.

[17] Shahar Mendelson, *Improving the sample complexity using global data*, IEEE Transactions on Information Theory **48(7)** (2002), 1977–1991.

[18] E. Nedelec and P. Massart, *Risk bounds for statistical learning*, unpublished (2003).
[19] Ingo Steinwart, *Support vector machines are universally consistent*, Journal of Complexity **18** (2002), 768–791.
[20] Ambuj Tewari and Peter L. Bartlett, *On the consistency of multiclass classification methods*, Proceedings of the 18th Annual Conference on Learning Theory, Springer, volume **3559** (2005), 143–157.
[21] A. B. Tsybakov, *Optimal aggregation of classifiers in statistical learning*, Annals of Statistics (2004), to appear.
[22] T. Zhang, *Statistical behavior and consistency of classification methods based on convex risk minimization*, Annals of Statistics **32** (2004), to appear.

Multiple testing: the problem and some solutions

María-Jesús Bayarri
(joint work with James O. Berger)

An increasingly common situation in practice is simultaneous screening of many (hundreds or thousands) of hypotheses to determine whether we have 'noise' or 'signals'. A typical example is in gene expression (microarrays), when many genes are tested for differential expression among different treatments. Other examples occur in 'Anomaly Discovery'; for example, in 'Syndromic Surveillance' many counties perform daily tests on the 'excess' of some symptoms, the goal being early detection of the outbreak of epidemics or of bio-terrorist attacks (Stoto et al., 2004).

Assume that, based on the observed value of $\boldsymbol{X} = (X_1, X_2, \ldots, X_M)$ (M usually very large) we wish to perform M tests of hypotheses $H_{0i} : X_i \sim f_{0i}$ versus $H_{1i} : X_i \sim f_{1i}$. f_{1i} and f_{0i} usually involve unknown parameters.

If the M tests are independent and each is tested at level α, then, even when all the M nulls are true, we expect $\alpha\, M$ rejections; In simultaneous testing, this is perceived as 'too many', unduly masking detection of incorrect null. Alternatively, this is stated as the problem of 'multiplicity in testing:' as the number of simultaneous tests being conducted increases, the criterion for rejection must become more strict.

Let $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_M)$, with $\gamma_i = 0$ if the ith null is true, and $\gamma_i = 1$ if the ith alternative is true. The multiple testing problem can thus be formulated as a model selection problem: to choose among the 2^M models indexed by the possible values of the vector $\boldsymbol{\gamma}$.

In these massive screenings, there tends to be strong prior probability that there are few signals, that is, that many of the γ_i's are 0. Bayesian as well as recent frequentist analyses take this into account. For summary and references on frequentist estimates of the proportion of true nulls, see Langaas et al. (2005). Note that in Bayesian analysis, estimation of this proportion, if of interest, is a byproduct of the overall analysis.

From the frequentist point of view, model selection techniques are rarely used; instead, generalizations of hypothesis testing ideas to the overall scenario involving the M tests are used. In particular, 'global' error rates are defined, and procedures that control or evaluate these error rates are developed.

Traditional error rates to control in multiple testing have been the per-comparison error rate, controlled by testing each hypothesis at level α, and family-wise error rate, the probability of at least one false 'discovery' (rejection), controlled by, among others, the popular Bonferroni method. While the former does not take multiplicity into account, the latter is very conservative, resulting in tests with very little power. For reviews and references, see Shaffer(1995); Dudoit et al.(2003); Yang & Rempala(2004).

A new type of error rate is being increasingly used lately and has become enormously popular. This is the 'false discovery rate' (FDR), first introduced by Benjamini and Hochberg (1995). They argued that the interesting quantity in multiple testing is the is the % of false discoveries (erroneous rejections) *among the rejected hypotheses*. This gave rise to several related FDR error types to control (or monitor), and the literature in the area has grown huge. Some procedures fix the error rate and select a procedure that guarantees that the error rate is bellow that fixed value for all combination of nulls and alternatives (Benjamini and Hochberg, 1995, 2000; Black, 2004); other procedures are based in errors that can not be 'controlled' in this classical sense, and work instead by choosing among all fixed rejection regions, seeking to 'control' an estimate of the error rate (Storey 2002, 2003); a partial comparison of procedures and many asymptotic properties can be found in Genovese and Wasserman (2002, 2003). In this scenario, a mixture of frequentist and empirical Bayes analyses can be found in Efron et al. (2001) and Efron and Tibshirani (2002). Empirical Bayes and full Bayes analyses can be found in Newton et al. (2001, 2004); Newton and Kendziorski (2003); Gönen et al. (2003); Müller et al. (2004); Do et al. (2005); House et al. (2006); Scott and Berger (2006).

A few important remarks:

- The (unknown) proportion p_0 of true nulls is a crucial ingredient in both Bayesian and FDR analyses: in Bayesian analyses to provide the solution to the multiple comparisons problem; in FDR analyses to increase 'power.' Modern analyses incorporate a real or conservative estimate of p_0; For Bayesian analyses, it is a required ingredient; for frequentist analyses, often a bound (instead of an estimate) is used, which might not be optimal.
- Storey shows that his 'positive FDR' can be shown to be equal to the probability of each null being true conditional on the observation being in the rejection region. Because of this conditioning on a subset of the sample space, he (and others) call this a 'Bayesian FDR'. Notice, however, that this is far from being a Bayesian quantity, since it is not conditional in the observed data.

- Because Bayesian decisions rules usually adopt the form of individual cut-of points for the posterior probabilities of the hypotheses, Bayesian solutions have been wrongly accused of 'ignoring the multiplicity issue' (since the cut-of points do not explicitly depend on the number of hypotheses being simultaneously tested). However, this is a wrong statement: the Bayesian machinery implicitly controls for multiplicity; no external adjustments are needed. This is nicely illustrated in Scott and Berger (2006).

In spite of its popularity, some authors (Finner and Roters, 2001) defend control of properly adjusted Type I errors, since FDR can be used to 'cheat' by simply adding spurious comparisons which are very likely to be rejected. On the other hand, even if Bayesian analyses do not need to resort to FDR or related quantities to control for multiplicities, many Bayesian analyses use averaged posterior FDR's to choose a cut-of point for the posterior probabilities (Genovese and Waserman, 2002; Newton et al., 2004; Broët et al., 2004; Do et al., 2005; House et al., 2006, among others).

As Bickel (2004) points out, neither Benjamini and Hochberg FDR, nor Storey 'positive' FDR have clear decision theoretical justification, unless, of course, they are forced as primitives into a 'global' loss function. But when this is done, resulting decision rules can be shown to have undesirable behavior (Müller et al., 2004). The loss functions considered are usually variant of the $0 - k_i$ loss; more realistic loss functions, in which the loss for a false acceptance depends on the strength of the signal erroneously 'missed', are contemplated in the Bayesian analyses of Duncan (1965), Waller and Duncan (1969), and Scott and Berger (2006).

References

[1] Y. Benjamini and Y. Hochberg, *Controlling the false discovery rate: a practical and powerful approach to multiple testing*, Journal of the Royal Statistical Society B **85** (1995), 289–300.

[2] Y. Benjamini and Y. Hochberg, *The adaptive control of the false discovery rate in multiple hypothesis testing with independent statistics*, J. Educ. Behav. Statist. **25** (2000), 60–83.

[3] D.R. Bickel, *Error-rate and decision-theoretic methods of multiple testing: Which genes have high objective probabilities of differential expression?* , Statistical Applications in Genetics and Molecular Biology **3(1) 8** (2004), http://www.bepress.com/sagmb/vol3/iss1/art8 (2004).

[4] M.A. Black, *A Note on the adaptive control of the false discovery rates*, J.R. Statist. Soc. B **66** (2004), 297–304.

[5] P. Broët, A. Lewin, S. Richardson, C. Dalmasso and H. Magdelenat, *A mixture model based strategy for selecting sets of genes in multiclass response microarray experiments*, Bioinformatics **20** (16) (2004), 2562-2571.

[6] K-A. Do, P. Müller and F. Tang, *A Bayesian Mixture Model for Differential Gene Expression*, Applied Statistics **54** (3) (2005), 627–644.

[7] S. Dudoit, J.P. Shaffer and J.C. Boldrick, *Multiple hypothesis testing in microarray experiments*, Statistical Science **18** (2003), 71–103.

[8] D.B. Duncan, *A Bayesian approach to multiple comparisons*, Technometrics **7** (1965), 171-222.

[9] B. Efron, R. Tibshirani, J.D. Storey and V. Tusher, *Empirical Bayes analysis of a microarray experiment*, Journal of the American Statistical Association **96** (2001), 1151–1160.

[10] H. Finner and M. Roters, *On the false discovery rate and expected Type I errors*, Biometrical Journal **43** (2001), 895–1005.

[11] C.R. Genovese and L. Wasserman, *Operating characteristics and extensions of the false discovery rate procedure*, Journal of the Royal Statistical Society **64** (2002), 499–518.

[12] C.R. Genovese and L. Wasserman, *Bayesian and frequentist multiple testing*, Bayesian Statistics **7** (2003) (J.M. Bernardo, M.J. Bayarri, J.O. Berger, A.P. Dawid, D. Heckerman, A.F.M. Smith, and M. West, eds.), Oxford University Press, 145–162.

[13] M. Gönen, P.H. Westfall, W.O. Johnson, *Bayesian multiple testing for two-sample multivariate endpoints*, Biometrics **59** (2003), 76–82.

[14] L.L.House, M.A. Clyde and Y.-C.T. Huang, *Bayesian Analysis* (2006), posted online July 29, 2005. http://ba.stat.cmu.edu/journal/forthcoming/house.pdf

[15] M. Langaas, B.H. Lindqvist and E. Ferkingstad, *Estimating the proportion of true null hypotheses with application to DNA microarray data*, J.R. Statist. Soc. B **67** (2005), 555–572.

[16] P. Müller, G. Parmigiani, C. Robert and J. Rouseau, *Optimal Sample Size for Multiple Testing: the Case of Gene Expression Microarrays*, Journal of the American Statistical Association **99** (2004), 990–1001.

[17] M.A. Newton and C.M. Kendziorski, *Parametric empirical Bayes methods for microarrays*. The Analysis of Gene Expression Data: Methods and Software (G. Parmigiani, E.S. Garret, R.A. Irizarri, and S.L. Zeger, ed.), Springer (2003), 254–271.

[18] M.A.Newton,C.M.Kendziorski,C.S.Richmon, F.R.Blattner and K.W.Tsui,*On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data*, Journal of Computational Biology **8** (2001), 37–52.

[19] M.A.Newton, D. Sarkar and P. Ahlquist, *Detecting differential gene expression with a semiparametric hierarchical mixture model*, Biostatistics **5** (2004), 155-176.

[20] G. Scott and J.O. Berger, *An exploration of Bayesian multiple testing*, Journal of Statistical Planning and Inference (2006)(in press).

[21] J.P. Shaffer, *Multiple hypothesis testing: a review*, Annual Review of Psychology **46** (1995), 561–584. Also Technical Report # 23, National Institute of Statistical Sciences.

[22] R.J. Simes, *An improved Bonferroni procedure for multiple tests of significance*, Biometrika **73** (1986), 751–754.

[23] J.D. Storey, *A direct approach to false discovery rates*, Journal of the Royal Statistical Society B **64** (2002), 479–498.

[24] J.D. Storey, *The positive false discovery rate: a Bayesian interpretation and the q-value*, Annals of Statistics **31** (2003), 2013–2035.
[25] M.A. Stoto, M. Schonlau, and L.T. Mariano, *Syndromic Surveillance: Is it worth the effort?*, Chance **17** (2004), 19–24.
[26] R.A. Waller and D.B. Duncan, *A Bayes rule for the symmetric multiple comparison problem*, Journal of the American Statistical Association **64** (1969), 1484-1503.
[27] Y. Yang and G.A. Rempala, *A note on multiple tests for gene expression data*, Unpublished manuscript (2004).

Regularized inversion methods and error bounds for general statistical inverse problems with application to density estimation of young massive cluster luminosities in the Antennae galaxies

Nicolai Bissantz
(joint work with Lutz Dümbgen, Hajo Holzmann, Axel Munk, Fritz Ruymgaart)

In this paper we are concerned with estimating a function of interest f in a Hilbert space $\mathbb{H}_1$ from indirect noisy measurements

$$Y = Kf + \sigma\varepsilon,$$

related to f by a known operator $K : \mathbb{H}_1 \to \mathbb{H}_2$ mapping $\mathbb{H}_1$ to another Hilbert space $\mathbb{H}_2$. Here, σ denotes the variance of the random noise, and the stochastic error ε is a Hilbert space process with $\|\mathbf{Cov}_\varepsilon\| \leq 1$. We introduce general regularization estimators for the estimation of f. This includes Tikhonov type and spectral cut-off estimators as well as iterative methods (e.g. Brackhage, 1987, Engl, Hanke & Neubauer, 1996, and Mair & Ruymgaart, 1996), such as ν-methods and the Landweber iteration. We analyse their convergence properties in statistical error models. It turns out that the latter estimators achieve the same (optimal) convergence rates as spectral cut-off, but do not require explicit spectral information on the operator and are often much faster to compute than Tikhonov regularization (Bissantz et al., 2005). We demonstrate application of a ν-method to estimation of the luminosity density of young massive star cluster in the Antennae galaxies (Anders et al, 2005). Here, the variance of the measurement error depends on the brightness of the specific cluster under consideration. Therefore, this is not a density deconvolution problem, and standard methods (e.g. Fourier-domain based algorithms) cannot be applied.

In the final part of the talk we discuss estimation of confidence bands for the function of interest f in ordinary smooth deconvolution problems, i.e. K is convolution with an error distribution ψ. To this end we study the study the supremum of the process

$$Y_n(t) := \frac{n^{1/2}h^{\beta+1/2}}{g(t)^{1/2}} \left(\hat{f}_n(t) - E\hat{f}_n(t) \right),$$

where n is the sample size, h is the bandwidth of the spectral cut-off estimator $\hat{f}_n$, $\beta \geq 0$ is such $\Phi_\psi(t)t^\beta \to C$, $t \to \infty$ for the characteristic function Φ_ψ of ψ and some constant $C \in \mathbb{C}$, and $g = f * \psi$. We derive the asymptotic distribution of $Y_n(t)$ and use it to construct asymptotic confidence bands for f (Bissantz, et al., 2005). Moreover, we establish a bootstrap version of the confidence bands, and give an application to measurements of the metallicity of local F and G dwarf stars where we confirm the "G dwarf problem".

References

[1] P. Anders, N. Bissantz, L. Boysen, U. Fritze-v. Alvensleben and R. de Grijs, *The luminosity distribution of young massive clusters in the Antennae galaxies* (2005), In preparation.

[2] N. Bissantz, T. Hohage, A. Munk and F. Ruymgaart, *Convergence rates of general regularization methods for statistical inverse problems and applications* (2005), Submitted.

[3] N. Bissantz, L. Dümbgen, H. Holzmann and A. Munk, *Nonparametric confidence bands in deconvolution density estimation* (2005), In preparation.

[4] H. Brakhage, *On ill-posed problems and the method of conjugate gradients*, Inverse and Ill-Posed Problems, eds. Engl, H. W. and Groetsch, C. W., Academic Press, Orlando (1987), 191–205.

[5] H.W. Engl, M. Hanke and A. Neubauer, *Regularization of Inverse Problems*, Kluwer Academic Publisher, Dordrecht, Boston, London (1996).

[6] B.A. Mair and F. Ruymgaart, *Statistical inverse estimation in Hilbert scales*, SIAM J. Appl. Math. **56** (1996), 1424-1444.

Bayesian Nonparametric Function Estimation Using Overcomplete Representations and Lévy Random Field Priors

Merlise A. Clyde

(joint work with Leanna House, Chong Tu, Robert L. Wolpert)

We consider the problem of nonparametric function estimation using overcomplete representations. The canonical setup for nonparametric regression problem consists of having n noisy measurements $\{Y_1, \ldots Y_n\}$ of an unknown real valued function $f : \mathbb{X} \to \mathbb{R}$ on some space $\mathbb{X}$,

$$Y_i = f(\mathbf{x}_i) + e_i \qquad e_i \overset{\text{iid}}{\sim} N(0, \sigma^2) \tag{1}$$

observed at points $\mathbf{x}_i \in \mathbb{X}$. The function $f(\cdot)$ is often regarded as an element of some separable Hilbert space $\mathbb{H}$ of real-valued functions on $\mathbb{X}$, and is expressed as a linear combination of basis functions $\phi_j \in \mathbb{H}$:

$$f(\mathbf{x}_i) = \sum_{j=1}^{J} \phi_j(\mathbf{x}_i)\beta_j \tag{2}$$

with unique coefficients $\{\beta_j\}$. Of interest are bases that lead to sparse representations, where only a few of the coefficients β_j in the expansion are nonzero. In applications where functions exhibit non-stationarity, no single (especially orthonormal) basis will necessarily lead to a sparse representation (4; 13). Overcomplete dictionaries and frames (3; 10) provide a larger collection of generating elements $\{\phi_{\boldsymbol{\omega}}\}_{\boldsymbol{\omega}\in\boldsymbol{\Omega}}$ than with a single basis for $\mathbb{H}$, potentially allowing for more effective signal extraction and data compression for functions. Because of the redundancy inherent in the overcomplete representation, coefficients for expansions using the complete dictionary are no longer unique. This lack of uniqueness is advantageous, as it is possible to find a more parsimonious representation (by shrinking or forcing coefficients to zero) than those obtained using any single basis.

We consider dictionaries created from rescaling and translating a single generating function g, such as a mother wavelet or kernel function,

$$\phi_{\boldsymbol{\omega}}(\mathbf{x}) \equiv g(\mathbf{x}, \boldsymbol{\omega}) \qquad \boldsymbol{\omega} \in \boldsymbol{\Omega} \tag{3}$$

where $\boldsymbol{\Omega}$ is a complete separable metric space and g is a Borel measurable function $g : \mathbb{X} \times \boldsymbol{\Omega} \to \mathbb{R}$, such as a kernel function,

$$g(\mathbf{x}, \boldsymbol{\omega}) = \exp\{-\omega_1 ||\mathbf{x} - \boldsymbol{\omega}_2||^p\} \quad \text{where } \boldsymbol{\omega}^T = (\omega_1, \boldsymbol{\omega}_2^T) \tag{4}$$

or a wavelet function. The expansion in (2) may be generalized to the overcomplete representation

$$f(\mathbf{x}) = \sum_{j=1}^{J} g(\mathbf{x}, \boldsymbol{\omega}_j)\beta_j \equiv \int_{\boldsymbol{\Omega}} g(\mathbf{x}, \boldsymbol{\omega}) L(d\boldsymbol{\omega}) \tag{5}$$

where $L(d\boldsymbol{\omega}) = \sum \beta_j \delta_{\boldsymbol{\omega}_j}(d\boldsymbol{\omega})$ is a (possibly signed) Borel measure on $\boldsymbol{\Omega}$.

We describe a Bayesian method for inference regarding the unknown f in the sparse regression problem using overcomplete dictionaries. To make posterior inference about the unknown function $f \in \mathbb{H}$ in (1), we must first propose a prior distribution on $\mathbb{H}$ for f. With the representation of $f(\mathbf{x})$ in (5), this is equivalent to specifying a random signed Borel measure $L(d\boldsymbol{\omega})$ on $\boldsymbol{\Omega}$. An intuitive construction of such random measures begins by choosing any positive number $\nu_+ > 0$ and assigning J a Poisson distribution, $J \sim \mathsf{Poisson}(\nu_+)$. Then, conditionally on J, accord the $(\beta_j, \boldsymbol{\omega}_j) \in \mathbb{R} \times \boldsymbol{\Omega}$ independent identical distributions, $(\beta_j, \boldsymbol{\omega}_j) \overset{\text{iid}}{\sim} \pi(d\beta, d\boldsymbol{\omega})$, where π is a probability distribution on $\mathbb{R} \times \boldsymbol{\Omega}$. In that case, L will assign independent infinitely-divisible random variables $L(A_i)$ to disjoint Borel sets $A_i \subset \boldsymbol{\Omega}$. Such a random measure L determines naturally a Lévy random field $L[g]$, a continuous (in probability) linear mapping $g \to L[g]$ from Borel measurable functions $g : \mathbb{X} \times \boldsymbol{\Omega} \to \mathbb{R}$ to random elements in $\mathbb{H}$,

$$L[g] \equiv \int_{\boldsymbol{\Omega}} g(\cdot, \boldsymbol{\omega})\, L(d\boldsymbol{\omega}) = \sum_{j=1}^{J} g(\cdot, \boldsymbol{\omega}_j)\beta_j \tag{6}$$

with Lévy measure $\nu(d\beta, d\boldsymbol{\omega}) = \nu_+ \pi(d\beta, d\boldsymbol{\omega})$, the product of the Poisson rate ν_+ for J and the distribution $\pi(d\beta, d\boldsymbol{\omega})$ for $\{(\beta_j, \boldsymbol{\omega}_j)\}$. Here $\nu(\mathbb{R} \times \boldsymbol{\Omega})$ is finite (by construction), and $L[g]$ is equivalent to a compound Poisson random field.

The stochastic expansions in terms of wavelets of (1), which utilize a compound Poisson random field with normal prior distributions for β_j, may be viewed as a special case of a Lévy random field prior.

More generally, the Lévy measure $\nu(d\beta, d\omega)$ need not be finite for the random field $L[g]$) to be finite and well-defined with infinite J; it is sufficient for ν to satisfy the bound

$$\iint_{\mathbb{R}\times K} \left(1 \wedge |\beta|^2\right) \nu(d\beta, d\boldsymbol{\omega}) < \infty \tag{7}$$

for every compact $K \subset \Omega$. Examples of Lévy random fields with infinite Lévy measures include the Gamma random field, with Lévy measure

$$\nu(d\beta, d\boldsymbol{\omega}) = \lambda\gamma(d\boldsymbol{\omega})\beta^{-1}e^{-\varphi\beta}\mathbf{1}_{\{\beta>0\}}\, d\beta \tag{8}$$

for $\lambda > 0$ and for some σ-finite measure $\gamma(d\boldsymbol{\omega})$ on $\boldsymbol{\Omega}$, giving

$$L[A] \sim \mathsf{Gamma}(\lambda\gamma(A), \varphi)$$

for $A \subset \boldsymbol{\Omega}$ of finite γ measure, and the symmetric α-stable (SαS) for $0 < \alpha < 2$ (including the Cauchy process with $\alpha = 1$), with Lévy measure

$$\nu(d\beta, d\boldsymbol{\omega}) = c_\alpha\lambda\gamma(d\boldsymbol{\omega})|\beta|^{-1-\alpha}\, d\beta \tag{9}$$

for some constant $c_\alpha > 0$, giving $L[A] \sim \mathsf{Stable}(\alpha, 0, \lambda\gamma(A), 0)$. While such measures lead to an infinite number of support points J a priori, finite measures which permit tractable computation may be obtained by the approximation $\nu_\epsilon(d\beta, d\omega) \equiv \nu(d\beta, d\omega)\mathbf{1}_{|\beta|>\epsilon}$. The Lévy random fields based on such an approximation converge in distribution to $L[g]$ as $\epsilon \to 0$. Although we are interested in sparse representations, and hence finite J (a posteriori), the representation using infinite measures provides robustness to miss-specification due to a perhaps poor choice of generating function. For more background on Lévy random fields and approximations, see (9; 11; 7; 12; 2; 14).

The model may be restated in hierarchical fashion as

$$\begin{aligned}
Y_i \mid f(\mathbf{x}_i) &\overset{\text{ind}}{\sim} N(f(\mathbf{x}_i), \sigma^2) \\
f(\mathbf{x}_i) &= \sum_{j=1}^{J} g(\mathbf{x}_i, \boldsymbol{\omega}_j)\beta_j \\
(\beta_j, \boldsymbol{\omega}_j) \mid J &\overset{\text{iid}}{\sim} \pi(d\beta_j, d\boldsymbol{\omega}_j) \equiv \frac{\nu_\epsilon(d\beta_j, d\boldsymbol{\omega}_j)}{\nu_\epsilon(\mathbb{R}, \boldsymbol{\Omega})} \qquad \text{for } j = 1, \ldots, J \\
J &\sim \mathsf{Poisson}(\nu_+) \quad \text{where } \nu_+ \equiv \nu_\epsilon(\mathbb{R}, \boldsymbol{\Omega})
\end{aligned}$$

where J is the random number of terms in the stochastic expansion, $(\beta_1, \ldots, \beta_J)$ represents the unknown coefficients and $(\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_J)$ represents the collection of generator specific parameters. We also place a prior distribution on parameters in the Lévy measure. A Gamma for λ in the ϵ approximation to the Lévy measure for the Gamma (8) or Stable (9) random field leads to J having a Negative Binomial distribution, which leads to robustness to a fixed choice of λ. A prior distribution on σ^2 completes the model specification.

Given observations $\mathbf{Y} \equiv (Y_1, \ldots, Y_n)^T$, the log of the (conditional) posterior distribution is

$$\log \pi\left(\{\beta_j, \boldsymbol{\omega}_j\}_{j=1}^J, J \mid \sigma^2, \mathbf{Y}\right) = \text{constant} - \tag{10}$$
$$\left\{ \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(Y_i - f(\mathbf{x}_i)\right)^2 + \log(J!) - \sum_{j=1}^{J} \nu_\epsilon(d\beta_j, d\omega_j) \right\}$$

which takes the form of a penalized or regularized likelihood. Model complexity is penalized directly through the $\log(J!)$ term as in a ℓ_0 penalty which penalizes the number of coefficients in the expansion. Model complexity is also indirectly penalized through the choice of Lévy measure ν.

While expressions for posterior modes or posterior distributions of quantities of interest do not exist in closed form, the prior construction using Lévy random fields permits tractable posterior simulation via a reversible jump Markov chain Monte Carlo algorithm (6). Efficient computation is possible because updates to f based on adding/deleting or updating single dictionary elements bypass the need to invert large matrices. Furthermore, because dictionary elements $g(\mathbf{x}, \boldsymbol{\omega})$ are only computed as needed, memory requirements scale linearly with the sample size.

We compare the performance of estimators using the Levy random field priors to estimators based on translational invariant wavelets (8) on simulated data using standard wavelet test functions Blocks, Bumps, Doppler and Heavysine (5). For the test functions Blocks, Bumps, and Heavisine, we achieve a gain in mean squared error efficiency of $13.7\% - 56.3\%$.

We explore Lévy random field priors in several challenging applications. In the first, we use Gamma random field priors to construct models for the latent relative abundance of proteins as a function of their mass/charge (or equivalently time of flight) using data from Matrix Assisted Laser Desorption/Ionization Time of Flight mass spectroscopy. Normalized Gaussian kernels with time varying scale parameters provide a natural choice of generating functions to capture the variation in time of flight of proteins of a given mass/charge. Unlike wavelets or spline models, the parameters in the adaptive kernel model have interesting biological interpretations: J is the number of unknown proteins in the sample, and β_j is the unknown concentration for a protein with expected time of flight τ_j and resolution ρ_j, here we take $\boldsymbol{\omega}_j \equiv (\tau_j, \rho_j)$. This interpretability is a key feature of the Lévy random field models, as it allows us to incorporate subjective prior information regarding resolution and time of flight (a transformation of the mass/charge).

The second area of application concerns development of non-stationary temporal, spatial and spatial-temporal models for concentrations of one or more criteria pollutants. As expected pollution concentrations are inherently non-negative, the Lévy random field priors based on a Gamma random field ensure that the expected functions are non-negative, and are a natural alternative to the commonly adopted Gaussian random field priors in Bayesian nonparametrics. The spatial-temporal locations of jumps in the Lévy random field may be interpreted as point sources of

pollution, with dispersal over time and space controlled by additional parameters in the kernels. Hierarchical models for parameters in the Lévy measure allow incorporation of meteorological variables which influence both the dispersal parameters and expected concentrations. An interesting extension with the multivariate pollution models is the use of marked random fields that allow common jumps (shared impulses) between two or more pollutants. In comparison with standard methods, the Lévy random field priors provide excellent performance in terms of both mean squared error and coverage for out-of-sample model predictions.

Papers describing the Lévy random field priors and these applications will be available from the authors' websites; please visit `http://www.isds.duke.edu`.

Acknowledgments

This work was made possible by National Science Foundation grants DMS–0342172, DMS–0422400, DMS–0406115.

References

[1] Felix Abramovich, Theofanis Sapatinas and Bernard W. Silverman, *Wavelet thresholding via a Bayesian approach*, J. Roy.Statist. Soc. Ser. B **60(1)** (1998), 1–52, .

[2] Rama Cont and Peter Tankov, *Financial modelling with jump processes*, Chapman & Hall, London, UK (2004).

[3] Ingrid Daubechies, *Ten Lectures on Wavelets*, CBMS-NSF Regional Conference Series in Applied Mathematics Volume, SIAM, Philadelphia, PA, **61** (1992).

[4] David L. Donoho and Michaell Elad, *Maximal sparsity representation via l_1 minimization*, Proc. Nat. Aca. Sci. **100** (2003), 2197–2202.

[5] David L. Donoho and Iain M. Johnstone, *Ideal spatial adaptation by wavelet shrinkage*, Biometrika **81** (1994), 425–455.

[6] Peter J. Green, *Reversible jump Markov chain Monte Carlo computation and Bayesian model determination*, Biometrika **82** (1995), 711–732.

[7] Jean Jacod and Albert N. Shiryaev, *Limit Theorems for Stochastic Processes*, Grundlehren der mathematischen Wissenschaften, Springer-Verlag, Berlin, DE, Volume **288** (1987).

[8] Iain M. Johnstone and Bernard W. Silverman, *Empirical Bayes selection of wavelet thresholds*, Ann. Statist. **33** (2005), 1700–1752.

[9] Alexander Ya. Khinchine and Paul Lévy, *Sur les lois stables*, C. R. Acad. Sci. Paris **202** (1936), 374–376.

[10] Stéphane G. Mallat and Zhifeng Zhang, *Matching pursuit with time-frequency dictionaries*, IEEE T. Signal Proces. **41** (1993), 3397–3415.

[11] G. Maruyama, *Infinitely divisible processes*, Theor. Probab. Appl. **15(1)** (1970), 1–22.

[12] Ken-iti Sato, *Lévy Processes and Infinitely Divisible Distributions*, Cambridge Univ. Press, Cambridge, UK, (1999).

[13] Patrick J. Wolfe, Simon J. Godsill and Wee-Jing Ng, *Bayesian variable selection and regularisation for time-frequency surface estimation*, J. Roy.Statist. Soc. Ser. B **66** (2004), 575–589.

[14] Robert L. Wolpert and Murad S. Taqqu, *Fractional Ornstein-Uhlenbeck Lévy processes and the Telecom process: Upstairs and downstairs*, Signal Processing **85(8)** (2005),1523–1545.

Empirical process techniques for locally stationary processes

Rainer Dahlhaus

(joint work with Wolfgang Polonik)

We consider inference for locally stationary processes (cf. Dahlhaus, 1997, 2000), that is for processes $X_{t,n}$ $(t = 1, \ldots, n)$ which have a slowly-varying moving average representation

$$X_{t,n} = \sum_{j=-\infty}^{\infty} a_{t,n}(j)\, \varepsilon_{t-j}. \tag{1}$$

We assume that the coefficient functions $a_{t,n}(j)$ are uniformly decaying to zero with rate $|j|^{-1} \log^{-(1+\kappa)} |j|$ and can approximately be rescaled to the unit interval, that is $a_{t,n}(j)$ can be approximated by $a(\frac{t}{n}, j)$ with a function $a(u, j)$ of bounded variation in u. The ε_t are assumed to be independent and identically distributed with $E\varepsilon_t \equiv 0$, $E\varepsilon_t^2 \equiv 1$ and $\kappa_4 := cum_4(\varepsilon_t)$ (for more details on these assumptions see Dahlhaus and Polonik, 2005). In addition we assume for some results that the sequence ε_t and therefore also the process $X_{t,n}$ is Gaussian. A simple example of a process which fulfills these assumptions is $X_{t,n} = \phi(\frac{t}{n})Y_t$ where $Y_t = \Sigma_j\, a(j)\varepsilon_{t-j}$ is stationary and ϕ is of bounded variation. Furthermore time varying ARMA models whose coefficient functions are of bounded variation are locally stationary in the above sense.

The function

$$f(u, \lambda) = \frac{1}{2\pi} |A(u, \lambda)|^2$$

with

$$A(u, \lambda) = \sum_{j=-\infty}^{\infty} a(u, j) \exp(-i\lambda j)$$

is the time varying spectral density, and

$$c(u, k) = \int_{-\pi}^{\pi} f(u, \lambda) \exp(i\lambda k) d\lambda = \sum_{j=-\infty}^{\infty} a(u, k + j)a(u, j)$$

is the time varying covariance of lag k at rescaled time u.

The goal now is to make statistical inference about the process - for example to estimate the coefficient functions of a time varying AR-process

$$X_{t,n} \,+\, a(\frac{t}{n})\, X_{t-1,n} = \sigma(\frac{t}{n})\, \epsilon_t$$

which can be represented in the form (1).

For such problems the so-called empirical spectral process plays a major role. It is defined by

$$E_n(\phi) = \sqrt{n}\Big(F_n(\phi) - F(\phi)\Big)$$

where

$$F(\phi) = \int_0^1 \int_{-\pi}^{\pi} \phi(u,\lambda)\, f(u,\lambda)\, d\lambda\, du$$

and

$$F_n(\phi) = \frac{1}{n}\sum_{t=1}^{n} \int_{-\pi}^{\pi} \phi(\frac{t}{n},\lambda)\, J_n(\frac{t}{n},\lambda)\, d\lambda$$

with the pre-periodogram

$$J_n\Big(\frac{t}{n},\lambda\Big) = \frac{1}{2\pi} \sum_{k:1\le[t+1/2\pm k/2]\le n} X_{[t+1/2+k/2],n} X_{[t+1/2-k/2],n} \exp(-i\lambda k).$$

If $X_{[t+1/2+k/2],n} X_{[t+1/2-k/2],n}$ is regarded as a (raw-) estimate of $c(\frac{t}{n},k)$ then $J_n(\frac{t}{n},\lambda)$ can be regarded as a (raw-) estimate of $f(\frac{t}{n},\lambda)$ - however, in order to become consistent $J_n(\frac{t}{n},\lambda)$ needs to be smoothed in time and frequency direction. The pre-periodogram J_n was first defined by Neumann and von Sachs (1997).

Many important statistics occurring in the analysis of locally stationary processes can be written as a functional of $F_n(\phi)$:

(i) For $\phi(u,\lambda) = \chi_{[0,v]\times[0,\mu]}(u,\lambda)$ one obtains the time-spectral measure.

(ii) For $\phi(u,\lambda) = \frac{\partial}{\partial\theta_j} f_{\theta_0}(u,\lambda)^{-1}$ one obtains the score function of the generalized Whittle likelihood for parametric locally stationary models (cf. Dahlhaus, 2000).

(iii) For kernel functions $\phi(u,\lambda) = \phi_n(u,\lambda)$ we obtain several other applications: If $k_n(x) = \frac{1}{b_n} K(\frac{x}{b_n})$ is some kernel with bandwidth b_n then $F_n(\phi)$ occurs as an estimate of the spectral density in the stationary case where $\phi(u,\lambda) = k_n(\lambda - \lambda_0)$, as an estimate of $f(u,\lambda)$ in the nonstationary case where $\phi(u,\lambda) = k_n(u - u_0)\, k_n(\lambda - \lambda_0)$ and as an estimate of the time varying covariance function at time u_0 and of lag k where $\phi(u,\lambda) = \cos(\lambda k)\, k_n(u - u_0)$.

(iv) For $\phi(\frac{t}{n},\lambda) = \tilde{\phi}(\lambda)$ we obtain with the relation

$$I_n(\lambda) := \frac{1}{2\pi n}\Big|\sum_{t=1}^{n} X_{t,n}\exp(-i\lambda t)\Big|^2 = \frac{1}{n}\sum_{t=1}^{n} J_n^{(h_n)}(\frac{t}{n},\lambda)$$

$$F_n(\phi) = \int_{-\pi}^{\pi} \tilde{\phi}(\lambda) I_n(\lambda)\, d\lambda$$

leading to well known statistics for stationary time series, such as the empirical covariance function of some lag k where $\tilde{\phi}(\lambda) = \cos(\lambda k)$, the score function of the parametric Whittle likelihood where $\tilde{\phi}(\lambda) = \frac{\partial}{\partial\theta_j} f_{\theta_0}(\lambda)^{-1}$ and the empirical spectral measure where $\tilde{\phi}(\lambda) = \chi_{[0,\alpha]}(\lambda)$.

The asymptotic properties for the empirical spectral process are derived under conditions on the richness of the index class Φ measured by metric entropy. For each $\epsilon > 0$, the covering number of Φ with respect to the metric $\rho_2(\phi) = \left(\int_0^1 \int_{-\pi}^{\pi} |\phi(u,\lambda)|^2 \, d\lambda \, du\right)^{1/2}$ is defined as

$$N(\epsilon, \Phi, \rho_2) = \inf\{n \geq 1 : \exists\, \phi_1, \ldots, \phi_n \in \Phi \text{ such that } \forall\, \phi \in \Phi \;\; \exists\, 1 \leq i \leq n \text{ with } \rho_2(\phi - \phi_i) < \epsilon\}.$$

The quantity $H(\epsilon, \Phi, \rho_2) = \log N(\epsilon, \Phi, \rho_2)$ is called the metric entropy of Φ with respect to ρ_2.

We now can formulate the following functional central limit theorem (for details about the assumptions see Dahlhaus and Polonik, 2005).

Theorem 1 *Suppose that $X_{t,n}$ is a Gaussian locally stationary process and let Φ be a class of functions with*

$$\int_0^1 H(u, \Phi, \rho_2)\, du \; < \; \infty.$$

Then the process $(E_n(\phi);\, \phi \in \Phi)$ converges weakly in $\ell^\infty(\Phi)$ to a tight mean zero Gaussian process $(E(\phi);\, \phi \in \Phi)$ with

$$cov\big(E(\phi_j), E(\phi_k)\big) = 2\pi \int_0^1 \frac{h^4(u)}{\|h\|_2^4} \int_{-\pi}^{\pi} \phi_j(u,\lambda)\,[\phi_k(u,\lambda) + \phi_k(u,-\lambda)]\, f^2(u,\lambda)\, d\lambda\, du.$$

Besides investigating the properties of the empirical process we have looked in Dahlhaus and Polonik (2005) at the nonparametric MLE (and related sieve estimates) defined by

$$\widehat{f}_n = \mathrm{argmin}_{g \in \mathcal{F}} \mathcal{L}_n(g)$$

where

$$\mathcal{L}_n(g) = \frac{1}{n} \sum_{t=1}^{n} \frac{1}{4\pi} \int_{-\pi}^{\pi} \{\log g(\frac{t}{n}, \lambda) + \frac{J_n(\frac{t}{n}, \lambda)}{g(\frac{t}{n}, \lambda)}\}\, d\lambda.$$

and $\mathcal{F}$ is a suitable class of spectral densities (e.g. under shape restrictions). A detailed example is given in Section 3 of Dahlhaus and Polonik (2005) where the estimation of a monotonic variance function in a time-varying AR-model is studied, including explicit algorithms involving isotonic regression.

By using again the empirical spectral process consistency and a rate of convergence is derived of $\widehat{f}_n$. Furthermore an optimal rate is obtained for sieve estimates.

A key role in the proof (as well as in the proof of the central limit theorem) is played by the following exponential inequalities (the first being a Bernstein-type inequality). Let

$$\rho_{2,n}(\phi) := \Big(\frac{1}{n} \sum_{t=1}^{n} \int_{-\pi}^{\pi} \phi(\frac{t}{n}, \lambda)^2 d\lambda\Big)^{1/2} \quad \text{and} \quad \rho_\infty(\phi) := \sum_{j=-\infty}^{\infty} \sup_u |\hat{\phi}(u, j)|.$$

where $\hat{\phi}(u,j)$ are the Fourier coefficients of $\phi(u,\lambda)$ in frequency direction.

Theorem 2 *Suppose that $X_{t,n}$ is a Gaussian locally stationary process and let the Fourier coefficients of ϕ be of uniformly bounded variation. Then we have for all $\eta > 0$*

$$P(|\sqrt{n}(F_n(\phi) - EF_n(\phi))| \geq \eta) \leq c_1 \exp\Big(- c_2 \frac{\eta^2}{\rho_{2,n}(\phi)^2 + \frac{\eta \rho_\infty(\phi)}{\sqrt{n}}} \Big)$$

and

$$P(|\sqrt{n}(F_n(\phi) - EF_n(\phi))| \geq \eta) \leq c_1 \exp\Big(- c_2 \frac{\eta}{\rho_{2,n}(\phi)} \Big)$$

with some constants c_1, $c_2 > 0$.

We mention that both theorems will we extended to non-Gaussian processes in Dahlhaus and Polonik (2006). Furthermore the case of ϕ depending on n shall be studied in future work.

References

[1] R. Dahlhaus, *Fitting time series models to nonstationary processes*, Ann. Statist. **25** (1997), 1–37.

[2] R. Dahlhaus, *A likelihood approximation for locally stationary processes*, Ann. Statist. **28** (2000), 1762-1794.

[3] R. Dahlhaus and W. Polonik, *Nonparametric quasi maximum likelihood estimation for Gaussian locally stationary processes*, preprint (2005).

[4] R. Dahlhaus and W. Polonik, *Empirical spectral processes for locally stationary time series*, in preparation (2006).

[5] M.H. Neumann and R. von Sachs, *Wavelet thresholding in anisotropic function classes and applications to adaptive estimation of evolutionary spectra*. Ann. Statist. **25** (1997), 38 –76.

Universal principles, approximation and model choice

Patrick Laurie Davies

Given data set $\mathbf{x}_n = (x_1, \ldots, x_n)$ and a family $\{P_\theta : \theta \in \Theta\}$ of probability models statisticians have developed different procedures for specifying one or more values of the parameter space Θ such that P_θ is in some sense an appropriate model for the data. Many of these procedures are what may be termed universal as the choice of parameter, in contrast to the choice of model family, is independent of the subject matter of the data. That is, the same procedure can be applied to data from physics, literature and sociology. Examples of such universal procedures are maximum likelihood, Bayes, AIC, BIC, MDL and cross validation. The majority

of these is likelihood based so we restrict the present discussion to such ones. If the density of the model P_θ is $f(x, \theta)$ then the log-likelihood is

$$l(\theta, \mathbf{x}_n) = \log f(\mathbf{x}_n, \theta).$$

We consider three applications of likelihood.

Example 1.
The model we use is the i.i.d. Gaussian one whose log-likelihood is

$$l(\mu, \sigma, \mathbf{x}_N) = -\frac{n}{2}\log(2\pi) - n\log(\sigma) - \frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i - \mu}{\sigma}\right)^2.$$

If now the data were i.i.d. Cauchy there is no way we could tell this just given the likelihood. The conclusion is that likelihood is blind.

Example 2:
The data is a large sample of numbers obeying the binomial distribution with parameters $n = 5000$ and $p = 1/2$. The models we consider are the Poisson model with parameter $\lambda = 1$ and the normal $N(2500, 1250)$ model. The latter is well supported by the data and the central limit theorem. Nevertheless all likelihood methods would choose the Poisson model. Likelihood is pathologically discontinuous.

Example 3:
A common maximum penalized likelihood estimator is the function f which minimizes

$$\sum_{i=1}^{n}(y_i - f_i)^2 + \lambda\int_0^1 f^{(2)}(t)^2\,dt$$

where λ is the smoothness parameters. Likelihood provides no help in choosing λ and indeed, there may well be no acceptable choice of λ. Figures 1 and 2 below show respectively a large and a small value of λ respectively. In summary

- Likelihood is blind.
- Likelihood reduces the measure of fit of a model to data to a *single number*.
- Likelihood is pathologically discontinuous.
- Likelihood is only useful when combined with a regularization.

In contrast to likelihood which is closely related to the idea of truth in statistics we propose to base model choice on a concept of approximation. The idea is that a model P is an adequate approximation for a data set $\mathbf{x}_n$ if "typical" samples of size n $\mathbf{X}_n(P)$ generated under P "look like" the real data. The notions of "typical" and "look like" have to be made precise and this will depend on probabilistic and subject matter considerations. We expound the approach in the context of nonparametric regression. Given data $(t_i, y(t_i)), i = 1, \ldots, n$ with the $t_i \in [0, 1]$ we look for a function f_n such that the data may well be represented by $(t_i, f_n(t_i))$. We do this in two steps. We define precisely what is meant by an adequate

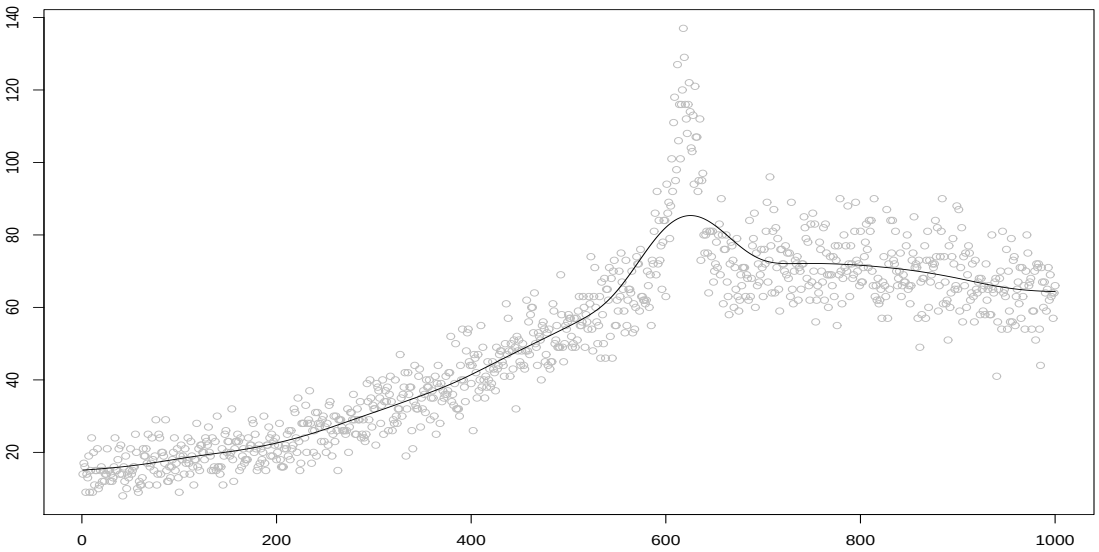


FIGURE 1.

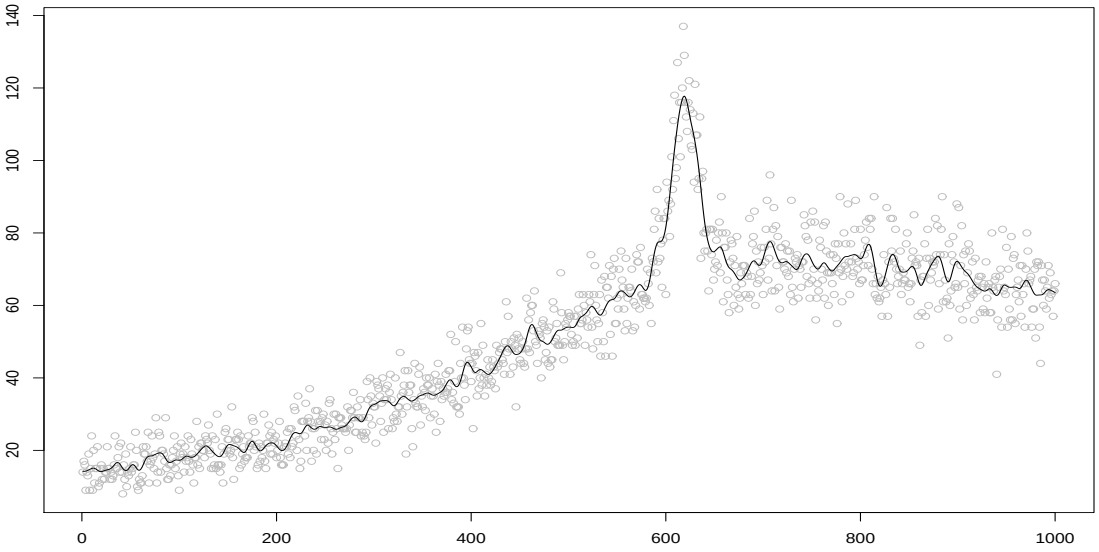


FIGURE 2.

approximation and then we regularize by specifying the simplest function which is an adequate approximation. We base our considerations on the model

$$Y(t) = f(t) + \sigma Z(t) \tag{1}$$

where Z denotes standard Gaussian white noise. If f_n is an adequate approximation to the data then based on(1) the corresponding residuals $r_n(t_i) = y(t_i) - f_n(t_i)$ must "look like" white noise with variance σ^2. If this is the case then for each interval $I \subset [0, 1]$

$$\frac{1}{\sqrt{|I|}} \sum_{t_i \in I} (y(t_i) - f_n(t_i)) \tag{2}$$

will behave like a $N(0, \sigma^2)$ random variable. Based on the maximum of Gaussian random variables we are lead to the requirement

$$\max_I \left| \frac{1}{\sqrt{|I|}} \sum_{t_i \in I} (y(t_i) - f_n(t_i)) \right| \leq \sigma \sqrt{\tau \log(n)} \tag{3}$$

which holds asymptotically for any $\tau > 2$. Our default value is 2.3. The value of σ can usually be obtained with sufficient accuracy from the data. In this sense the function of Figure 2 is an adequate approximation but not that of Figure 1. The second step is to define what is meant by a simple function. One useful measure is the number of local extremes. The problem now is to minimize the number of local extremes of f_n subject to the approximation inequalities (3). The taut string method to do this was developed in Davies and Kovac (2001). Figure 3 shows the taut string approximation to the same data. Another possibility is to maximize

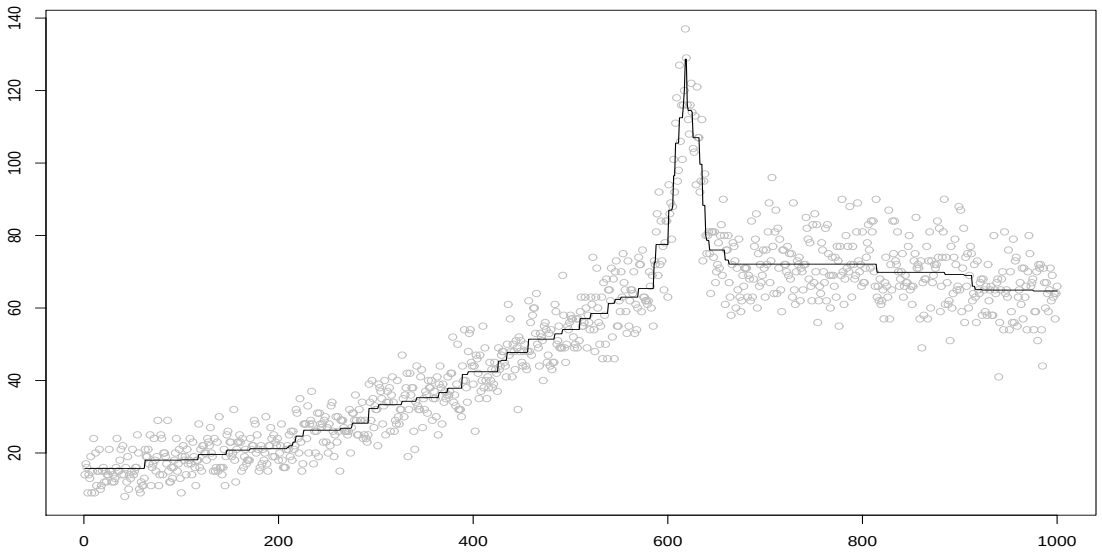


FIGURE 3.

the smoothness of the function. This may be accomplished by minimizing the total variation of the third derivative of f_n subject to the constraints (3). This leads to the following linear programming problem:

$$\text{(4)} \qquad \text{minimize } \sum_{i=1}^{n-4} |f_n(t_{i+4}) - 4f_n(t_{i+3}) + 6f_n(t_{i+2}) - 4f_n(t_{i+1}) + f_n(t_i))|$$

subject to (4). If we also incorporate the monotonicity constraints from the taut string the resulting approximation is shown in Figure 4).

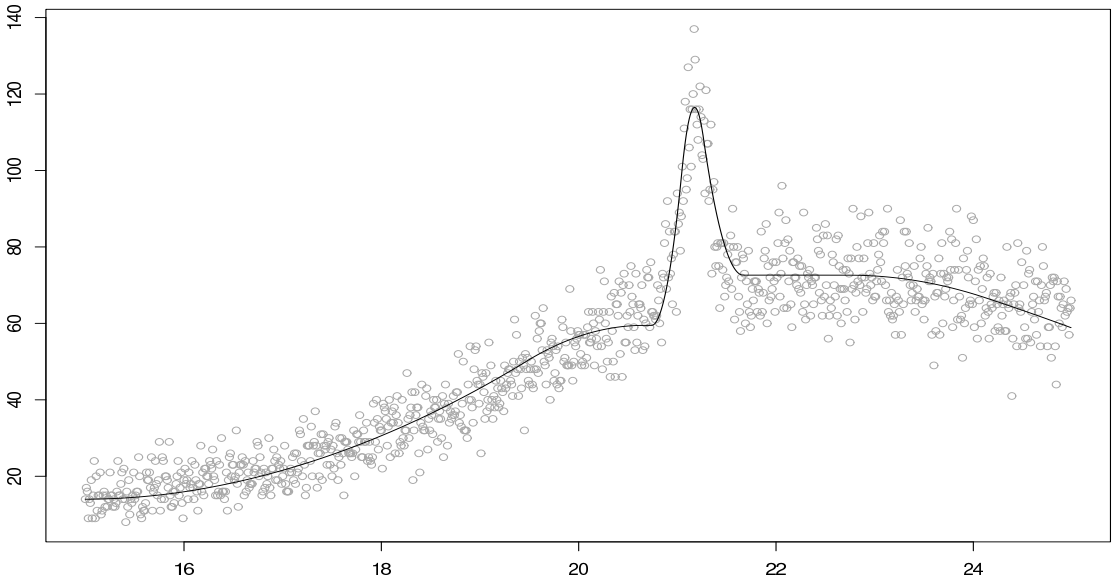


FIGURE 4.

REFERENCES

[1] P. L.. Davies and A. Kovac, *Local extremes, runs, strings and multiresolution (with discussion)*, Annals of Statistics **29** (2001), 1–65.

Prediction Regions for Nested Model Selection

LUTZ DUEMBGEN
(joint work with Angelika Rohde)

This talk is concerned with a new type of prediction regions in connection with model selection in a gaussian shift model. Before starting with this framework let us consider an even simpler setting.

A Toy problem. Suppose we observe a stochastic process $Y = (Y(t))_{t\in[0,1]}$, where

$$Y(t) \;=\; F(t) + W(t), \quad t \in [0,1],$$

with a fixed or random continuous function F on $[0,1]$ and a brownian motion $W = (W(t))_{t\in[0,1]}$ such that F and W are independent. We are interested in the set

$$S(F) \;:=\; \arg\min_{t\in[0,1]} F(t).$$

Precisely, we want to construct a $(1-\alpha)$–prediction region $\hat{S} = \hat{S}(Y) \subset [0,1]$ for $S(F)$, i.e.

$$P(S(F) \subset \hat{S}) \;\geq\; 1-\alpha, \tag{1}$$

regardless of (the distribution of) F. A possible solution is as follows: Let κ be the $(1-\alpha)$–quantile of

$$T(W) \;:=\; \sup_{s,t\in[0,1]} \Big(\frac{|W(s)-W(t)|}{\sqrt{|s-t|}} - \Gamma(|s-t|) \Big),$$

where $\Gamma(\delta) := \sqrt{2\log(e/\delta)}$. Then

$$\hat{S} \;:=\; \Big\{ s\in[0,1] : Y(s) \leq Y(t) + \sqrt{|s-t|}\big(\Gamma(|s-t|)+\kappa\big) \text{ for all } t\in[0,1] \Big\}$$

satisfies the constraint (1).

This method is motivated by multiscale methods introduced by Dümbgen and Spokoiny (2001). In order to understand its power, consider a sequence of fixed functions $F = F_n$ such that for some parameters $t_n \in [0,1]$, $\delta_n \in (0,1]$ and $c_n > 0$,

$$F_n(t_n+h) \;\geq\; c_n h^2 \quad \text{whenever } |h| \leq \delta_n,$$

where $c_n\delta_n^2 \to \infty$. Then

$$\hat{S}_n \cap [t_n \pm \delta_n] \;\subset\; \Big[t_n \pm O_p(\log(c_n)^{1/3} c_n^{-2/3}) \Big].$$

Nested Models. Now consider a random vector $X_n \sim \mathcal{N}_n(\theta_n, \sigma^2 I_n)$, where $\sigma > 0$ is assumed to be known (for simplicity), whereas θ_n is an unknown vector in some set $\Theta_n \subset \mathbb{R}^n$. Given an estimator $\hat{\theta}_n = \hat{\theta}_n(X_n)$ for θ_n, let

$$L(\hat{\theta}_n, \theta_n) \;:=\; \|\hat{\theta}_n - \theta_n\|^2 \quad \text{and} \quad R(\hat{\theta}_n, \theta_n) \;:=\; \mathbb{E}L(\hat{\theta}_n, \theta_n)$$

be its loss and risk, respectively.

Depending on Θ_n, various adaptivity results are known for point estimators of θ_n, many of which have the following form: Let $\mathcal{C}_n$ be a family of candidate estimators $\check{\theta} = \check{\theta}(X_n)$ for θ_n. Then there exist estimators $\hat{\theta}_n$ and constants $A_n, B_n = O(\log(n)^\kappa)$ for some $\kappa \geq 0$ such that for arbitrary $\theta_n \in \Theta_n$,

$$R(\hat{\theta}_n, \theta_n) \;\leq\; A_n \inf_{\check{\theta}\in\mathcal{C}_n} R(\check{\theta}, \theta_n) + B_n\,.$$

Results of this type are provided, for instance, by Polyak and Tsybakov (1991) and Donoho and Johnstone (1994).

By way of contrast, when aiming at adaptive confidence sets one faces severe limitations.

Here is a result of Li (1989), slightly reformulated: Let $\hat{C}_n = \hat{C}_n(X_n) \subset \mathbb{R}^n$ be a $(1-\alpha)$–confidence set for $\theta_n \in \Theta_n$. Suppose that Θ_n contains a closed euclidean ball $B(\theta_n^o, cn^{1/4})$. Then

$$\begin{aligned} \Pr_{\theta_n^o}\Big(\sup_{\eta \in \hat{C}_n} \|\eta - \theta_n^o\| < cn^{1/4} \Big) &\leq \Pr(\chi_n^2 \leq \chi_{n;\alpha}^2(c^2 n^{1/2})) \\ &= \Phi(\Phi^{-1}(\alpha) + 2^{-1/2} c^2) + o(1) \end{aligned}$$

as $n \to \infty$, uniformly in $c \geq 0$. Thus a reasonable confidence set cannot have a diameter of order $o_p(n^{-1/4})$. Despite these limitations, there is some literature on confidence sets in the present or similar settings; see for instance Beran and Dümbgen (1998), Baraud (2004), Genovese and Wassermann (2005), Robins and van der Vaart (2005), Cai and Low (2005).

The question is, whether one can bridge this gap between confidence sets and point estimators. More precisely, we would like to understand the possibility of adaptation for point estimators in terms of some confidence set or prediction region. To this end we consider the standard nested sequence of approximating models with candidate estimators

$$\hat{\theta}_n^{(k)} := (X_{n,1}, \ldots, X_{n,k}, 0, \ldots, 0)^\top$$

for $k \in \{0, 1, \ldots, n\}$. The risk of such an estimator is given by

$$R_n(k) := R(\hat{\theta}_n^{(k)}, \theta_n) = k\sigma^2 + \sum_{i>k} \theta_{n,i}^2,$$

while its loss equals

$$\begin{aligned} L_n(k) := L(\hat{\theta}_n^{(k)}, \theta_n) &= \sum_{i \leq k} (X_{n,i} - \theta_{n,i})^2 + \sum_{i>k} \theta_{n,i}^2 \\ &= R(k) + \sum_{i \leq k} (\epsilon_{n,i}^2 - \sigma^2), \quad \epsilon_n := X_n - \theta_n \,. \end{aligned}$$

Moreover, an unbiased estimator of risk is given by

$$\hat{R}_n(k) := k\sigma^2 + \sum_{i>k} (X_{n,i}^2 - \sigma^2).$$

Now our goal is to construct a $(1-\alpha)$–prediction region for the random set

$$S_n := \arg \min_{k \in \{0,1,\ldots,n\}} L_n(k).$$

For this purpose we consider the process $W_n = (W_n(k))_{k=0}^n$ with

$$W_n(k) := \hat{R}_n(k) - L_n(k).$$

Note that for $0 \leq j < k \leq n$,

$$W_n(k) - W_n(j) = 2 \sum_{i=j+1}^{k} (\epsilon_{n,i}^2 - \sigma^2 + \theta_{n,i} \epsilon_{n,i}).$$

Thus $W_n - W_n(0)$ is simply a random walk, but its increments are neither i.i.d. nor subgaussian. In fact,

$$W_n(k) - W_n(j) \;=_{\mathcal{L}}\; 2\sigma^2 \left(\chi^2_n(\delta^2_{n,j,k}) - (k-j) - \delta^2_{n,j,k}\right)$$

with noncentrality parameter $\delta^2_{n,j,k} := \sum_{i=j+1}^{k} \theta^2_{n,i}/(4\sigma^2)$. Nevertheless, by means of exponential inequalities for noncentral χ^2–distributions and recent results of Dümbgen and Walther (2005), one can show that the distribution of the random quantity

$$T_n \;:=\; \max_{0\le j<k\le n} \Big(\frac{|W_n(k) - W_n(j)|}{\gamma_n(j,k)} - \Gamma\Big(\frac{\gamma_n(0,n)^2}{\gamma_n(j,k)^2}, \gamma_n(j,k)\Big)$$

is asymptotically less than or equal (w.r.t. stochastic order) to the distribution of $T(W)$ introduced earlier. Here

$$\Gamma(u,\delta) \;:=\; \Gamma(u) + \frac{4\log(1/u)}{\delta^2},$$

while

$$\gamma_n(k,j) = \gamma_n(j,k) \;:=\; 2\sqrt{2|k-j| + \sum_{i=j+1}^{k} \theta^2_{n,i}/\sigma^2}$$

is the standard deviation of $(W_n(k) - W_n(j))/\sigma^2$.

In order to construct a confidence set by means of T_n, we are facing the problem that the numbers $\gamma_n(j,k)$ involve the unknown signal θ_n. Fortunately it suffices to consider the least favourable case when $|\theta_{n,i}| = \sigma$ for all i. In this case,

$$\gamma_n(j,k) \;=\; \sqrt{12|k-j|}.$$

Now let κ_n be the $(1-\alpha)$–quantile of T_n in this least favourable case, and let

$$\hat{S}_n \;:=\; \Big\{ j \in \{0,\ldots,n\} :\; \text{for all } k \in \{0,\ldots,n\},$$
$$\hat{R}_n(j) \le \hat{R}_n(k) + \sigma^2\sqrt{12|k-j|}\Big(\Gamma\big(\frac{|k-j|}{n}\big) + \frac{K\log(n/|k-j|)}{|k-j|} + \kappa_n\Big)\Big\}$$

for some universal constant K. Then for any fixed $c > 0$ and uniformly in $\theta_n \in B(0, c\sqrt{n})$,

$$\Pr\Big(S_n \subset \hat{S}_n\Big) \;\ge\; 1 - \alpha + o(1)$$

while

$$\max_{k\in\hat{S}_n} L_n(k) \;\le\; \Big(\min_{k\in\{0,1,\ldots,n\}} L_n(k) + \sigma^2\Big) O_p(\log n).$$

References

[1] Y. Baraud, *Confidence balls in gaussian regression*, Annals of Statistics **32** (2004), 528–551.

[2] R. Beran and L. Dümbgen, *Modulation of estimators and confidence sets*, Annals of Statistics **26** (1998), 1826–1856.

[3] T.T. Cai and M.G. Low, *Adaptive confidence balls*, Annals of Statistics, to appear.
[4] D.L. Donoho and I.M. Johnstone, *Ideal spatial adaptation by wavelet shrinkage*, Biometrika **81** (1994), 425–455.
[5] L. Dümbgen and V.G. Spokoiny, *Multiscale testing of qualitative hypotheses*, Annals of Statistics **29** (2001), 124–152.
[6] L. Dümbgen and G. Walther, *Multiscale inference about densities*, Preprint (2005).
[7] C.R. Genovese and L. Wassermann, *Confidence sets for nonparametric wavelet regression*, Annals of Statistics **33** (2005), 698–729.
[8] K.-C. Li, *Honest confidence sets for nonparametric regression*, Annals of Statistics **17** (1989), 1001–1008.
[9] B.T. Polyak and A.B. Tsybakov, *Asymptotic optimality of the C_p-test for the orthogonal series estimation of regression*, Theory of Probability and its Applications **35** (1991), 293–306.
[10] J. Robins and A. van der Vaart, *Adaptive nonparametric confidence sets*, Annals of Statistics (2005), to appear.

A survey on penalized empirical risk minimization

Sara A. van de Geer

We address the question how to choose the penalty in empirical risk minimization. Roughly speaking, this penalty should be a good bound for the estimation error. Main point is however that the estimation error depends on unknown parameters. We discuss a nonlocal estimate of the estimation error. Moreover, we show that the ℓ_1 penalty allows one to avoid explicitly estimating the estimation error.

The framework is as follows. Let the data $X_1, \ldots, X_n$ be i.i.d. copies of a random variable $X \in \mathbf{X}$ with distribution P. The empirical distribution is $P_n = \sum_{i=1}^n \delta_{X_i}/n$. We are interested in the parameter $f_0 \in \Lambda$, (Λ, d) being a metric space. This parameter f_0 is defined as the minimizer of the theoretical loss $R(f) := P\gamma_f$, $f \in \Lambda$, where $\gamma_f : \mathbf{X} \to \mathbf{R}$ is a given loss function. To estimate f_0, we replace $R(f)$ by its empirical counterpart $R_n(f) := P_n\gamma_f$. Next, we choose a model class $\mathbf{F} \subset \Lambda$, and define the penalized empirical risk minimizer

$$\hat{f}_n = \arg\min_{f \in \mathbf{F}} R_n(f).$$

Generally, it is necessary to choose a model class $\mathbf{F}$ which is strictly smaller than Λ. This is because Λ may be a very rich set, and empirical risk minimization over Λ may lead to overfitting the data.

Given the model class $\mathbf{F}$, the approximation error is defined as

$$B_n^2 = R(f_*) - R(f_0),$$

where

$$f_* = \arg\min_{f \in \mathbf{F}} R(f)$$

is the minimizer over the class $\mathbf{F}$. The estimation error is

$$V_n = R(\hat{f}_n) - R(f_*).$$

The excess risk of $\hat{f}_n$ is

$$R(\hat{f}_n) - R(f_0).$$

Thus we have a "bias-variance" type decomposition for the excess risk:

$$R(\hat{f}_n) - R(f_0) = B_n^2 + V_n.$$

Note that both the approximation error and the estimation error depend on $\mathbf{F}$. We express this by writing $B_n^2 = B_n^2(\mathbf{F})$ and $V_n = V_n(\mathbf{F})$. Consider now a collection of candidate models $\{\mathbf{F}\}$. The optimal model $\mathbf{F}_{\text{oracle}}$ is then the one which optimally trades off approximation error and estimation error, i.e.,

$$\mathbf{F}_{\text{oracle}} = \arg \min_{\mathbf{F} \in \{\mathbf{F}\}} \left\{ B_n^2(\mathbf{F}) + V_n(\mathbf{F}) \right\}.$$

Our aim is to find an estimator that mimics this trade off.

The following elementary lemma tells us that we can bound the estimation error by the empirical process ν_n, defined by $\nu_n(f) = \sqrt{n}(R_n(f) - R(f))$.

Elementary lemma 1. *Let $\hat{f}_n = \arg\min_{f \in \mathbf{F}} R_n(f)$ and $f_* = \arg\min_{f \in \mathbf{F}} R(f)$. Then we have the following bound for the estimation error $V_n := R(\hat{f}_n) - R(f_*)$:*

$$V_n \leq -[\nu_n(\hat{f}_n) - \nu_n(f_*)]/\sqrt{n}.$$

The next lemma indicates that in penalized empirical risk minimization, one should take the penalty, $\hat{\text{pen}}(\mathbf{F})$, equal to a good bound for the estimation error.

Elementary lemma 2. *Let $\hat{f}_n(\mathbf{F}) = \arg\min_{f \in \mathbf{F}} R_n(f)$ and*

$$\hat{\mathbf{F}}_n = \arg \min_{\{\mathbf{F}\}} \left\{ R_n(\hat{f}_n(\mathbf{F})) + \hat{\text{pen}}(\mathbf{F}) \right\}.$$

Fix some $\mathbf{F}_* \in \{\mathbf{F}\}$ and some $f_* \in \mathbf{F}_*$, and define the "approximation error" $B_n^2(\mathbf{F}_*) = R(f_*) - R(f_0)$ and "estimation error bound"

$$\text{V}_n(\mathbf{F}) = -[\nu_n(\hat{f}_n(\mathbf{F})) - \nu_n(f_*)]/\sqrt{n}. \tag{1}$$

Suppose that with probability at least $1 - \epsilon$, we have

$$\hat{\text{pen}}(\mathbf{F}) \geq \text{V}_n(\mathbf{F}), \ \forall \ \mathbf{F}.$$

Then with probability at least $1 - \epsilon$,

$$R(\hat{f}_n(\hat{\mathbf{F}}_n)) - R(f_0) \leq B_n^2(\mathbf{F}_*) + \hat{\text{pen}}(\mathbf{F}_*).$$

Concentration inequalities provide exponential probability inequalities for the concentration of the supremum of the empirical process around its mean (see e.g. (9)).

One may now derive a nonlocal bound for $V_n(\mathbf{F})$ defined in (1). Note first that for a non-random choice of f_*,

$$\mathbf{E}V_n(\mathbf{F}) = -\mathbf{E}\nu_n(\hat{f}_n)]/\sqrt{n} \leq \mathbf{E}\|R_n - R\|_{\mathbf{F}},$$

where we use the notation $\|\cdot\|_{\mathbf{F}}$ for the sup-norm of a class of functions on $\mathbf{F}$. Moreover,

$$\mathbf{E}\|R_n - R\|_{\mathbf{F}} \leq 2\mathbf{E}\|R_n^\sigma\|_{\mathbf{F}},$$

with $R_n^\sigma(f) = \sum_{i=1}^n \sigma_i \gamma_f(X_i)/n$ being the symmetrized version involving the Rademacher sequence $\{\sigma_i\}_{i=1}^n$. The latter is defined as a sequence of i.i.d. random variables, independent of $\{X_i\}_{i=1}^n$, with $\mathbf{P}(\sigma_i = 1) = \mathbf{P}(\sigma_i = -1) = 1/2$ $(i = 1, \ldots, n)$. Finally,

$$\mathbf{E}\|R_n^\sigma\|_{\mathbf{F}} = \mathbf{E}\mathbf{E}_{X_1,\ldots,X_n}\|R_n^\sigma\|_{\mathbf{F}},$$

where $\mathbf{E}_{X_1,\ldots,X_n}$ denotes conditional expectation given $X_1, \ldots, X_n$. Concentration inequalities (see (5)) now tell us (under conditions) that, with probability $1 - \epsilon$, up to a $n^{-1/2}$ term involving ϵ, $2\mathbf{E}_{X_1,\ldots,X_n}\|R_n^\sigma\|_{\mathbf{F}}$ is a bound for $V_n(\mathbf{F})$. If we use this bound, it is rather difficult to get rid of the $n^{-1/2}$ term and establish rates faster than $n^{-/12}$. The reason is that our estimate of the estimation error is a nonlocal one.

We will now illustrate that generally, the estimation error is smaller than $O(n^{-1/2})$. More details are e.g. in (3), (4), (5) and (8). We introduce the following two conditions, which both involve the same parameter $0 < \beta \leq 1$.

Margin condition. Let $G = \int_0^{\cdot} g(x)dx$, with g a strictly increasing function on the positive halfline, having $g(0) = 0$. Suppose

$$R(f) - R(f_0) \geq G(d^\beta(f, f_0)), \ \forall \ f \in \Lambda.$$

Empirical process condition. Let $f_* = \arg\min_{f\in\mathbf{F}} R(f)$. Suppose that for some positive constants d_n and C_n, we have with probability at least $1 - \epsilon$

$$\sup_{f\in\mathbf{F}} \frac{|\nu_n(f) - \nu_n(f_*)|}{d^\beta(f, \tilde{f}_*) + d_n^\beta} \leq C_n.$$

Lemma 3. *Assume the margin condition and the empirical process condition. Let* $\hat{f}_n = \arg\min_{f\in\mathbf{F}} R_n(f)$, *and* $B_n^2 = R(f_*) - R(f_0)$. *Let* $0 < \delta < 1$. *With probability at least* $1 - \epsilon$, *we have*

$$R(\hat{f}_n) - R(f_0) \leq \frac{1+\delta}{1-\delta}\{B_n^2 + \mathbf{V}_n + n^{-1/2} d_n^\beta C_n\},$$

where

$$\mathbf{V}_n = 2\delta H(\frac{C_n}{\delta\sqrt{n}}),$$

and $H = \int_0^{\cdot} g^{-1}(x)dx$.

As a typical example, suppose we have $\beta = 1$ and that g is the identity. Then $G(x) = H(x) = x^2/2$, and we find

$$\mathbf{V}_n = \frac{C_n^2}{n\delta}.$$

the constant C_n^2 is typically something like "dimension" or a more general measure of "complexity" of $\mathbf{F}$. If it does not grow too fast in n, and if in addition d_n decreases fast in n, we indeed arrive at estimation error of order smaller than $n^{-1/2}$.

It will be clear however that in general it is not obvious to verify the conditions, as they depend on the underlying distribution. In particular, it is often not clear what the function g is the margin condition. Thus, we do not know how large $\mathbf{V}_n$ is. However, as is shown in literature (see for example (1), (2), (5), (6), (7), (11)), there are ways to obtain a good local estimate.

We now turn to ℓ_1 penalization, to avoid the problem of unknown margin behavior. Let $\gamma_f = \gamma \circ f$, and suppose γ is convex, and Lipschitz with Lipschitz constant 1. Suppose $\Lambda \subset L_2(\nu)$, with ν some measure on $\mathbf{X}$. Let $\mathbf{F}_m$ be a convex subset of $\{f_\alpha = \sum_{k=1}^m \alpha_k \psi_k\}$, where $\{\psi_k\}_{k=1}^m \subset L_2(\nu)$ are given base functions. We assume that $m \leq n^D$ for some $D \geq 1$. Also, we assume

$$\max_{k=1,\ldots,m} \|\psi_k\|_\infty \leq \sqrt{\frac{n}{\log n}}.$$

We consider the estimator

$$\hat{f}_n = \arg \min_{f_\alpha \in \mathbf{F}_m} \{R_n(f_\alpha) + \hat{\lambda}_n \sum_{k=1}^m |\alpha_k|\}.$$

Here, we take

$$\hat{\lambda}_n \geq 864 \hat{\Psi}_n D \sqrt{\frac{\log n}{n}},$$

with

$$\hat{\Psi}_n^2 = \max_{k=1,\ldots,n} P_n \psi_k^2 \vee 4^2.$$

We let

$$\Psi_0^2 = \max_{k=1,\ldots,m} P \psi_k^2 \vee 4^2,$$

and let λ_n be the theoretical counterpart of the smoothing parameter $\hat{\lambda}_n$, i.e.

$$\lambda_n = \hat{\lambda}_n \frac{\Psi_0}{\hat{\Psi}_n}.$$

Now, our further conditions depend on the unknown underlying distribution, so we call them non-verifiable conditions. Note however that our estimation procedure does not require them to be verifiable.

Non-verifiable conditions.

- The margin condition holds.
- It holds that $\|f - \tilde{f}\|_{2,\nu} \leq d^\beta(f, \tilde{f})$ for all $f, \tilde{f} \in \mathbf{F}_m$. Here β is from the margin condition, and $\|\cdot\|_{2,\nu}$ denotes the $L_2(\nu)$-norm.

• It holds that $\|f - \tilde{f}\|_\infty \leq K_n d(f, \tilde{f}) \vee 2$ for all $f, \tilde{f} \in \mathbf{F}$. Here K_n is a sequence satisfying a growth condition (see Theorem 4).
• For some diagonal matrix $W = \mathrm{diag}(w_1, \ldots, w_m)$ of positive weights, the matrix $W\Sigma_\nu W$ has smallest eigenvalue equal to one. Here $\Sigma_\nu = \int \psi\psi^T d\nu$ with $\psi = (\psi_1, \ldots, \psi_m)^T$.

We now define the "estimation error bound" as

$$\mathbf{V}_n(\alpha) = 2\delta H(18\lambda_n C(\alpha)/\delta),$$

with $H = \int_0^{\cdot} g^{-1}(x)dx$, and with

$$C^2(\alpha) = D \sum_{k:\alpha_k \neq 0} w_k^2.$$

Let

$$\epsilon_n = \frac{1+\delta}{1-\delta} \min_{f_\alpha \in \mathbf{F}} \left\{ R(f_\alpha) - R(f_0) + \mathbf{V}_n(\alpha) + 2\lambda_n \sqrt{\frac{\log n}{n}} \right\}.$$

The following theorem is a generalization of the result in (10).

Theorem 4. *Consider the estimator*

$$\hat{f}_n = \arg \min_{f_\alpha \in \mathbf{F}_m} \{R_n(f_\alpha) + \hat{\lambda}_n \sum_{k=1}^m |\alpha_k|\}.$$

Assume the non-verifiable conditions with growth rate condition $K_n^\beta G^{-1}(\epsilon_n) \leq 1$. Then there is a universal constant c, such that with probability at least $1 - c/n^2$, we have

$$R(\hat{f}_n) - R(f_0) \leq \epsilon_n.$$

References

[1] J.-Y. Audibert, *Classification under polynomial entropy and margin assumptions and randomized estimators* , Preprint, Laboratoire de Probabilités et Modèles Aléatoires (2004).

[2] P.L. Bartlett, O. Bousquet and S. Mendelson, *Local Rademacher complexities*, Ann. Statist. **33** (2005), 1497–1537.

[3] G. Blanchard, G. Lugosi and N. Vayatis, *On the rate of convergence of regularized boosting classifiers*, J. Machine L. Research **4** (2003), 861–894.

[4] G. Blanchard, O. Bousquet and P. Massart, *Statistical performance of support vector machines*, Manuscript (2004).

[5] O. Bousquet, S. Boucheron and G. Lugosi, *Theory of classification: a survey of recent advances*, (2005). To appear in ESAIM: Probability and Statistics.

[6] V. Koltchinskii, *Local Rademacher complexities and oracle inequalities in risk minimization* (2003). To appear in Ann. Statist.

[7] G. Lugosi and M. Wegkamp, *Complexity regularization via localized random penalties*, Ann. Statist. **32** (2004), 1679–1697.

[8] P. Massart, *Some applications of concentration inequalities to statistics*, Annales de la Faculté de Toulouse **9** (2000), 245–303.
[9] P. Massart, *About the constants in Talagrand's concentration inequalities for empirical processes*, Ann. Probab. **28** (2000), 863–884.
[10] B. Tarigan and S.A. van de Geer, *Classifiers of support vector machine type, with ℓ_1 complexity regularization*, submitted (2005).
[11] A.B. Tsybakov, *Optimal aggregation of classifiers in statistical learning*, Ann. Statist. **32** (2004), 1203–1224.

Bayesian Constructions for the General Regression Problem

Edward I. George, Robert E. McCulloch
(joint work with Hugh Chipman)

We describe implementations of the Bayesian approach for model uncertainty problems where a large number of different models are under consideration for data, (1) (2), (3). A joint distribution is obtained by introducing prior distributions on all the unknowns, here the parameters of each model and the models themselves, and then combining them with the likelihood. Conditioning on the data then induces a posterior distribution of model uncertainty that can be used for model selection and other inference and decision problems. After laying out the main details of the general Bayesian approach for model uncertainty, we focus on the construction of methods for addressing three central regression problems where one wants to model the relationship between between a variable of interest y, and a set of potential predictor variables $x_1, \ldots, x_p$.

We first consider the variable selection problem for the linear model which arises when one wants to model the linear relationship between y and a subset of $x_1, \ldots, x_p$, but there is uncertainty about which subset to use, (4). By embedding this setup in a hierarchical mixture model on the regression coefficients, promising subset models are identified with high posterior models, (5), (6). We next turn to regression tree modeling where the goal is to select a binary tree that partitions the predictor space into regions where the distribution of y is homogeneous. By describing tree uncertainty with a tree generating stochastic process, a posterior distribution is obtained on the set of trees, (7). Posterior computation and exploration in both of these problems can be obtained by MCMC (Markov chain Monte Carlo) simulation, (8).

Combining and extending these formulations, we propose BART (Bayesian Additive Regression Trees), a new approach to discover the form of $f(x_1, \ldots, x_p) \equiv E(Y \mid x_1, \ldots, x_p)$ and draw inference about it, (9). BART approximates f by a Bayesian "sum-of-trees" model where each tree is constrained by a prior to be a weak learner as in boosting. Fitting and inference are accomplished via an iterative backfitting MCMC algorithm. By using a large number of trees, which yields an overcomplete basis for f, we have found BART to be remarkably effective at

finding highly nonlinear relationships hidden within a large number of irrelevant potential predictors.

BART is motivated by ensemble methods in general, and boosting algorithms in particular. Like boosting, each weak learner (i.e., each weak tree) contributes a small amount to the overall model, and the training of a weak learner is conditional on the estimates for the other weak learners. The differences from boosting algorithms are just as striking as the similarities: BART is defined by a statistical model: a prior and a likelihood, while boosting is defined by an algorithm. MCMC is used both to fit the model and to quantify inferential uncertainty through the variation of the posterior draws.

The BART modelling strategy can also be viewed in the context of Bayesian non-parametrics. The key idea is to use a model which is rich enough to respond to a variety of signal types, but constrained by the prior from overreacting to weak signals. The ensemble approach provides for a rich base model form which can expand as needed via the MCMC mechanism. The priors are formulated so as to be interpretable, relatively easy to specify, and provide results that are stable across a wide range of prior hyperparameter values. The MCMC algorithm, which exhibits fast burn-in and good mixing, can be readily used for model averaging and for uncertainty assessment.

After introducing BART, we proceed to illustrate how it opens up a new approach to variable selection when one wants to model the relationship between y and a subset of $x_1, \ldots, x_p$, but there is uncertainty about which subset to use. This selection problem is typically treated by assuming that the relationship between y and $x_1, \ldots, x_p$ belongs to a parametric family such as the normal linear models. If incorrect, however, such an assumption can at the outset defeat the ultimate goal; subsets of $x_1, \ldots, x_p$ may be excluded simply because their relationship to y is far outside the assumed parametric family. To avoid this limitation, we show how BART may be used to discover the nature of the relationship between y and $x_1, \ldots, x_p$ before attempting to find relevant variables and a suitable parametric form.

To begin with, BART automatically screens for relevant predictors. As the BART algorithm moves through the model space, different potential predictors enter the model with different frequencies. Those that enter rarely or not at all are candidates for elimination, and those that enter frequently are candidates for inclusion. Based on such information, we consider various strategies for rerunning BART on subsets of $x_1, \ldots, x_p$ which lead to a stable subset for selection. Note that BART also provides an omnibus test: the absence of any relationship between y and any subset of $x_1, \ldots, x_p$ is suggested when BART posterior intervals for f reveal no signal.

Going further, let $\hat{f}$ be a BART estimate of f based on the selected subset of $x_1, \ldots, x_p$. Intuitively, $\hat{f}$ may be regarded as a sufficient statistical summary of the systematic relationship between y and $x_1, \ldots, x_p$. Thus $\hat{f}$ and the selected subset can be used, instead of the raw data, to find a parametric model for this relationship. For example, let $M_1, \ldots, M_m$ be m different parametric model classes

under consideration such as the normal linear models or other exponential family models. Partial dependence plots applied to $\hat{f}$ may be useful for suggesting the form of such model classes as well as useful transformations of the predictors. Basically, the goal is to find the model within any of these model classes that is "best supported" by $\hat{f}$. For this purpose, we consider the strategy of selecting the model corresponding to the projection of $\hat{f}$ onto the nearest model class with respect to a utility criterion such as the Kullback-Leibler discrepancy. Yet another strategy is to construct a likelihood over the model space based on the probability distribution of $\hat{f}$ for each model. This opens the door to $\hat{f}$ based Bayesian approaches for model selection and averaging over $M_1, \ldots, M_m$.

References

[1] E.I. George, *Bayesian model selection*, Encyclopedia of Statistical Sciences (eds. S. Kotz, C. Read and D. Banks), Wiley, N.Y., Update Volume **3** (1998), 39–46.

[2] H. Chipman, E.I. George and R.E. McCulloch. *The Practical Implementation of Bayesian Model Selection*, Model Selection, (P. Lahiri, ed.) IMS Lecture Notes – Monograph Series Volume **38** (2001), 65–134.

[3] M. Clyde and E.I. George, *Model Uncertainty*, Statistical Science **19 1** (2004), 81–94.

[4] E.I. George, *The Variable Selection Problem*, Journal of the American Statistical Association **95** (2000), 1304-1307.

[5] E.I. George and R.E. McCulloch, *Variable selection via Gibbs sampling*, Journal of the American Statistical Society **88** (1993), 881–889.

[6] E.I. George and R.E. McCulloch, *Approaches for Bayesian variable selection*, Statist. Sinica **7** (1997), 339-373.

[7] H.A. Chipman, E.I. George and R.E. McCulloch, *Bayesian CART model search (with discussion)*, Journal of the American Statistical Association **93** (1998), 935–960.

[8] W.R. Gilks, S. Richardson and D.J. Spiegelhalter, *Markov Chain Monte Carlo in Practice*, London: Chapman & Hall (1996).

[9] H. Chipman, E.I. George and R.E. McCulloch, *BART: Bayesian Additive Regression Trees* (2005), (submitted).

Different Roles of Penalties in Penalized Likelihood Model Selection Rules

Jayanta Ghosh

Introduction : We argue that each penalty has a specific role in the sense of addressing a specific problem and attaining optimality there. We illustrate our thesis through examples and results from Bayesian Analysis, Parametric Empirical Bayes Analysis (PEB), Machine Leaning and Classical Statistics. In particular, we

examine in detail the performance of AIC. We also examine BIC to some extent since the AIC and BIC remain the most popular as well as the most misused.

We begin with Bayesian model selection because basic issues are clearest in this set up. Bayesian model selection rules do not involve penalties but it is clear that different objectives as measured by different loss functions lead to different rules.

If the object is to select a correct model the appropriate loss is 0-1. If the object is to make prediction, squared error seems appropriate. Moreover, the nature of the model space is relevant for choice of priors - a uniform prior for nested models and the Binomial prior $\pi(q) = w^p(1-w)^{p-q}$ for all subsets model selection. Here, $0 < w < 1$, $q =$ dimension of the model under consideration, $p =$ dimension of the most complex model.

For a 0-1 loss the Bayes rule selects the posterior mode, i.e., the model with highest posterior probability. On the other hand, for squared error, the Bayes rule is to select the posterior median, cf. Barbieri and Berger (2004). The remarks below address this dichotomy from different points of view.

First, having two optimal rules is quite natural and serves two different basic purposes. The posterior mode is part of our description of the truth whereas the posterior median is part of our description of the action we ought to take. So we need two optimally selected models to satisfy two different needs.

Secondly, it is hard to believe that the 0-1 loss is a good choice when the cardinality of the model space is large and some models differ by relatively few parameters. It would be easier to reconcile two optimal models if 0-1 loss is changed to reflect better the topology of the model space or one changes the first model selection problem. For example, have a credibility set for the model multi-index. The median model index would lie in this subset.

Model Selection in Nested PEB Regression with orthogonal design : In a nested orthogonal version of the problem formulated in George and Foster (2000), Ghosh and Mukhopadhyay (2003) show through asymptotics and extensive simulation that an optimal Bayes rule for one loss function will perform poorly when the loss is changed. In particular it is shown that the AIC is asymptotically optimal in the sense of attaining an oracle for prediction loss and that AIC does very well in prediction, in some cases doing substantially better that the (PEB) posterior median and the (PEB) posterior mode. On the other hand this very fact is related to its poor performance in situations involving 0-1 loss and in its failure to be consistent for all values of hyperparameters.

More about AIC and BIC : Schwarz (1978) derived the BIC essentially as an approximation to the Bayes rule for 0-1 loss in low dimensional problems whereas AIC was proposed by Akaike (1973) as an appropriate rule in high dimensional prediction problems when the true model is so complex that it is not in the model space.

Thus AIC and BIC solve quite different problems. The penalty of BIC arises from the 0-1 loss and the prior while the penalty of AIC is such that the difference

of AIC for two models is an unbiased estimate of the difference of the two prediction risks. One may say the penalty of AIC arises from the squared error loss.

In view of this it is not surprising that performance of AIC and BIC can vary from "pretty good" to "pretty bad" depending on the context. In Chakrabarti and Ghosh (2005a) it is shown that AIC attains an oracle and the minimax rate for nonparametric regression. It also seems to do better than the Cai-Low-Zhao shrinkage estimate for smooth functions. BIC fails for both squared error and 0-1 loss in high dimensional problems (cf. Berger (2005), Berger, Ghosh and Mukhopadhyay (2003), Chakrabarti and Ghosh (2005a,b), Ghosh and Mukhopadhyay (2003), Mukhopadhyay (2000)). Here we follow Vapnik's tentative definition of high dimension as meaning $n \asymp p$.

Machine Leaning and Classical Statistics : Vapnik (1995) has suggested the principle of structured risk minimization. This may be interpreted as penalized empirical risk minimization with the penalty arising from application of the Cervonenkis-Vapnik Uniform Strong Law of Large Numbers. Vapnik obtains a tight oracle-like upper bound that proves the optimality of his principle. A more recent similar example is Lugosi and Vayatis (2004). An interesting new example was presented at the conference by Kolchinski in the context of a modification of LASSO.

In classical statistics Fan *et al.* (2001, 2002) discuss three desirable properties for a penalty and (almost) necessary and sufficient conditions for these properties. They choose their non-concave penalty on the basis of these considerations and prove oracle like properties as well indicate advantages over LASSO and Hard Thresholding.

Given all these examples from different paradigms, it seems one may now look for a Grand Unifying Theory.

References

[1] H. Akaike, *Information theory as an extension of the maximum likelihood principle*, Second International Symposium on Information Theory, B. N. Petrov and F. Csaksi editors (1973) 267–281, Akademiai Kiado, Budapest, Hungary.

[2] M. Barbieri and J. Berger, *Optimal predictive model selection*, The Annals of Statistics **32** (2004), 870–897.

[3] J. Berger, *Generalization of BIC*, Personal communications (2005).

[4] J. Berger, J.K . Ghosh and N. Mukhopadhyay, *Approximations to the Bayes factor in model selection problems and consistency issues*, Journal of Statistical Planning and Inference **112** (2003), 241–258.

[5] A. Chakrabarti and J.K. Ghosh, *Optimality of AIC in Inference About Brownian Motion*, The Annals of Institute of Statistical Mathematics **To appear** (2005a).

[6] A. Chakrabarti and J.K. Ghosh, *A Generalization of BIC for the General Exponential Family*, Journal of Statistical Planning and Inference **To appear** (2005b).
[7] J. Fan and R.Z. Li, *Variable selection via penalized likelihood*, Journal of American Statistical Association **96** (2001), 1348–1360.
[8] J. Fan and R.Z. Li, *Variable Selection for Cox's Proportional Hazards Model and Frailty Model*, The Annals of Statistics **30** (2002), 74–99.
[9] E. George and D. Foster, *Calibration and empirical Bayes variable selection*, Biometrika **87** (2000), 731–747.
[10] J.K. Ghosh and N. Mukhopadhyay, *Parametric empirical Bayes model selection - some theory, methods and simulation*, IMS Lecture Notes in honor of Rabi Bhattacharya, Athreya *et al.* editors (2003).
[11] G. Lugosi and N. Vayatis, *On the Bayes-risk consistency of regularized boosting methods*, The Annals of Statistics **32** (2004), 30–55.
[12] N. Mukhopadhyay, *Bayesian Model Selection for High Dimensional Models with Prediction Error Loss and 0-1 Loss*, Thesis submitted to Purdue University (2000).
[13] G. Schwarz, *Estimating the dimension of a Model*, The Annals of Statistics **6** (1978), 461–464.
[14] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer Verlag (1995).

Testing and model selection for density estimation

László Györfi
(joint work with Gérard Biau, Benoît Cadre, Luc Devroye)

Consider two mutually independent samples of $\mathbb{R}^d$-valued random vectors $X_1, \ldots, X_n$ and $X'_1, \ldots, X'_n$ with *i.i.d.* components defined on the same probability space and distributed according to unknown probability measures μ and μ'. We are interested in testing the null hypothesis that the two samples are homogeneous, that is

$$\mathcal{H}_0 : \mu = \mu'.$$

Denote by μ_n and μ'_n the empirical measures associated with the samples $X_1, \ldots, X_n$ and $X'_1, \ldots, X'_n$, respectively, so that

$$\mu_n(A) = \frac{\#\{i : X_i \in A, i = 1, \ldots, n\}}{n}$$

for any Borel subset A, and, similarly,

$$\mu'_n(A) = \frac{\#\{i : X'_i \in A, i = 1, \ldots, n\}}{n}.$$

Based on a finite partition $\mathcal{P}_n = \{A_{n1}, \ldots, A_{nm_n}\}$ of $\mathbb{R}^d$ ($m_n \in \mathbb{N}^*$), we let the test statistic comparing μ_n and μ'_n be defined as

$$T_n = \sum_{j=1}^{m_n} |\mu_n(A_{nj}) - \mu'_n(A_{nj})|.$$

Theorem 1. *(Biau, Györfi (1)). Assume that*

$$\lim_{n\to\infty} m_n = \infty, \qquad \lim_{n\to\infty} \frac{m_n}{n} = 0,$$

and

$$\lim_{n\to\infty} \max_{j=1,\ldots,m_n} \mu(A_{nj}) = 0.$$

Then, under $\mathcal{H}_0$, for all $0 < \varepsilon < 2$,

$$\lim_{n\to\infty} \frac{1}{n} \ln \mathbf{P}\{T_n > \varepsilon\} = -g_T(\varepsilon),$$

where

$$g_T(\varepsilon) = (1+\varepsilon/2)\ln(1+\varepsilon/2) + (1-\varepsilon/2)\ln(1-\varepsilon/2).$$

Consider a model selection problem for density estimation. Let $(\mathcal{F}_k)_{k\geq 1}$ be a sequence of nested parametric models of density functions on $\mathbb{R}^d$. Define

$$\mathcal{F} = \bigcup_{k\geq 1} \mathcal{F}_k\,.$$

In the union above, $\mathcal{F}_k$ denotes, for each fixed $k \geq 1$, a given class of densities such that $\mathcal{F}_k \subset \mathcal{F}_{k+1}$ for all k. The general problem is to estimate a density f which belongs to $\mathcal{F}$. Formally, we let the complexity associated with f be defined as

$$k^* = \min\{k \geq 1 : f \in \mathcal{F}_k\}.$$

Clearly, as it is supposed that $f \in \mathcal{F}$, we have $k^* < \infty$.

We assume that the sample of independent random vectors distributed according to the probability measure μ with density f is of even size $2n$. Let μ_{2n} be its empirical measure, *i.e.*, $\mu_{2n}(A) = \big(1/(2n)\big)\sum_{i=1}^{2n} \mathbf{1}_{\{X_i \in A\}}$. Split the sample into two subsamples: $X_1, \ldots, X_n$ and $\{X'_1, \ldots, X'_n\} = \{X_{n+1}, \ldots, X_{2n}\}$, and denote by μ_n and μ'_n the respective empirical measures. Let $\mathcal{P}_n = \{A_{n,j}, j \geq 1\}$ be a cubic partition of $\mathbb{R}^d$ with volume h_n^d. Introduce the statistic

$$d_{n,k} = \inf_{g\in\mathcal{F}_k} \sum_{A\in\mathcal{P}_n} \left| \int_A g - \mu_{2n}(A) \right|.$$

Let the threshold be

$$T_n = \sum_{A\in\mathcal{P}_n} |\mu_n(A) - \mu'_n(A)|.$$

Our estimate of k^* is

$$K_n = \min\{k \geq 1 \,:\, d_{n,k} \leq T_n\},$$

with the convention $\min\{\emptyset\} = 1$.

Theorem 2. *(Biau, Cadre, Devroye, Györfi (2)) Assume that, for each $k \geq 1$, the set of Fourier transforms of the densities in $\mathcal{F}_k$ is closed with respect to the pointwise convergence. Suppose that $\lim_{n\to\infty} h_n = 0$ and $\lim_{n\to\infty} nh_n^d = \infty$. Then there exists a constant $\kappa > 0$, depending only on f, such that, for all $n \geq 1$,*

$$\mathbf{P}\left\{K_n \neq k^*\right\} \leq 8 \exp\left(-\kappa\, h_n^{-d}\right).$$

In particular, if $\kappa h_n^{-d} \geq (1+\delta)\log n$ for some $\delta > 0$, then

$$\lim_{n\to\infty} K_n = k^*$$

almost surely.

Based on the strong consistent model selection K_n, we apply a minimum distance density estimation. Introduce the class of sets

$$\mathcal{A}_k = \left\{\{x : g_1(x) \geq g_2(x)\} : g_1, g_2 \in \mathcal{F}_k\right\}$$

($\mathcal{A}_k$ is the so-called Yatracos class associated with $\mathcal{F}_k$) and the goodness criterion for a density $g \in \mathcal{F}_k$:

$$\Delta_k(g) = \sup_{A\in\mathcal{A}_k} \left| \int_A g - \mu_{2n}(A) \right|.$$

Then the minimum distance estimate $\hat{f}_k$ is defined as any density estimate selected from among those densities $f_k \in \mathcal{F}_k$ with

$$\Delta_k(f_k) < \inf_{g\in\mathcal{F}_k} \Delta_k(g) + \frac{1}{n}.$$

For the model selection K_n, our density estimate is $\hat{f}_{K_n}$.

Theorem 3. *(Biau, Cadre, Devroye, Györfi (2)) Assume that, for each $k \geq 1$, the set of Fourier transforms of the densities in $\mathcal{F}_k$ is closed with respect to the pointwise convergence. Assume that the Vapnik-Chervonenkis dimension of $\mathcal{A}_{k^*}$ is finite. Suppose that $\lim_{n\to\infty} nh_n^d = \infty$ and $\kappa h_n^{-d} \geq (1/2)\log n$, where κ is defined in Theorem 2. Then the minimum distance estimate $\hat{f}_{K_n}$ satisfies*

$$\mathbf{E}\left\{ \int |\hat{f}_{K_n} - f| \right\} = \mathrm{O}\left(\frac{1}{\sqrt{n}}\right).$$

References

[1] G. Biau, L. Györfi, *On the asymptotic properties of a nonparametric L_1-test of homogeneity*, IEEE Trans. Information Theory **51** (2005), 3965–3973.

[2] G. Biau, B. Cadre, L. Devroye, L. Györfi, *Strongly consistent model selection for densities*, (submitted for publications).

Focussed information criteria and model averaging

Nils Lid Hjort
(joint work with Gerda Claeskens)

1. Background, examples and questions

There is a great variety of model selection criteria and model average techniques, coming from different brands of motivation and traditions of thought. Often the underlying aims of methods, whether to be understood implicitly or stated explicitly, also differ. To introduce the themes concentrated on in my survey talk I start with a list of examples, or situations.

Example A. Suppose independent observations $Y_1, \ldots, Y_n$ stem from the parametric family with cumulative distribution function

$$F(y) = \Phi((y-\xi)/\sigma)^\gamma,$$

where Φ is the standard normal cumulative. Assume one wishes to estimate e.g. the upper quartile

$$\mu = \xi + \sigma\Phi^{-1}((3/4)^{1/\gamma}).$$

Using the simple normal model corresponds to $\gamma = 1$, with consequent estimate $\widehat{\mu}_{\rm narr} = \widehat{\xi}_{\rm narr} + 0.675\,\widehat{\sigma}_{\rm narr}$, in terms of the familiar maximum likelihood estimators of the $\mathrm{N}(\xi, \sigma^2)$ model. Using the wider three-parameter model corresponds on the other hand to using $\widehat{\mu}_{\rm wide} = \widehat{\xi} + \widehat{\sigma}\Phi^{-1}((3/4)^{1/\widehat{\gamma}})$, involving the maximum likelihood estimators in that wider model. Which estimator is best?

Example B. Consider survival regression data of the usual form (t_i, x_i, δ_i), where t_i is the possibly censored survival time for an individual with covariate vector $x_i = (x_{i,1}, \ldots, x_{i,p})^{\rm t}$ and δ_i is an indicator for non-censoring. The proportional hazards regression model holds that the hazard rates for individuals $i = 1, \ldots, n$ take the form

$$\alpha_i(s) = \alpha_0(s)\exp(x_i^{\rm t}\beta).$$

Assume one wishes to estimate the median remaining survival time $\mu = \mu(x_0, t_0)$ for a given patient who has already survived up to t_0 and who has covariates x_0; one finds

$$\mu = A_0^{-1}\big(A_0(t_0) + (\log 2)/\exp(x_0^{\rm t}\beta)\big),$$

in terms of the cumulative hazard rate A_0 for α_0. Which among the p covariates ought to be included in the model?

Example C. I have collected data on 190 football matches from four grand occasions: the $31+31$ matches from the European Championships 2004 and 2000 and the $64+64$ matches from the World Championships 2002 and 1998. In addition to the match results I have recorded the official FIFA scores for each team, as of two weeks prior to the championships in question. Various statistical models may be put up to model probabilistically the result (y, y') of a football match between opponents with FIFA scores (x, x'). If I wish to estimate the probability

$\mu = \mu(x, x')$ that Norway beats Belgium in tomorrow's game, which of the candidate models should I employ?

EXAMPLE D. Consider temperature time series data $y_1, \ldots, y_n$ at Oberwolfach, and suppose one needs to estimate the probability that tomorrow's temperature y_{n+1} falls below zero. Again, which of the many time series models should be used?

Natural questions emerging from these and similar situations and examples would include the following:

(i) For cases where a narrow model (with say p parameters) is extended to a suitable richer model (with say $p+q$ parameters), how much can the simpler model tolerate, in order for the narrow based estimators to still be more precise than the wide model based estimators?

(ii) Each of the examples have a well-defined 'focus parameter' μ. For different estimators $\widehat{\mu}_S$, with S indexing which of the full set of parameters to include in the model, what is the (approximate) mean squared error, or somewhat more ambitiously the (approximate) distribution? Answers to this and similar questions ought to depend on the location in the parameter space as well as on the sample size n.

(iii) Which of the various $\widehat{\mu}_S$ will actually be best, as measured e.g. by mean squared error?

(iv) How well can we estimate these theoretical mean squared errors, and can we use such estimates to select the tentatively best estimator?

(v) For model selection methods like the AIC, what are the model choice probabilities, say $p_n(S)$ for the different submodels indexed by S?

(vi) Can anything be won by averaging estimates across submodels, perhaps using data-dependent weights, compared to using the best submodel?

(vii) In statistical practice, one tends to gloss over the elaborations of the model selection step, and to more simply report both estimates, confidence intervals and p-values computed under the finally selected model. This 'hides uncertainty', and yields too optimistic confidence intervals and p-values. What would be the real coverage probability of a confidence interval with claimed level 95%, if that interval has been constructed using say the AIC or the FIC strategy to arrive at a model?

(viii) Are there ways of 'repairing' the implicit over-optimism in confidence and p-value statements alluded to in the previous point?

(ix) The last decade has seen several hundred technical and applied papers on 'Bayesian model averaging', but these have mainly focussed on computation, algorithms, specification of models and priors, and interpretation. Very few have dealt with the actual distributional aspects of the model average estimators. What are the limit distributions involved in Bayesian model average procedures?

(x) How does standard theory about e.g. the AIC and BIC stand up to challenges that involve an increasing number of parameters as a function of sample size?

2. **Discussion**

My survey talk did perhaps not manage to cover the full convex hull spanned by these ten questions, but did provide methods and results of relevance for each of them. The talk was primarily based on Hjort and Claeskens (2003a,b, 2006) and Claeskens and Hjort (2003), but did also touch on work of a different nature from Hjort, Dahl and Steinbakk (2006).

I would stress that a common thread in our approaches is 'the focus'; we carefully distance ourselves from the perhaps too commonly found viewpoint, whether made implicitly or explicitly, that 'one good model' should be selected to cater for all needs of analysis, interpretation and prediction. In our work, three different foci parameters μ_1, μ_2, μ_3, corresponding to three different questions of interest, might call for three different models. Thus good statistical decisions in a complicated regression setup might depend on where one is in the covariate space, so to speak. In a study of survival of Danish melanoma patients we found that 'the best model', as measured by accuracy of the median remaining survival time as in Example B, is different for men and for women. We do not view this as a paradox. Similarly, a good model for understanding and analysing the mean structure might not be a good model for understanding and analysing the variance of skewness structure in a model for life quality responses to socio-economic background factors. And for the football prediction game of Example C, I find different optimal models for different prospective matches: one could be good for estimating the Norway beats Belgium probability, but not as good as another one for estimating the winning odds for Germany against Brazil.

As Longford (2005) opines in his Editorial, the 'which model?' question might not always be the right question; and any answer needs to take on board aspects not only of 'for what purpose?' but also the component 'followed by which analyses and techniques?'. This might appear obvious to most statisticians, when phrased in such ways, but one needs to realise that big chunks of even the modern statistical literature have been concerned with properties of the more 'automatic' overall selection criteria like the AIC and the BIC.

Inside a precise local asymptotic framework, covering in essence all regular parametric families, we have reached relevant answers to each of the ten questions posed above. For this framework, see Hjort and Claeskens (2003a), Claeskens and Hjort (2003) and the ensuing discussion contributions and rejoinder. The model bias inherent in choosing amongst models that are not entirely perfect is explicitly taken into account. Our techniques involve limit distributions of submodel-based estimators as well as of convex data-based averages of such.

References

[1] G. Claeskens and N.L. Hjort, *The focussed information criterion* [with discussion], Journal of the American Statistical Association **98** (2003), 900–916.

[2] N.L. Hjort and G. Claeskens, *Frequentist model averaging estimators* [with discussion], Journal of the American Statistical Association **98** (2003a), 879–899.

[3] N.L. Hjort and G. Claeskens, *Rejoinder to the discussion of the FIC and FMA articles*, Journal of the American Statistical Association **98** (2003b), 938–945.
[4] N.L. Hjort and G. Claeskens, *Focussed information criteria and model averaging for Cox's hazard regression model*, Journal of the American Statistical Association **101** (2006), to appear.
[5] N.L. Hjort, F.A. Dahl and G.H. Steinbakk, *Post-processing posterior predictive p-values*, Journal of the American Statistical Association **101** (2006), to appear.
[6] N.T. Longford, *Editorial: Model selection and efficiency – is 'Which model...?' the right question?*, Journal of the Royal Statistical Society A **168** (2005), 469–472.

A Note on the Consistency and Interpretation of Bayes Factors Based on Test Statistics

Valen E. Johnson

A method for defining Bayes factors based on the sampling distributions of test statistics was proposed in Johnson (2005). Although distributions of test statistics are completely specified under a null model, the use of my method also requires specification of the distribution of test statistics under alternative hypotheses. For Bayes factors based on χ^2 and F statistics, these distributions can be naturally defined as noncentral versions of the null distribution. In this talk, I describe criteria for setting hyperparameters that determine these noncentral distributions so that the resulting Bayes factors are consistent.

Let $BF(1|2)$ denote the Bayes factors between models 1 and 2, i.e. the ratio of the marginal density of the data under model 1 to the marginal density of the data under model 2. Then $BF(1|2)$ is *consistent* if (a) $BF(1|2) \xrightarrow{p} \infty$ as the sample size $n \to \infty$ when model 1 is true, and (b) $BF(1|2) \xrightarrow{p} 0$ as $n \to \infty$ when model 2 is true.

For the remainder of this note, 'J5" refers to Johnson (2005) and, unless otherwise stated, notation and regularity conditions stated in J5 apply here also.

χ^2 tests for multinomial data. Let $\mathbf{p}$ denote a multinomial probability vector which satisfies a given null hypothesis, and suppose that under the alternative hypothesis the multinomial probability vector $\mathbf{q}$ is drawn from a Dirichlet distribution with parameter $c\,\mathbf{p}$. Letting $K - s - 1$ denote the degrees of freedom of the χ^2 statistic x_n (as defined in Sec. 2 of J5), the logarithm of the Bayes factor between the alternative hypothesis (model 2) and null hypothesis (model 1) may be written

$$\left[\frac{nx_n}{2(1+c+n)}\right] + \left(\frac{K-s-1}{2}\right)\log\left(\frac{1+c}{1+c+n}\right). \tag{1}$$

Under the null model, the distribution of x_n is χ^2_{K-s-1}. For fixed $c > 0$, the first term of (1) is bounded in probability, while the second tends to $-\infty$ as $n \to \infty$. Thus, $BF(2|1) \xrightarrow{p} -\infty$ under the null model.

Under the alternative hypothesis, the distribution of x_n is approximately noncentral $\chi^2_{K-s-1}(\lambda)$ with noncentrality parameter $\lambda = n\boldsymbol{\kappa}'\boldsymbol{\kappa}$, $\boldsymbol{\kappa} = \{(q_i - p_i)/\sqrt{p_i}\}$. If this hypothesis pertains, the first term in (1) dominates and $BF(2|1) \xrightarrow{p} \infty$ as $n \to \infty$. It follows that the Bayes factor is consistent for constant c. Note that if marginal maximum likelihood estimation (MMLE) is used to estimate $\alpha = (c+1)/n$, the resulting Bayes factor tends to ∞ under the alternative hypothesis, but remains bounded under the null hypothesis and so is not consistent.

F tests for linear models. Suppose that

$$\mathbf{y} \,|\, \boldsymbol{\beta}, \sigma^2 \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}),$$

where $\mathbf{y}$ is an $n \times 1$ observation vector, $\boldsymbol{\beta}$ is an $r \times 1$ regression parameter, $\mathbf{X}$ is an $n \times r$ matrix of rank r, and σ^2 is a scalar variance parameter. Assume further that under the null hypothesis, $\mathbf{H}'\boldsymbol{\beta} = \xi$, where $\mathbf{H}$ is an $r \times k$ matrix of rank k whose range space is contained in the range space of $\mathbf{X}'$, and let f_n denote the standard F statistic for testing the null hypothesis against the alternative that $\boldsymbol{\beta}$ does not satisfy this constraint. Under the alternative hypothesis, if $\boldsymbol{\beta}$ is drawn from an r-variate normal distribution centered on a value that does satisfy the null constraint and having covariance matrix $\tau\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$, then the logarithm of the Bayes factor in favor of the alternative, say $\log(BF(2|1))$, can be written

$$(2) \quad -\frac{k}{2}\log(1+n\tau^*) + \frac{k+m}{2}\log\left(1+\frac{kf_n}{m}\right) - \frac{k+m}{2}\log\left(1+\frac{kf_n}{m(1+n\tau^*)}\right)$$

where $m = n - r$ and $n\tau^* = \tau$.

Under the null hypothesis, $f_n = O_p(1)$. This implies that the first term in (2) dominates the sum, so that $BF(2|1) \xrightarrow{p} -\infty$ as $n \to \infty$.

Under the alternative hypothesis, $f_n/(1+n\tau^*)$ has a $F_{k,m}$ distribution. Consequently, $f/m = O_p(1)$ and $f/m > 0$ with probability 1. The second term in (2) is thus linear in m; the remaining terms are $O_p(\log(n))$ or less. It follows that $BF(2|1) \xrightarrow{p} \infty$ under the alternative hypothesis. Therefore, the Bayes factor based on the F statistic is consistent for fixed values of τ^* (but not for fixed values of τ).

Stomach cancer data revisited. White and Eisenberg (1959) provided a cross-classification of stomach cancer site with blood type for 707 cancer patients (Table 1). The purpose of their study was to determine whether there was an association between blood type and cancer site. The χ^2 test for independence for these data is 12.65 on 6 degrees of freedom.

Because White and Eisenberg did not specify an alternative hypothesis, it is not clear what value of c should be used to define the distribution of the χ^2 test statistic under the alternative model. This difficulty can be partially circumvented

	Blood Group		
Site	O	A	B or AB
Pylorus and antrum	104	140	52
Body and fundus	116	117	52
Cardia	28	39	11
Extensive	28	12	8

TABLE 1. White and Eisenberg's classification of cancer patients

by reporting test results for a range of alternative models corresponding to different values of c. This strategy is particularly appealing if the "weight of evidence" criteria suggested in Kass and Raftery (1995) are used. According to their scheme (which represents a variation on criteria proposed by Jeffreys (1961)), relative evidence in favor of one of the tested hypotheses is classified according to the value of twice the natural logarithm of Bayes factors. Based on this value, experimental evidence can be classified as "not worth more than a bare mention" (0-2), "positive"' (2-6), "strong " (6-10) or "very strong" (> 10).

Figure 1 illustrates the relative weight of evidence in favor of the independence versus dependence models for White and Eisenberg's data using these classifications. Using these classifications, the Bayes factor based on the χ^2 statistic suggests that White and Eisenberg's data provide (a) very strong evidence against alternative hypotheses generated from values of c in (0,16.5); (b) strong evidence against alternatives generated from values of c in (16.5,35.9); (c) positive evidence against alternatives generated from values of c in (35.9,86.0); and (d) evidence not worth mentioning for alternative models generated from values of c in (86,412) or values of $c > 1050$. There is positive evidence for alternative models generated with c in the range (412,1050), and The maximum evidence in favor of the alternative hypothesis occurs when c=636, and evidence favors the alternative model (though often in a way barely worth mentioning!) for $c > 150$.

When $c = 636$, the Bayes factor in favor of the alternative hypothesis is slightly less than 3 (twice the log of the Bayes factor is 2.17). Prior standard deviations of cell probabilities generated from this alternative hypothesis are approximately $\sqrt{p_i(1-p_i)}/25$, where $\{p_i\}$ denotes a probability vector satisfying the independence assumption. Such deviations ($\approx 4\%$) may or may not be regarded as substantively important.

REFERENCES

[1] H. Jeffreys, *Theory of Probability*, London: Oxford University Press (1961).
[2] V.E. Johnson, *Bayes Factors Based on Test Statistics*, **to appear** in Journal of the Royal Statistical Society, Series B (2005).
[3] R.E. Kass and A.E. Raftery, *Bayes Factors*, Journal of the American Statistical Association **90** (1995), 773–795 .

[4] C. White and H. Eisenberg, *ABO blood groups and cancer of the stomach*, Yale Journal of Biology and Medicine **32** (1959), 58–61 .

Model selection and aggregation in sparse classification problems

Vladimir Koltchinskii

Let (X,Y) be a random couple in $S \times \{-1,1\}$, the first component X being an observable instance in some space S and the second component Y being an unobservable label. The goal of binary classification is to predict the label based on the observation of the instance in the cases when the joint distribution P of (X,Y) is unknown, but a sample $(X_1,Y_1),\dots(X_n,Y_n)$ of i.i.d. copies of (X,Y) is available. Given a classifier $g : S \mapsto \{-1,1\}$, its generalization error is defined as

$$\mathbb{P}\{Y \neq g(X)\} = P\{(x,y) : y \neq g(x)\}$$

and it can be estimated by the corresponding training error

$$n^{-1}\sum_{j=1}^{n} I(Y_j \neq g(X_j)) = P_n\{(x,y) : y \neq g(x)\},$$

where P_n denotes the empirical measure based on the data. In principle, one would like to find a classifier with a small (ideally, minimal) generalization error and a naive approach to this would be to minimize the training error (which is the empirical risk with respect to the binary loss) over a suitable (not too large and not too small) class of functions g. However, modern theory (and practice) of classification is based on replacing the binary loss by its smooth and most often convex approximation. Classification algorithms then are looking for classifiers that minimize the penalized empirical risk with such a convex loss and with a properly chosen complexity penalty.

We will be looking at the following version of this problem. Let $h_1,\dots,h_N : S \mapsto [-1,1]$ be given functions. They can be viewed as features, base classifiers, previously trained classifiers in aggregation problems, etc. Functions

$$f_\lambda := \sum_{j=1}^{N} \lambda_j h_j,\ \lambda \in \mathbb{R}^N$$

will be viewed as real valued classifiers, the corresponding binary classifier being $S \ni x \mapsto \mathrm{sign}(f_\lambda(x)) \in \{-1,1\}$.

Let $\ell : \mathbb{R} \mapsto \mathbb{R}_+$ be a decreasing convex loss function such that $\ell(u) \to 0$ as $u \to +\infty$, $\ell(u) \to +\infty$ as $u \to -\infty$ and $\ell \geq I_{(-\infty,0]}$. In addition to this, we will be assuming (and these assumptions are restrictive) that ℓ is twice continuously differentiable, ℓ' and ℓ'' are uniformly bounded in $\mathbb{R}$ and $\ell''(u) > 0$ for all $u \in \mathbb{R}$. The so called logit loss $\ell(u) = \log_2(1+e^{-u})$ is a typical example of such a loss function.

Denote $(\ell \bullet f)(x,y) := \ell(yf(x))$ and define the ℓ-risk of a real valued classifier f :

$$P(\ell \bullet f) = \int (\ell \bullet f) dP = \mathbb{E}\ell(Yf(X))$$

Its empirical version (the empirical ℓ-risk) is:

$$P_n(\ell \bullet f) = \int (\ell \bullet f) dP_n = n^{-1} \sum_{j=1}^{n} \ell(Y_j f(X_j))$$

Suppose $\varepsilon \geq 0$ and $p \geq 1$.

$$\lambda^{\varepsilon} := \mathrm{argmin}_{\lambda \in \mathbb{R}^N} \left[P(\ell \bullet f_\lambda) + \varepsilon \sum_{j=1}^{N} |\lambda_j|^p \right]$$

and

$$\hat{\lambda}^{\varepsilon} := \mathrm{argmin}_{\lambda \in \mathbb{R}^N} \left[P_n(\ell \bullet f_\lambda) + \varepsilon \sum_{j=1}^{N} |\lambda_j|^p \right].$$

In particular,

$$\lambda^0 := \mathrm{argmin}_{\lambda \in \mathbb{R}^N} P(\ell \bullet f_\lambda).$$

We assume in what follows that λ_0 exists (which is the case if $\mathbb{P}\{Yf_\lambda(X)\} > 0$ for all $\lambda \in \mathbb{R}^N$).

This can be viewed as an approach to optimal linear aggregation of given base classifiers $h_1, \ldots, h_N$ in spirit of Nemirovski (2000) or Tsybakov (2003) who dealt mostly with aggregation of regression estimates (see also more recent work of Bunea, Tsybakov and Wegkamp (2004) who obtained a number of results on aggregation in regression context that are close to what we are trying to do here in the case of classification). The problems of this nature have been often looked at in the case of ℓ_1-penalty, i.e. for $p = 1$ (in regression, such a penalization is often called LASSO, see Tibshirani (1996)). We will consider the case when $p > 1$, but it is close enough to 1, so that $p-1$ is of the order $1/\log N$. To be specific, suppose that $\frac{1}{p} + \frac{1}{q} = 1$ and take q such that $N^{1/q} = 2$. Under this assumption

$$\|\lambda\|_{\ell_p} \leq \|\lambda\|_{\ell_1} \leq 2\|\lambda\|_{\ell_p}.$$

Our goal is to provide (partial) answers to the following questions:

- suppose λ^{ε} is "sparse". Is $\hat{\lambda}^{\varepsilon}$ "sparse"?
- what impact does "sparsity" of λ^{ε} have on the size of the excess ℓ-risk

$$P(\ell \bullet f_{\hat{\lambda}^{\varepsilon}}) - P(\ell \bullet f_{\lambda^0})$$

and of

$$\|\hat{\lambda}^{\varepsilon} - \lambda^0\|_{\ell_1} = \sum_{j=1}^{N} |\hat{\lambda}_j^{\varepsilon} - \lambda_j^0|?$$

- how the "sparsity bounds" depend on ε and how to choose ε?

Note that sparsity of the solution λ^ε of the "true" problem might occur naturally, for instance, in the following situation. Suppose there exists a small subset $J \subset \{1, 2, \dots, N\}$ such that the sets of random variables

$$\{Y, h_j(X), j \in J\} \text{ and } \{h_j(X), j \notin J\}$$

are independent. In other words, $\{h_j, j \in J\}$ are "relevant features" for classification whereas $\{h_j, j \notin J\}$ are "irrelevant." In addition, suppose that $\mathbb{E}h_j(X) = 0, j \notin J$. Then, using Jensen's inequality, it is easy to check that for all $\varepsilon \geq 0$ $\lambda^\varepsilon_j = 0, j \notin J$ (so, λ^ε is "sparse").

In what follows, we will use the following definition of a "sparsity function" of vector $\lambda \in \mathbb{R}^N$: for $d = 0, 1, \dots, N$, let

$$\gamma_d(\lambda) := \min\left\{\sum_{j \notin J} |\lambda_j| : \#(J) = d,\ J \subset \{1, 2, \dots, N\}\right\} = \sum_{j=d+1}^{N} |\lambda_{[j]}|,$$

where

$$|\lambda_{[1]}| \geq |\lambda_{[2]}| \geq \cdots \geq |\lambda_{[N]}|$$

is a nonincreasing rearrangement of the coefficients.

Clearly, $\gamma_d(\lambda)$ is a nonincreasing function of d and, if $\gamma_d(\lambda) = 0$, then there are at most d nonzero coordinates of vector $\lambda \in \mathbb{R}^N$.

Theorem 1. *There exist constants $c, C > 0$ depending only on ℓ such that for all $A > 0$ and for all*

$$\varepsilon \geq c\sqrt{\frac{A \log N}{n}},$$

the following inequalities hold:

$$\mathbb{P}\left\{\|\hat{\lambda}^\varepsilon\|_{\ell_1} \geq C(\|\lambda^0\|_{\ell_1} + 1)\right\} \leq N^{-A}$$

and also with probability $\geq 1 - N^{-A}$

$$C^{-1}(\|\lambda^{2\varepsilon}\|_{\ell_1} - N^{-1}) \leq \|\hat{\lambda}^\varepsilon\|_{\ell_1} \leq C\|\lambda^{\varepsilon/2}\|_{\ell_1} + N^{-1}.$$

Theorem 2. *There exist constants $c > 0$ depending only on ℓ and $K > 0$ depending on ℓ and on $\|\lambda^0\|_{\ell_1}$ such that for all $\varepsilon \geq c\sqrt{\frac{d + A \log N}{n}}$, we have with probability $\geq 1 - N^{-A}$*

$$\gamma_d(\hat{\lambda}^\varepsilon) \leq \min_{0 \leq m \leq d}\left[2\gamma_m(\lambda^\varepsilon) + K \log N \sqrt{\frac{m + A \log N}{n}}\right]$$

and

$$\gamma_d(\lambda^\varepsilon) \leq \min_{0 \leq m \leq d}\left[2\gamma_m(\hat{\lambda}^\varepsilon) + K \log N \sqrt{\frac{m + A \log N}{n}}\right].$$

Moreover, if $\gamma_d(\lambda^\varepsilon) = 0$, then

$$\mathbb{P}\left\{\gamma_d(\hat{\lambda}^\varepsilon) \geq K\sqrt{\frac{d + A \log N}{n}}\right\} \leq N^{-A}.$$

In what follows we assume (*the sparsity assumption*) that there exists $J^* \subset \{1, 2, \ldots, N\}$ with $\#(J^*) = d^*$ such that, for all $\varepsilon \geq 0$, $\lambda_j^\varepsilon = 0, j \notin J^*$. This implies that $\gamma_{d^*}(\lambda^\varepsilon) = 0$ for all $\varepsilon \geq 0$. In addition, assume that $\{h_j : j \in J^*\}$ are linearly independent functions in $L_2(\Pi)$ (Π is the distribution of X).

Theorem 3. *Under the sparsity assumption, there exists a constant $c > 0$ depending only on ℓ such that, for all $A > 0$ and for*

$$\varepsilon = \varepsilon_d = c\sqrt{\frac{d + A\log N}{n}},$$

with probability $\geq 1 - N^{-A}$

$$\|\hat{\lambda}^\varepsilon - \lambda^0\|_{\ell_1} \leq K\left[\gamma_d(\lambda^\varepsilon) + \log N\sqrt{\frac{d + A\log N}{n}}\right]$$

and

$$P(\ell \bullet f_{\hat{\lambda}^\varepsilon}) - P(\ell \bullet f_{\lambda^0}) \leq K\left[\gamma_d(\lambda^\varepsilon)\sqrt{\frac{d + A\log N}{n}} + \log N\frac{d + A\log N}{n}\right]$$

with a constant $K > 0$ depending on $\ell, \|\lambda^0\|_{\ell_1}$ and $\{h_j : j \in J^*\}$.

Building upon the above results, we now suggest adaptive choices of regularization parameter $\varepsilon > 0$. Define for a fixed $A > 0$

$$\hat{d} := \mathrm{argmin}_{0 \leq d \leq N}\left[\gamma_d(\hat{\lambda}^{\varepsilon_d})\sqrt{\frac{d + A\log N}{n}} + \frac{d + A\log N}{n}\right],$$

$$\check{d} := \mathrm{argmin}_{0 \leq d \leq N}\left[\gamma_d(\hat{\lambda}^{\varepsilon_d}) + \sqrt{\frac{d + A\log N}{n}}\right]$$

and

$$\hat{\lambda} := \hat{\lambda}^{\varepsilon_{\hat{d}}}, \check{\lambda} := \hat{\lambda}^{\varepsilon_{\check{d}}}.$$

Theorem 4. *With probability $\geq 1 - N^{-A}$*

$$\|\check{\lambda} - \lambda^0\|_{\ell_1} \leq K\log N\sqrt{\frac{d^* + A\log N}{n}}$$

and

$$P(\ell \bullet f_{\hat{\lambda}}) - P(\ell \bullet f_{\lambda^0}) \leq K\log N\frac{d^* + A\log N}{n}$$

with a constant $K > 0$ depending on $\ell, \|\lambda^0\|_{\ell_1}$ and $\{h_j : j \in J^*\}$.

References

[1] F. Bunea, A. Tsybakov and M. Wegkamp, *Aggregation for regression learning*, Preprint (2004).

[2] A. Nemirovski, *Topics in non-parametric statistics*, In P. Bernard, editor, Ecole d'Eté de Probabilités de Saint-Flour, 1998, Lecture Notes in Mathematics, Springer, New York, **XXVIII** (2000).

[3] R. Tibshirani, *Regression shrinkage and selection via the Lasso*, J. Royal Statist. Soc. Ser B. **58** (1996), 267-288.

[4] A. Tsybakov, *Optimal rates of aggregation*, In Proc. 16th Annual Conference on Learning Theory (COLT) and 7th Annual Workshop on Kernel Machines, Lecture Notes in Artificial Intelligence, Springer, New York, **2777** (2003), 303–313.

Nonparametric Estimation in a Stochastic Volatility Model

Jens-Peter Kreiss
(joint work with Andreas Dürkes)

The talk started with a discussion of the possible nonparametric extension of the GARCH-model for log-returns $X_t = \log S_t$, of an financial asset price S_t. Bühlmann and McNeil (2002) proposed an algorithmic approach in order to estimate the nonparametric GARCH-function. Their result, which more or less is related to a non-stochastic situation, could be regarded as a kind of a fixed-point theorem. However, estimation in nonparametric GARCH-models has to face an errors-in-variables situation.

The paper then lead the focus to a nonparametric extension of a simple stochastic volatility model (the so-called lagged autoregressive variance model) introduced by Taylor (1994) for log-returns X_t of the following type

$$X_t = \mu + \sigma_{t-1} \cdot e_t \ ,$$

where $\xi_t := \log \sigma_t$ is assumed to satisfy a first order nonparametric autoregressive scheme

$$\xi_t = m(\xi_{t-1}) + \tau \cdot \eta_t \ .$$

Both η_t and e_t are assumed to be (possibly correlated) random variables with zero mean and unit variance. Typically the so-called volatility process ξ_t is not observable. The assumption of a strictly positive Lebesgue density of η_1 and

$$\limsup_{|x|\to\infty} \left| \frac{m(x)}{x} \right| | < 1$$

ensures geometric ergodicity of the volatility and the return process.

Concerning nonparametric estimation of the function m on the basis of the log-returns X_t Franke et al. (2003), after a logarithmic transformation of the returns, successfully applied so-called nonparametric deconvolution kernel smoothers known from nonparametric regression models with errors-in-variables (cf. Fan (1991) and Fan et al. (1993)). For the case where (η_t, e_t) possesses a bivariate normal distribution it is obtained that we are in the so-called super-smooth case which leads to relatively poor logarithmic rates of convergence for this deconvolution kernel estimator (cf. Franke et. al (2003)). A further drawback of this nonparametric deconvolution estimator is, that it has to be assumed that the density of the convoluting random variable, which is e_t, is known. To overcome this latter drawback one may follow a suggestion of Horowitz (1998) for panel data in order to estimate the distribution of the convoluting random variables. Following

this idea one obtains for a specific situation in which at least two returns with exactly the same volatility can be observed that we are able to consistently estimate the wanted distribution (cf. Dürkes and Kreiss (2005)). However, the rather poor logarithimic rate of convergence of the estimator is still there.

A more promising proposal is to make use of the so-called realized volatility (cf. Andersen et al. (2001), Barndorff-Nielsen et al. (2002) or Ait-Sahalia et al. (2004)) calculated from intraday returns in order to estimate the wanted daily volatility. In other words the idea is to try to estimate the daily volatility from higher (higher than daily) frequency returns. If we, for example, are able to observe an increasing (with the sample size n) number of intraday returns, e.g. hourly or 30 minutes returns and so on, then we receive an approximation of the daily volatility which is precise enough in order to overcome the deconvolution dilemma. Such a situation is comparable to a deconvolution problem in which the variance of the convoluting random variable converges to zero with increasing sample size. Under specific assumptions it is shown that we may end up with the same asymptotic results (asymptotic normality) as if the volatility process itself would be observable.

A drawback of this proposal is that we definitely need to assume that the increasing number of intraday returns converges to infinity with growing sample size. Since it is known that so-called microstructure noise appears for returns beyond a certain time grid of 15 to 30 minutes, say, and that this microstructure noise leads to inconsistent estimates of the integrated volatility (as is shown in Ait-Sahalia et al. (2004)) we have to use in real applications intraday returns for computing the realized volatility, which are not sampled more frequently than every 15 or 30 minutes. Another possibility is to follow a proposal by Ait-Sahalia et al. (2004) which overcomes the problem of inconcistency of realized volatility due to microstructure noise.

References

[1] Y. Ait-Sahalia, L. Zhang and P. Mykland, *A Tale of Two Time Scales: Determining Integrated Volatility with Noisy High-Frequency Data*, Journal of the American Statistical Association (2004), to appear.

[2] T.G. Andersen, T. Bollerslev, F.X. Diebold and P. Labys, *The distribution of exchange rate realized volatility*, Journal of the American Statistical Association **96** (2001), 42–55.

[3] O.E. Barndorff-Nielsen and N. Shephard, *Econometric analysis of realized volatility and its use in estimating stochastic volatility models*, Journal of the Royal Statistical Society Ser. B **64** (2002), 253–280.

[4] P. Bühlmann and A.J. McNeil, *An algorithm for nonparametric GARCH modelling*, Computational Statistics and Data Analysis **40** (2002), 665–683.

[5] A. Dürkes and J.-P. Kreiss, *Nonparametric modelling and estimation of stochastic volatility*, Prepint (2005).

[6] J. Fan, *On the optimal rates of convergence for nonparametric deconvolution problems*, The Annals of Statistics **19** (1991), 1257–1272.

[7] J. Fan and Y.K. Truong, *Nonparametric regression with errors in variables*, The Annals of Statistics **21** (1993), 1900–1925.
[8] D. Feldmann, W. Härdle, C. Hafner, M. Hoffmann, O. Lepski and A. Tsybakov, *Testing Linearity in an AR Errors-in-variables Model with Application to Stochastic Volatility*, Applicationes Mathematicae **30** (2003), 389–412.
[9] J. Franke, W. Härdle and J.-P Kreiss, *Nonparametric estimation in a Stochastic Volatility model*, Recent advances and trends in Nonparametric Statistics, M. Akritas and D.N. Politis (eds.) Elsevier (North Holland) (2003), 303–314.
[10] J.L. Horowitz, *Semiparametric methods in econometrics*, Lecture Notes in Statistics **85** (1998), Springer-Verlag.
[11] S.J. Taylor, *Modelling stochastic volatiliy: A review and comparative study*, Mathematical Finance **4** (1994), 183–204.

Model selection for ill-posed inverse problems

Jean-Michel Loubes
(joint work with Ana Karina Fermin, Carenne Ludeña)

We are interested in recovering an unobservable signal x_0 based on observations

$$y(t_i) = F(x_0)(t_i) + \varepsilon_i, \tag{1}$$

where $F : X \to Y$ is a regular functional, with X, Y Hilbert spaces and $t_i, i = 1, \ldots, n$ is a fixed observation scheme. $x_0 : \mathbb{R} \to \mathbb{R}$ is the unknown function to be recovered from the data $y(t_i)$, $i = 1, \ldots, n$. The regularity condition over the unknown parameter of interest is expressed through the assumption $x_0 \in X$. We assume that the observations $y(t_i) \in \mathbb{R}$ and that the observation noise ε_i are i.i.d. realizations of a certain random variable ε. Throughout the paper, we shall denote $\mathbf{y} = (y(t_i))_{i=1}^n$. We assume F is Fréchet differentiable and ill posed in the sense that our noise corrupted observations might lead to large deviations when trying to estimate x_0. In a deterministic framework, the statistical model is formulated as the problem of approximating the solution of $F(x) = y$, when y is not known, and is only available through an approximation y^δ,

$$\|y - y^\delta\| \leq \delta.$$

It is important to remark that whereas in this case consistency of the estimators depends on the approximation parameter δ, which depends on the number of observations n.

In the linear case, the best L^2 approximation of x_0 is $x^+ = F^+ y$, where F^+ is the Moore-Penrose (generalized) inverse of F. We will say the problem is ill-posed if F^+ is unbounded. This might entail, and is generally the case, that $F^+(y^\delta)$ is not close to x^+. Hence, the inverse operator needs to be, in some sense, regularized. In the nonlinear case, by ill-posedness we will always mean that the solutions do not depend continuously on the data. If F is a compact operator, local injectivity around x^+ a solution of $F(x^+) = y$, implies ill-posedness of the problem.

Regularization methods replace an ill-posed problem by a family of well-posed

problems. Their solution, called regularized solutions, are used as approximations of the desired solution of the inverse problem. These methods always involve some parameter measuring the closeness of the regularized and the original (unregularized) inverse problem. Rules (and algorithms) for the choice of these regularization parameters as well as convergence properties of the regularized solutions are central points in the theory of these methods, since they allow to find the right balance between stability and accuracy.

Our goal in this article is to develop algorithms providing estimators that achieve optimal rates of convergence when the smoothness of the true solution is not known a priori. We will assume operator F can be linear or non linear, but in the latter case we will assume it satisfies a local linear invariance condition.

For this we consider penalized M-estimators minimizing quantities of the form

$$\hat{x}_n = \arg \min_{\alpha_n \in \Theta} \arg\min_{x \in \mathcal{X}} \left(\gamma(y - F(x)(t)) + \alpha_n \mathrm{pen}(x, \mathcal{X}) + pen(\alpha_n)\right), \tag{2}$$

where $\mathcal{X}$ is a specific set, $\gamma(.)$ is a loss-function, $\mathrm{pen}(.,.)$ is a penalty over x and/or $\mathcal{X}$, and $\alpha_n \in \Theta$ is a decreasing sequence all of which will be defined precisely later. The idea of penalized M-estimators is to find an estimator close enough to the data, close in the sense defined by γ and with a regularity property induced by the choice of the penalty pen. Adaptivity means here, that the construction of the estimator does not require knowing beforehand the regularity of the function of interest to be recovered x_0. In the inverse problems literature this is known as a posteriori methods. But, we do assume that the inverse operator is known as well as some assumptions, such as its degree of ill-posedness.

When F is linear, the statistical problem has been extensively studied, although in general efficient adaptive regularization-parameter choice is still under active research. Two main kinds of estimators have been considered. First regularized estimators such as Tikhonov type estimators, then non linear thresholded estimators. We refer to (5), (6), (2) or (1) for more references.

1. Definitions and notations

We introduce certain standard assumptions on the observation noise

AN moment condition for the errors: ε is a centered random variable satisfying the moment condition $\mathbf{E}(|\varepsilon|^p/\sigma^p) \leq p!/2$ and $\mathbf{E}(\varepsilon^2) = \sigma^2$.

The smoothness of the function is expressed through the following assumption.

SC source condition: There exists $0 < \nu \leq 1/2$ such that

$$x_0 \in Range((T^*T)^\nu) = \mathcal{R}((T^*T)^\nu).$$

We assume the regularity of the problem is defined by that of $F'(x_0)$. In the linear case we will write $F = T$. This linear operator acts with a degree of ill-posedness defined by an index p. This is generally expressed by the fact that T maps L^2 into some Sobolev space H_p, or by assuming that T acts along a Hilbert scale H_s.

IP ill posedness of the operator: There exists $p > 0$ such that $F'(x_0)(H_s) = H_{s+p}$.

Now consider approximation subsets, namely $\forall m \in \mathcal{M}_n$, $Y_m \subset Y$ with $d_m = \dim(Y_m)$. We start out with a big enough subspace Y_{m_0} and in order to deal with the ill posedness of $(T^*\Pi^n_{Y_m})^+$ use Tikhonov-like regularization methods. Consider the corresponding sets in space X, defined by $X_m = (\Pi^n_{Y_m} T)^+ X$. To begin assume that m_0 is such that

$$\|(I - \Pi_{X_{m_0}})x_0\| \leq \inf_m [\|(I - \Pi_{X_m})x_0\| + \sqrt{\frac{d_m}{n}}\frac{1}{\gamma_m}].$$

This quantity can be chosen so as not to depend on the unknown regularity of the solution x_0. Under assumption **SC** the above inequality is satisfied if the dimension of the set is such that

$$d_{m_0}^{2\nu p} \geq n^{\frac{2\nu p}{4\nu p + 2p + 1}}.$$

Thus it is enough to choose m_0 such that $d_{m_0} \geq n^{1/(2p+1)}$.

2. Estimation procedure

Next consider for a given $k \in \mathcal{K}$ a sequence of (diagonal) matrixes $A_k(n) \in \mathbb{M}_{m_0 \times m_0}(\mathbb{R})$ as and define the following penalized estimator:

$$\hat{x}_{A_k(n)} = \arg\min_{x \in X_m} \left[\|\Pi_{Y_{m_0}}(y - F(x))\|_n^2 + \|A_k(n)x\|^2\right]. \tag{3}$$

The behaviour of this general estimator depends on the choice of the smoothing matrixes $A_k(n)$. For particular choices of the penalty, this estimator can be seen as a Tikhonov regularized estimator or a projection estimator. We would like to choose $A_k(n)$, among all the $A_k(n)$, $k \in \mathcal{K}$ based on the data in such a way that optimal rates are maintained. This choice must also not depend on a priori regularity assumptions.

First we note that, for each fixed $A_k(n)$, the expression (3) can be written in the following way:

$$\hat{x}_{A_k(n)} = \arg\min_{x \in X_{m_0}} \|(T^t\Pi_{Y_{m_0}}T + A_k(n)^t A_k(n))^{-1}T^t\Pi_{Y_{m_0}}(y - T(x))\|^2. \tag{4}$$

In practice the second one is more complicated (the matrix to inverse might be big), but it is simpler to deal with in order to show our results concerning the selection of A_k. With this notation set

$$R_{A_k(n)} = (T^t\Pi_{Y_{m_0}}T + A_k(n)^t A_k(n) I_{m_0})^{-1}T^t\Pi_{Y_{m_0}}.$$

Now set $\gamma(x, \alpha_k) = \|R_{A_k}(y - F(x))\|^2$ and

$$\text{pen}(A_k) = r\sigma^2(1 + L_k)[Tr(R_{A_k}^t R_{A_k}) + \rho^2(R_{A_k})],$$

with $r > 2$. We choose $\hat{x}_{A_{\hat{k}}}$ such that

$$\hat{x}_{A_{\hat{k}}} = \arg\min_{k \in \mathcal{K}, x \in X_{m_0}} \left(\gamma(x, A_k(n)) + \text{pen}(A_k(n))\right).$$

Let $x_{A_k} = R_{A_k} T x_0$. Set

$$\Sigma(d) = \sum_k 2 \left[\sqrt{\frac{dTr(R_{A_k}^t R_{A_k})}{\rho^2(R_{A_k})}} + 1 \right] \frac{d}{\rho^2(nR_{A_k})} e^{-\sqrt{dL_k \frac{Tr(R_{A_k}^t R_{A_k}) + \rho^2(R_{A_k})}{\rho^2(R_{A_k})}}},$$

for d. We have the following result.

Theorem. There exists a constant C which depends on r and on T, such that the following inequality holds true

$$\mathbf{E}\|\hat{x}_{A_{\hat{k}}} - x_0\|^2 \leq 2\|(I - \Pi_{X_{m_0}})x_0\|^2 + C \inf_{k \in \mathcal{K}} \left[\|x_{A_k} - x_0\|^2 + 2\mathrm{pen}(A_k) \right] + \frac{\Sigma(d)}{n}. \tag{5}$$

Hence, the estimator is optimal in the sense that the adaptive estimator achieves the best rate of convergence among all the regularized estimators.

Proof. In the linear case $F = T$, for any x_{A_k} and any $k \in \mathcal{N}$

$$\|R_{A_{\hat{k}}}(y - T\hat{x}_{A_{\hat{k}}})\|^2 + pen(A_{\hat{k}}) \leq \|R_{A_k}(y - Tx_{A_k})\|^2 + pen(A_k)$$

and

$$\|R_{A_k}(y - Tx_{A_k})\|^2 = \|R_{A_k}T(x_0 - x_{A_k})\|^2 + 2\langle R_{A_k}T(x_0 - x_{A_k}), R_{A_k}\varepsilon\rangle + \|R_{A_k}\varepsilon\|^2$$

Thus, following standard arguments we have

$$\begin{aligned}\|R_{A_{\hat{k}}}T(x_0 - \hat{x}_{A_{\hat{k}}})\|^2 \leq\ & \|R_{A_k}T(x_0 - x_{A_k})\|^2 \\ & - 2 < R_{A_{\hat{k}}}T(x_0 - \hat{x}_{A_{\hat{k}}}), R_{A_{\hat{k}}}\varepsilon > + 2 < R_{A_k}T(x_0 - x_{A_k}), R_{A_k}\varepsilon > \\ & - \|R_{A_{\hat{k}}}\varepsilon\|^2 + \|R_{A_k}\varepsilon\|^2 + \mathrm{pen}(A_k) - \mathrm{pen}(A_{\hat{k}}).\end{aligned}$$

Let $0 < \kappa < 1$. Since $2ab \leq \kappa a^2 + \frac{1}{\kappa} b^2$, for any a, b we have for any k and $x_{A_k} \in X_m$

$$\begin{aligned}(1 - \kappa)\|R_{A_{\hat{k}}}T(x_0 - \hat{x}_{A_{\hat{k}}})\|^2 \leq\ & (1 + \kappa)\|R_{A_k}T(x_0 - x_{A_k})\|^2 \\ & + (2 + \frac{1}{\kappa})\mathrm{pen}(A_k) + 2\sup_k\{(1 + \frac{1}{\kappa})\|R_{A_k}\varepsilon\|^2 - \mathrm{pen}(A_k)\},\end{aligned}$$

On the other hand, using that is $1 \leq R_{A_k}T \leq C$, we have that for any $x_{A_k} \in X_{m_0}$ and any $k \in \mathbb{N}$,

$$\begin{aligned}(1 - \kappa)\|x_0 - \hat{x}_{A_{\hat{k}}}\|^2 \leq\ & C(1 - \kappa)\|x_0 - x_{A_k}\|^2 \\ & + (2 + \frac{1}{\kappa})\,\mathrm{pen}(A_k) + 2\,C_1 \sup_k \{\|R_{A_k}\varepsilon\|^2 - (1 + \frac{1}{\kappa})^{-1}\mathrm{pen}(A_k)\}.\end{aligned}$$

As above, the proof then follows directly from the concentration of the supremum of the empirical process under the linear application as defined by the regularization family, see (4). The choice of κ will depend on r in the penalty. □

Particular choises of penalty enable to handle nonlinear operatores, see (4) and iterative methods, see (3).

References

[1] L. Cavalier, G. K. Golubev, D. Picard and A. B. Tsybakov, *Oracle inequalities for inverse problems*, Ann. Statist. **30(3)** (2002), 843–874.

[2] Laurent Cavalier and Alexandre Tsybakov, *Sharp adaptation for inverse problems with random noise*, Probab. Theory Related Fields **123(3)** (2002), 323–354.

[3] A-K. Fermin, C. Ludeña, *Iterative methods for inverse problems*, working paper (2005).

[4] J-M. Loubes, C. Ludenña, *Complexity regularization for general inverse problems*, Prépublications de l'Université Paris Sud (2005).

[5] Finbarr O'Sullivan, *A statistical perspective on ill-posed inverse problems*, Statist. Sci. **1(4)** (1986), 502–527.

[6] Bernard A. Mair and Frits H. Ruymgaart, *Statistical inverse estimation in Hilbert scales*, SIAM J. Appl. Math. **56(5)** (1996), 1424–1444.

FDR and Bayesian Multiple Comparison Rules

Peter Müller

(joint work with Giovanni Parmigiani)

We discuss a Bayesian decision theoretic approach to multiple comparison problems for a large number of comparisons, and the relationship with the false discovery rate (FDR). The motivating example is inference in group comparison microarray experiments. Among genes, $i = 1, \ldots, n$, for large n, we need to identify those that are differentially expressed across two biologic conditions of interest. It can be argued that Bayesian posterior inference already accounts for multiplicities, and no further adjustment is required (Scott and Berger, 2006). The argument is valid with respect to the evaluation of posterior probabilities of differential expresison. But this only solves half the problem. We still need to address the second step of the inference problem, namely the identification of differentially expressed genes. Berry and Hochberg (1999) discuss this perspective. This identification is most naturally discussed as a decision problem. Let $\delta_i \in \{0, 1\}$ denote the decision for gene i, with $\delta_i = 1$ indicating that the gene is flagged as differentially expressed. Let $r_i \in \{0, 1\}$ denote an indicator for the (unknown) truth, i.e., a parameter. The following discussion is valid for any probability model that includes parameters r_i with positive prior probabilty, $0 < p(r_i = 1) < 1$. In a decision theoretic approach, a loss function $L(\delta, r)$ is used to formalize the relative preference for a possible decisions δ, for assumed hypothetical true values r. The loss function is implicitely a function of the data through δ. We write $\delta(y)$ when we want to highlight the nature of δ as a function of the data. In the context of a decision problem with a probability model on the random varibles (r, y) and a loss function $L(\delta(y), r)$, the optimal decision is the rule $\delta(y)$ that maximizes L in expectation. The relevant expectation is the probability model on r conditional on the observed

data, leading to the optimal rule

$$\delta^*(y) = \arg\min_{\delta} \int L(\delta(y), r)\; p(r \mid y)\; \mathrm{d}r.$$

Usually, the probability model includes additional parameters besides r. We therefore interpret the probability model $p(r \mid y)$ as the marginal posterior distribution of the indicators r given the observed data.

Let $v_i = E(r_i \mid y)$ denote the marginal posterior probability of gene i being differentially expressed. The assumption of non-zero prior probabilities, $0 < p(r_i = 1) < 1$, ensures non-trivial posterior probabilities. In Müller et al. (2004) we show that for several reasonable choices of $L(\delta(y), r)$ the optimal rule is of the form

$$\delta_i^*(y) = I(v_i > t). \tag{1}$$

In words, the optimal decision is to mark all those genes as differentially expressed that have marginal posterior probability v_i beyond a certain threshold t. The value of the threshold depends on the specific loss function. The optimal rule (1) is valid for several loss functions defined in Müller et al. (2004). Essentially, all are variations of basic 0-1 loss functions. Let $\mathrm{FD} = \sum \delta_i\,(1-r_i)$ and $\mathrm{FN} = \sum(1-\delta_i)\,r_i$ denote false discovery and negative counts, and let $\mathrm{FDR} = \mathrm{FD}/\sum \delta_i$ and $\mathrm{FNR} = \mathrm{FN}/\sum(1-\delta_i)$ denote false discovery and false negative rates. The definitions FD(R) and FN(R) are summaries of parameters, r, and data, $\delta(y)$. Taking an expectation with respect to y and conditioning on r one would arrive at the usual definition of false discovery rates, as used, among many others, in Benjamini and Hochberg (1995), Efron and Tibshirani (2002), Storey (2002, 2003), and Storey et al. (2004). Instead we use posterior expectations, defining $\overline{\mathrm{FD}} = E(\mathrm{FD} \mid y)$, etc. See, Genovese and Wasserman (2002,2003) for a discussion of posterior expected FDR. Using these posterior summaries we define the following losses: $L_N(\delta, z) = c\,\overline{\mathrm{FD}} + \overline{\mathrm{FN}}$, and $L_R(\delta, z) = c\,\overline{\mathrm{FDR}} + \overline{\mathrm{FNR}}$. The loss function L_N is a natural extension of $(0, 1, c)$ loss functions for traditional hypothesis testing problems (Lindely 1971). Alternatively, we consider bivariate loss functions that explicitly acknowledge the two competing goals: $L_{2R}(\delta, z) = \overline{\mathrm{FNR}}$, subject to $\overline{\mathrm{FDR}} < \alpha_R$, and $L_{2N}(\delta, z) = \overline{\mathrm{FN}}$, subject to $\overline{\mathrm{FD}} < \alpha_N$. Under all four loss functions, L_N, L_R, L_{2R} and L_{2N}, the nature of the optimal rule is (1). See Müller et al. (2004) for the definition of the thresholds.

One can argue that not all false negatives and all discoveries are equally important. False negatives of genes that are massively differentially expressed are more serious than only marginally differentially expressed genes. To formalize this notion we need to assume that the probability model includes parameters that can be interpreted as extent of differential expression, or strength of the signal. Assume that the model includes paramters ρ_i, $i = 1, \ldots, n$, with $\rho_i > 0$ if $r_i = 1$ and $\rho_i = 0$ if $r_i = 0$. For example, a popular class of sampling models for microarray data assumes that recorded gene expressions for gene i under the two conditions of interest are Gamma distributed with equal shape parameters, and scale parameters θ_{i0} and θ_{i1} (Newton et al. 2001, 2004). In this model a reasonable definition would use $\rho_i = \log(\theta_{i1}/\theta_{i0})$. Assuming parameters ρ_i that can be interpreted as

level of differential expression for gene i, we define

$$L_\rho(\rho, \delta, z) = -\sum \delta_i \, \rho_i + k \sum (1 - \delta_i)\rho_i + c \sum \delta_i.$$

The definition includes rewards for correct discoveries and penalties for false negatives that are proportional to the size of the signal. Without including the last term, $c \sum \delta_i$, the loss function would lead to the trivial solution $\delta_i = 1$ for all i. For $c > 0$ the optimal solution is easily found as $\delta_i^* = I\{\bar{\rho}_i \geq c/(1+k)\}$.

Additional assumptions on the probability model allow to further generalize the loss function. For example, if we assume that the sampling model for the observed gene expressions include parameters for the dependence structure, one could include terms related to the dependence.

References

[1] Y. Benjamini and Y. Hochberg, *Controlling the false discovery rate: A practical and powerful approach to multiple testing*, Journal of the Royal Statistical Society B **57** (1995), 289–300.

[2] D.A. Berry and Y. Hochberg, *Bayesian perspectives on multiple comparisons*, Journal of Statistical Planning and Inference **82**(1999), 215–227 .

[3] S. Berry and D. Berry, *Accounting for multiplicities in assessing drug safety: A three-level hierarchical mixture model*, Biometrics **60** (2004), 418–426.

[4] K. Do, P. Müller and F. Tang, *A bayesian mixture model for differential gene expression*, Journal of the Royal Statistical Society C **54** (2005), 627–644.

[5] B. Efron and R. Tibshirani, *Empirical bayes methods and false discovery rates for microarrays*, Genetic Epidemiology **23** (2002), 70–86.

[6] B. Efron, R. Tibshirani, J.D. Storey and V. Tusher, *Empirical Bayes analysis of a microarray experiment*, Journal of the American Statistical Association **96** (2001), 1151–1160.

[7] M. Fortini, B. Liseo, A. Nuccitelli and M. Scanu, *On bayesian record linkage*, Research in Official Statistics **4** (2001), 185–198.

[8] C. Genovese and L. Wasserman, *Operating characteristics and extensions of the false discovery rate procedure*, Journal of the Royal Statistical Society B **64** (2002), 499–518.

[9] C. Genovese and L. Wasserman, *Bayesian and Frequentist Multiple Testing*, Bayesian Statistics **7** (2003), 145–162. Oxford: Oxford University Press (J.M. Bernardo, M.J. Bayarri, J.O. Berger, A.P. Dawid, D. Heckerman, A.F.M. Smith, and M. West, eds.).

[10] R. Gopalan and D.A. Berry, *Bayesian multiple comparisons using Dirichlet process priors*, Journal of the American Statistical Association **93** (1998), 1130–1139.

[11] R.L. Keeney, H.A. Raiffa and R.F.C. Meyer, *Decisions With Multiple Objectives: Preferences and Value Tradeoffs*, New York: John Wiley & Sons(1976).

[12] D.V. Lindley, *Making decisions*, New York: Wiley, second edn (1971).

[13] P. Müller, G .Parmigiani, C. Robert and J. Rouseau, *Optimal sample size for multiple testing: the case of gene expression microarrays*, Journal of the American Statistical Association **99** (2004), 990-1001.

[14] M. Newton, A. Noueriry, D. Sarkar and P. Ahlquist, *Detecting differential gene expression with a semiparametric heirarchical mixture model*, Biostatistics **5** (2004), 155–176.

[15] M.A. Newton, C.M. Kendziorsky, C.S. Richmond, R., B. F. and K.W. Tsui, *On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data*, Journal Computational Biology **8** (2001), 37–52.

[16] J. Scott and J. Berger, *An exploration of aspects of bayesian multiple testing*, Journal of Statistical Planning and Inference **in press** (2006).

[17] J. Storey, *A direct approach to false discovery rates*, Journal of the Royal Statistical Society B **64** (2002), 479–498.

[18] J.D. Storey, *The positive false discovery rate: A Bayesian interpretation and the q-value*, The Annals of Statistics **31** (2003), 2013–2035.

[19] J.D. Storey, J.E. Taylor and D. Siegmund, *Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach*, Journal of the Royal Statistical Society Series B: Statistical Methodology **66** (2004), 187–205.

Bootstrap versions for tests based on residual empirical processes in nonparametric regression models

Natalie Neumeyer

Since a few decades in statistical research nonparametric regression models with independent observations,

$$Y_i = m(X_i) + \sigma(X_i)\varepsilon_i \quad (i = 1, \dots, n) \tag{1}$$

have been investigated intensively (here we assume X_i and ε_i to be independent and $E[\varepsilon_i] = 0$, $E[\varepsilon_i^2] = 1$ for the iid errors). Research focused mainly on nonparametric estimation of the regression function m and variance function σ^2 and corresponding hypotheses tests. Since a few years only there exist results on estimation of the smooth distribution F of the unobserved errors $\varepsilon_1, \dots, \varepsilon_n$. Weak convergence of the empirical process $\sqrt{n}(\hat{F}_n - F)$ based on nonparametrically estimated residuals

$$\hat{\varepsilon}_i = \frac{Y_i - \hat{m}(X_i)}{\hat{\sigma}(X_i)} \tag{2}$$

was shown by Akritas and Van Keilegom (1). Further, the empirical distribution function $\hat{F}_n$ of estimated errors $\hat{\varepsilon}_1, \dots, \hat{\varepsilon}_n$ recently turned out to be valuable for goodness-of-fit tests concerning the regression or variance function, see for instance (2), (9), (10). In all mentioned problems the asymptotic distribution of estimators and test statistics depends heavily on unknown features of the data generating

process such as the error density. In situations like these to circumvent problems with accuracy of the critical values resampling procedures such as bootstrap are applied. The presented work deals with different bootstrap approaches that can be used in the context of empirical processes based on nonparametric residuals. Based on a sample (1) bootstrap observations in residual bootstrap procedures are built as

$$Y_i^* = \hat{m}(X_i) + \hat{\sigma}(X_i)\varepsilon_i^* \quad (i = 1, \ldots, n). \tag{3}$$

It turns out that when using the empirical process based on residuals the classical residual bootstrap (i. e. to draw ε_i^* with replacement from standardized residuals) as often used in (homoscedastic) nonparametric regression (see, e. .g., Härdle and Bowman (3)) is not suitable because, given the original sample, then the bootstrap errors have a discrete distribution. However, in theory in our context the smoothness of the error distribution is crucial and the classical residual bootstrap leads to problems in theory and also does not work well in simulations. In heteroscedastic nonparametric regression models often wild bootstrap (see, e. .g., Härdle and Mammen (4)) is used. Here bootstrap errors ε_i^* are built by multiplying the residual $\hat{\varepsilon}_i$ by a bounded and centered random variable v_i, independent of the sample. But in the context of residual based empirical processes this wild bootstrap is not suitable in most problems because it changes the error distribution even asymptotically (in general $v_i\varepsilon_i$ has a different distribution than ε_i). We suggest to use a smooth residual boootstrap in the context of residual based empirical processes. To this end we build bootstrap observations as in (3), where the $\varepsilon_1^*, \ldots, \varepsilon_n^*$ are, given the sample (1), independent with a density f_n. Here f_n denotes a kernel density estimator of standardized versions of the residuals $\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n$ given by (2). In the context of homoscedastic linear models with fixed design a similar smooth residual bootstrap was considered by Koul and Lahiri (5).

The main result (see (6)) presented in the talk is the following. The empirical distribution function of residuals estimated from the smooth residual bootstrap sample centered by F_n (the bootstrap error distribution corresponding to density f_n) and multiplied by $\sqrt{n}$ converges weakly, given the original sample, to a centered Gaussian process, in probability, with the same covariance as given by the limit distribution of the residual empirical process built from the original sample.

The results about the smooth residual bootstrap version of the residual based empirical process have plenty of applications in model tests for nonparametric regression models.

In detail a test for equality of regression functions in two regression models is discussed that is based on a Wilcoxon rank statistic of residuals. The asymptotic distribution is presented (joint work with Holger Dette (8)) and the smooth residual bootstrap is applied and shown to be consistent. Some more applications of the smooth residual bootstrap in this context of regression model tests based on the residual empirical process are mentioned. In particular we show that the smooth residual bootstrap versions of the goodness-of-fit tests by Van Keilegom, Gonzalez Manteiga and Sanchez Sellero (10), Pardo Fernandez, Van Keilegom and Gonzalez Manteiga (9) and Dette and Van Keilegom (2) are consistent (as shown

in simulation studies but not in asymptotic theory by these authors). In more detail we consider the problem of testing the symmetry of the error distribution in a nonparametric regression model. We propose as a test statistic the difference between the two empirical distribution functions of estimated residuals and their counterparts with opposite signs. In this context the performance of a symmetric version of the smooth residual boostrap is discussed in asymptotic theory.

We also consider a symmetric wild bootstrap procedure. Here the bootstrap errors are defined as $\varepsilon_i^* = v_i\hat{\varepsilon}_i$ where the v_i are independent Rademacher variables. Then, $v_i\varepsilon_i$ always has a symmetric distribution. This symmetric wild bootstrap is consistent in testing symmetry of the error distibution as described above (joint work with Holger Dette (7)). In general, even when the error distribution is symmetric (such that ε_i and $v_i\varepsilon_i$ have the same distribution) the wild bootstrap version of the residual empirical process has a different limit distribution than the original version of this process. Therefore, it is not applicable in general in this context. However, for very specific testing problems and specific test statistics, the symmetric wild bootstrap can be used. This is the case for the Wilcoxon rank statistic for equality of regression functions as described above when the error distribution is known to be symmetric (see (7)).

References

[1] M. Akritas and I. Van Keilegom, *Nonparametric estimation of the residual distribution*, Scand. J. Statist. **28** (2001), 549–567.

[2] H. Dette and I. Van Keilegom, *A new test for the parametric form of the variance function in nonparametric regression*, preprint. http://www.rub.de/mathematik3/preprint.htm (2005).

[3] W. Härdle and A. W. Bowman, *Bootstrapping in nonparametric regression: Local adaptive smoothing and confidence bands*, J. Amer. Statist. Assoc. **83** (1988), 102–110.

[4] W. Härdle and E. Mammen, *Comparing nonparametric versus parametric regression fits*, Ann. Statist. **21** (1993), 1926–1947.

[5] H. L. Koul and S. N. Lahiri, *On Bootstrapping M-Estimated Residual Processes in Multiple Linear Regression Models*, J. Multivariate Analysis **49** (1994), 255–265.

[6] N. Neumeyer, *Bootstrap procedures for empirical processes based on nonparametric regression residuals*, Habilitationsschrift, Ruhr-Universität Bochum, work in progress.

[7] N. Neumeyer, H. Dette, *Testing for symmetric error distribution in nonparametric regression models*, http://www.rub.de/mathematik3/preprint.htm (2003). In revision for Statistica Sinica.

[8] N. Neumeyer, H. Dette, *A note on one-sided nonparametric analysis of covariance by ranking residuals*, Mathematical Methods of Statistics **14** (2005), 80–104.

[9] J. C. Pardo–Fernandez, I. Van Keilegom, W. González–Manteiga, *Comparison of regression curves based on the estimation of the error distribution*, preprint,

Institut de Statistique (Université catholique de Louvain), Louvain-la-Neuve. (2004).

[10] I. Van Keilegom, W. Gonzalez–Manteiga, C. Sanchez Sellero, *Goodness-of-fit tests in parametric regression based onestimation the error distribution*, preprint, Institut de Statistique (Université catholique de Louvain), Louvain-la-Neuve. (2004)

Significance and Recovery of Block Structures in Binary Matrices with Noise

Andrew Nobel

(joint work with Xing Sun)

1. Introduction

A number of common and well-studied data mining problems can be viewed as special cases of a more general problem, namely that of finding distinguished submatrices in a rectangular data matrix. These include frequent itemset and market basket analyses and bi-clustering (also known as subspace clustering). In the frequent itemset problem (*cf.* (4) (5)), the available data is described by a collection $S = \{s_1, \ldots, s_n\}$ of items and a set $T = \{t_1, \ldots, t_m\}$ of transactions. Each transaction t_i is associated with a subset of S, which can be thought of as a shopping list without multiplicity. The transaction database can naturally be expressed as an $m \times n$ binary matrix $\mathbf{Y} = \{y_{i,j}\}$ with $y_{i,j} = 1$ if transaction t_i contains item s_j and $y_{i,j} = 0$ otherwise. Frequent itemset mining (FIM) algorithms identify every submatrix of 1s in $\mathbf{X}$ having a minimum number of rows. (The set of items associated with the columns of the submatrix is then said to be frequent).

Here we consider a number of statistical issues related to frequent itemset mining, including (i) the significance of the submatrices found by FIM, (ii) the effects of noise on standard FIM algorithms, and (iii) the ability of noise tolerant algorithms to recover frequent itemsets in presence of noise.

2. Significance of Bi-clusters

Let $\{z_{i,j} : i, j \geq 1\}$ be an infinite array of independent Bernoulli random variables, $P(z_{i,j} = 1) = p = 1 - P(z_{i,j} = 0)$. For each $n \geq 1$, let $\mathbf{Z}_n = \{z_{i,j} : 1 \leq i, j \leq n\}$ be the $n \times n$ upper left corner of the infinite array, which we will write as $\mathbf{Z}_n \sim \text{Bern}(p)$. Let $M(\mathbf{Z}_n)$ be the largest k such that $\mathbf{Z}_n$ contains a $k \times k$ submatrix of 1's. The stochastic behavior of $M(\mathbf{Z}_n)$ enables us to study FIM under the null hypothesis that the available data is random, and does not contain any structure. To this end, define $s_0(n)$ to be any solution of the equation

$$1 = \phi_n(s) = (2\pi)^{-\frac{1}{2}}\, n^{n+\frac{1}{2}}\, s^{-s-\frac{1}{2}}\, (n-s)^{-(n-s)-\frac{1}{2}}\, p^{\frac{s^2}{2}} \tag{1}$$

over $s \in \mathbb{R}^+$. Routine but involved calculations show that, when n is sufficiently large, $s_0(n)$ is uniquely defined and that

$$s(n) \;=\; 2\,\log_b n \;-\; 2\,\log_b \log_b n + C + o(1), \tag{2}$$

where C is a positive constant and $b = p^{-1}$. For positive integer k, $\phi_n(k)$ is the Stirling approximation to EU_k, where U_k is the number of $k \times k$ submatrices of ones in $\mathbf{Z}_n$. Define $k(n) = \lceil s(n) \rceil$. An application of the first moment method yields a useful deviation inequality for $M(\mathbf{Z}_n)$. Related work can be found in Tanay *et al.* (13) and Koyutürk *et al.* (10).

Proposition 1. *Fix $0 < \gamma < 1$. When n is sufficiently large, $P\{M(\mathbf{Z}_n) \geq k(n) + k\} \leq 2\,n^{-2\,k}\,(\log_b n)^{3k}$ for any $1 \leq k \leq \gamma\,n$ and $k \in \mathbb{Z}^+$.*

One may readily generalize Proposition 1 to obtain similar results for non-square submatrices and non-square matrices. Viewing $\mathbf{Z}_n$ as the adjacency matrix of a bi-partite graph G, it can readily be seen that $M(\mathbf{Z}_n)$ is the size of the largest bi-clique in G. Following the work of Bollobás and Erdős (7) and Matula (12) on the size of maximal cliques in random graphs, one may establish a three point concentration result for the asymptotic behavior of $M(\mathbf{Z}_n)$.

Theorem 1. *With probability one, $|M(\mathbf{Z}_n) - k(n)| < \frac{3}{2}$ when n is sufficiently large.*

Theorem 1 extends earlier work of Dawande *et al.* (8), who showed that $P(\log_b n \leq M(\mathbf{Z}_n) \leq 2\log_b n) \to 1$.

3. Recoverability

Standard frequent itemset algorithms do not explicitly account for random noise in their search for submatrices of 1s. In order to study the potential effects of noise on FIM, we consider the simple statistical model

$$\mathbf{Y} \;=\; \mathbf{X} \oplus \mathbf{Z} \tag{3}$$

under which the observed $n \times n$ data matrix $\mathbf{Y}$ is equal to the modulo 2 sum of an unobserved "true" data matrix $\mathbf{X}$, plus random noise $\mathbf{Z} \sim \mathrm{B}ern(p)$ with $0 < p < 1/2$. Under the model (3), each entry $y_{i,j}$ is the modulo 2 sum of $x_{i,j}$ and $z_{i,j}$ The next result follows readily from Proposition 1 and a simple coupling argument.

Proposition 2. *Let $b' = (1-p)^{-1}$. With probability one, when n is sufficiently large, $M(\mathbf{Y}) \leq 2\log_{b'} n$ regardless of the values in $\mathbf{X}$.*

In particular, even if the unobserved matrix $\mathbf{X}$ contains interesting structure (*e.g.*, a large submatrix of 1s), frequent itemset mining algorithms which look for submatrices of 1s in $\mathbf{Y}$ will not detect this structure. It is natural then to consider noise tolerant criteria for frequent itemsets that may be more successful at recovering structure in (3) and related models. A number of noise tolerant criteria and related algorithms have been proposed in the data mining literature (1; 2; 3). Here we consider the approximate frequent itemset (AFI) criterion

proposed in (11). Let $0 < \tau < 1$. A submatrix C of $\mathbf{Y}$ is a τ-AFI, denoted $C \in \mathrm{AFI}_\tau(\mathbf{X})$, if every row and column of C has at least $100\,\tau\%$ ones.

Suppose that $\mathbf{X}$ is $n \times n$ and consists of an $l \times l$ submatrix C of 1s, with all other entries equal to 0. (The rows and columns of C need not be contiguous.) Suppose that we observe $\mathbf{Y}$ in (3), and wish to accurately recover C. Let p_0 be any number such that $p < p_0 < 1/2$, and let $\tau = 1 - p_0$ be an associated error threshold. Let $\mathcal{C}$ be the family of all square submatrices C' of $\mathbf{X}$ such that $C' \in \mathrm{AFI}_\tau(\mathbf{X})$, and define

$$\hat{C} \;=\; \mathrm{argmax}_{C' \in \mathcal{C}} |C'|$$

to be any maximal sized submatrix in $\mathcal{C}$. Note that $\hat{C}$ depends only on the observed matrix $\mathbf{Y}$. Let

$$\Lambda \;=\; |\hat{C} \cap C| / |\hat{C} \cup C|$$

measure the overlap between the estimated index set $\hat{C}$ and the true index set C. Then $0 \le \Lambda \le 1$, and values of Λ close to one indicate better overlap.

Theorem 2. *When n is sufficiently large, for any $\alpha > 0$ such that $16\alpha^{-1}(\log_b n + 2) < l$ we have*

$$P\left(\Lambda \le \frac{1-\alpha}{1+\alpha}\right) \;\le\; 2\exp\left\{-\frac{3l(p-p_0)^2}{4}\right\} \;+\; 2n^{-\frac{l}{4}\alpha l - 4\log_b n},$$

where the logarithms are to the base $b = \exp\{3(1-2\,p_0)^2/8p\}$.

The conditions of Theorem 2 require that the noise level $p < 1/2$ and that the user-specified parameter p_0 satisfy $p < p_0 < 1/2$. In advance, one only needs to know an upper bound on the noise level p.

4. Acknowledgement

A.B. Nobel's work was partly supported by US NSF Grant DMS 0406361.

References

[1] J. Pei, G. Dong, W. Zou, J. Jan, *Mining Condensed Frequent-Pattern Bases*, Knowledge and Information Systems Volume **6** Issue 5 (2002).

[2] C. Yang, U. Fayyad, P.S. Bradley, *Efficient discovery of error-tolerant frequent itemsets in high dimensions*, SIGKDD (2001).

[3] M. Steinbach, P.-N. Tan, V. Kumar, *Support envelopes: a technique for exploring the structure of association patterns*, SIGKDD (2004).

[4] R. Agrawal, T. Imielinski and A. Swami, *Mining association rules between sets of items in large databases*, SIGMOD (1993).

[5] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen and A. Verkamo, *Fast discovery of association rules*, U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy, editors, Advances in Knowledge Discove and Data Mining, chapter **12** 307–328. AAAI Pre (1996).

[6] B. Bollobás, *Random Graphs*, second edition, Cambridge Univ. Pre (2001).

[7] B. Bollobás, P. Erdős, *Cliques in random graphs*, Mathematical Proceedings of the Cambridge Philosophical Society **80** (1976), 419–427.

[8] M. Dawande, P. Keskinocak, J. M. Swaminathan, S. Tayur, *On Bipartite and Multipartite Clique Problems*, J. Algorithms **41(2)** (2001), 388-403.
[9] J. Han, J. Pei and Y.Yin, *Mining Frequent Patterns without Candidate Generation*, SIGMOD (2000).
[10] M. Koyutürk, W. Szpankowski, A. Grama, *Biclustering Gene-Feature Matrices for Statistically Significant Dense Patterns*, CSB (2004), 480-484.
[11] J.Liu, P. Susan, W. Wang, A.B. Nobel and J. Prins, *Mining Approximate Frequent Itemset from Noisy Data*, ICDM (2005).
[12] D. Matula, *The largest clique size in a random graph*, Southern Methodist University, Tech. Report, CS 7608 (1976).
[13] A. Tanay, R. Sharan, R. Shamir, *Dicovering statistically significant biclusters in gene expression data*, Bioinformatics Volume **18**.

A Synthesis and Unification of (Objective) Bayes Factors for Model Selection and Hypothesis Testing

Luis Raúl Pericchi

The basics of the Bayesian approach to model selection are presented and compared in Pericchi (2005). Eight objective methods of developing default Bayesian approaches that have undergone considerable recent development are reviewed and analyzed in a general framework:

- Well Calibrated Priors (WCP)
- Conventional Priors (CP)
- Intrinsic Bayes Factor (IBF)
- Intrinsic Priors (IPR)
- Expected Posterior Priors (EP)
- Fractional Bayes Factor (FBF)
- asymptotic methods and the Bayesian Information Criterion (BIC)
- Lower Bounds (LB) on Bayes Factors.

These approaches are illustrated and commented on how to use and how *not* to use them. Despite the apparent inordinate multiplicity of methods, there are important connections and similarities among different Bayesian methods. Most important, typically the results obtained by any of the methods, are closer among themselves than to results from non-Bayesian methods, and this is typically more so as the information accumulates.

1 Overview

The unifying concepts for Bayes Factors are mainly the following:

1. Training Samples: Real, simulated or imaginary

2. Predictively Matched Priors
3. Intrinsic Priors
4. Bayesian Principle: "Correspondance of Methods with *Sensible Priors".*
5. Asymptotics and Consistencies of Bayes Factors:
   (Large Sample) Consistency as the sample size accumulates for fixed number parameters.
   (Finite Sample) Consistency unter fixed sample size and fixed number of parameters as increases the distance of the sampling model to a candidate model.

2 Conclusions

- Here 8 methods are analyzed for calculation or approximation of Bayes Factors. These may be taken as a proof of widespread disagreement. However, there is deep source of agreements: often these methods share the same asymptotics and give results which are close to each other, apart from deep theoretical connections among them. On the other hand, frequentist methods, like p-values or significance testing with fixed type I errors, have different asymptotics, and thus are increasingly at odds with Bayesian methods, by an increasing amount as the data accumulates.
- There are general concepts which offer a unifying framework:
  i) **Principle 1:** *Testing and model selection methods should correspond, in some sense, to actual Bayes factors, arising from reasonable default prior distributions*, and Theorem 1 in Berger and Pericchi (1996), stating that the Intrinsic Prior arising from the Arithmetic IBF, is proper or conditionally proper under absolutely continuous distributions and mild conditions. ii) The different kinds of well calibrated priors. iii) The so called **Assumption 0** in Berger and Pericchi (2004), i.e. for all models and all parameter values the probability of the set of training samples under consideration should be unity. iv) The concept of Intrinsic Priors for different methods. v) The concept of training sample, real and imaginary; deterministic and aleatory.
- It should be remembered that a unifying view is that Bayes Factors can be seen as "*Un-Normalized Bayes Factors* $\times$ *Correction Factors*". The methods discussed here are mostly about the second right hand term which is **not** the dominant asymptotic factor, but that should be given careful consideration. Still all methods considered here have the first (or an approximation of it) factor embedded in their formulae. So the methods discussed here, evolving around the Un-Normalized Bayes Factor share a dominant (asymptotically) common term.

Bayes Factor is a powerful probabilistic tool, which is here to stay. We rather learn how to use it at its best. It is time to emphasize the agreements among the Bayesian approaches to form and approximate Bayes Factors. The 8 approaches

visited here are most of the time able to give sensible results if we avoid potentially harmful priors, like vague proper priors. Specially in new problems, it is advisable to compare several of them to check the reassuring agreement, a sort of robustness with respect to the method.

References

[1] J.O. Berger and L.R. Pericchi, *The Intrinsic Bayes Factor for Model Selection and Prediction*, Journal of the American Statistical Association **91** 433 (1996), 109–122.

[2] J.O. Berger and L.R. Pericchi, *Objective Bayesian Methods for Model Selection: Introduction and Comparison (with discussion)*, in Lahiri P. (Ed.) IMS Lecutre Notes - Monograph Series Volume **38** (2001), 135–207.

[3] J.O. Berger and L.R. Pericchi, *Training samples in objective Bayesian model selection*, Annals of Statistics **32 3** (2004), 841–869.

[4] L.R. Pericchi, *Model Selection and Hypothesis Testing based on Objective Probabilities and Bayes Factors*, Elsevier B.V. Handbook of Statistics Volume **25** (2005) in press.

Prior selection and model choice

Christian P. Robert

(joint work with J.A. Cano, J.M. Marin and D. Salmeró)

1. Priors for Bayesian testing and model selection

The selection of prior distributions is always an issue in Bayesian Statistics but it gets particularly crucial when dealing with model choice, because of a variety of sensitive issues that are detailed in Robert (6). One of these issues is that the central tool of Bayesian model choice, the Bayes factors

$$B_{12} = \frac{\Pr(\mathcal{M}_1|x)}{\Pr(\mathcal{M}_2|x)} \Big/ \frac{\Pr(\mathcal{M}_1)}{\Pr(\mathcal{M}_2)}$$

$$= \frac{\int f_1(x|\theta_1)\pi_1(\theta_1)\mathrm{d}\theta_1}{\int f_2(x|\theta_2)\pi_2(\theta_2)\mathrm{d}\theta_2}$$

that are used to compare models $\mathfrak{M}_1$ vs. $\mathfrak{M}_2$, is not compatible with improper priors $\pi_1(\theta_1)$ or $\pi_2(\theta_2)$. This difficulty is most troublesome in that reference (or noninformative) priors are usually improper. Moreover, using vague proper priors, that is, proper priors with large variances, is not an acceptable solution in that the dependence on the variance most often does not vanish as the variance goes to infinity or leads to $0, 1$ solutions because of Lindley's paradox Robert (6, Chapter 5).

There have been many attempts in the Bayesian literature to overcome this difficulty, most of those proposing a particular protocol to transform improper priors into proper priors. A first solution is to use *training samples*, that is, a part $x_{[i]}$ of the data x to make the prior proper in the sense that $\pi_i(\cdot|x_{[i]})$ is proper and thus

$$\frac{\int f_i(x_{[n/i]}|\theta_i)\ \pi_i(\theta_i|x_{[i]})\mathrm{d}\theta_i}{\int f_j(x_{[n/i]}|\theta_j)\ \pi_j(\theta_j|x_{[i]})\mathrm{d}\theta_j}$$

is independent of normalizing constants, where $x_{[n/i]}$ denotes the remaining part of the data. This defines a new Bayes factor

$$\begin{aligned} B_{12}(x_{[n/i]}) &= \frac{\int f^1_{[n/i]}(x_{[n/i]}|\theta_1)\pi_1(\theta_1|x_{[i]})\mathrm{d}\theta_1}{\int f^2_{[n/i]}(x_{[n/i]}|\theta_2)\pi_2(\theta_2|x_{[i]})\mathrm{d}\theta_2} \\ &= \frac{\int f_1(x|\theta_1)\pi_1(\theta_1)\mathrm{d}\theta_1}{\int f_2(x|\theta_2)\pi_2(\theta_2)\mathrm{d}\theta_2}\ \frac{\int \pi_2(\theta_2)f^2_{[i]}(x_{[i]}|\theta_2)\mathrm{d}\theta_2}{\int \pi_1(\theta_1)f^1_{[i]}(x_{[i]}|\theta_1)\mathrm{d}\theta_1} \\ &= B^N_{12}(x)B_{21}(x_{[i]}) \end{aligned}$$

that is more appropriately called *pseudo-Bayes factor* because of the replacement of the prior π with the posterior $\pi_i(\cdot|x_{[i]})$. Since this pseudo-factor depends on $x_{[i]}$, several ways of combining the quantities $B_{12}(x_{[n/i]})$ have been proposed, including for instance the arithmetic intrinsic Bayes factor of Berger and Pericchi (1; 2). However, these solutions are often lacking a true Bayesian nature in that the corresponding quantities usually are not Bayes factors for any prior pair (π_1, π_2).

2. Expected posterior priors

A solution found in Berger and Perez (5) is to avoid using real observations by relying instead on imaginary observations that are integrated against a reference measure. Starting with an improper prior π_1 of interest, the *expected posterior prior* is defined as

$$\pi^*_1(\theta) = \int \pi_1(\theta|x)\, m(x)\, \mathrm{d}x\,,$$

where $m(x)$ is the reference measure, often derived from a reference model $\mathfrak{M}_0$ and a reference prior as

$$m(x) = \int f_0(x|\theta)\, \pi_0(\theta)\mathrm{d}\theta\,.$$

Obviously, the selection of this reference measure is central to the construction of the expected posterior prior and it is often the case that there is neither reference measure nor reference model $\mathfrak{M}_0$ that everyone would agree on. In that setting, we may thus have two models $\mathfrak{M}_i$ $(i = 1, 2)$ that are *equally* valid and equiped with ideal reference priors π^N_i. In that case, we can either consider solving the

equation

$$\pi_1(\theta_1) = \int_{\mathscr{X}} \pi_1^N(\theta_1 \mid x) m_2(x) \mathrm{d}x$$

where m_2 is the marginal associated with π_2^N, or solving

$$\pi_2(\theta_2) = \int_{\mathscr{X}} \pi_2^N(\theta_2 \mid x) m_1(x) \mathrm{d}x,$$

where $m-1$ is the marginal associated with π_1^N. Since neither model is to be prefered, we propose in Cano, Robert and Salmeron (3) to iterate the process, namely to solve the system of integral equations

$$\pi_1(\theta_1) = \int_{\mathscr{X}} \pi_1^N(\theta_1 \mid x) m_2(x) \mathrm{d}x$$

and

$$\pi_2(\theta_2) = \int_{\mathscr{X}} \pi_2^N(\theta_2 \mid x) m_1(x) \mathrm{d}x,$$

where x is an imaginary minimal training sample and m_1, m_2 are the marginals associated with π_1 and π_2 respectively. that yield "true" marginals balancing each model wrt the other.

A sufficient condition for the existence of these symmetrised expected posterior priors is as follows: when both the observations and the parameters in both models are continuous, if the Markov chain with transition

$$Q\left(\theta_1' \mid \theta_1\right) = \int g\left(\theta_1, \theta_1', \theta_2, x, x'\right) dx\, dx'\, d\theta_2$$

where

$$g\left(\theta_1, \theta_1', \theta_2, x, x'\right) = \pi_1^N\left(\theta_1' \mid x\right) f_2\left(x \mid \theta_2\right) \pi_2^N\left(\theta_2 \mid x'\right) f_1\left(x' \mid \theta_1\right),$$

is *recurrent,* then there exists a solution to the integral equations, unique up to a multiplicative constant.

This result is interesting both from a theoretical point of view and from a computational point of view since, when the symmetrised expected posterior priors cannot be found, the Bayes factor can be approximated by MCMC simulation Casella and Robert (7).

3. Compatible prior

From another perspective, when dealing with multiple models $\mathfrak{M}_i$ $(i \in \mathfrak{I})$, it is fairly unrealistic to expect the priors on all models to be determined simultaneously from a subjective point of view. Rather, a single prior on an encompassing model should be enough to derive the other priors from a coherence principle as in, for instance Dawid and Lauritzen (4).

A rudimentary version of this coherence principle is to select the (sub)prior of a (sub)model $\mathfrak{M}_2$, given a prior π_1 on a model $\mathfrak{M}_1$, is as follows: we look for a

prior π_2 on $\mathfrak{M}_2$ which achieves the minimum Kullback divergence between the corresponding marginals:

$$m_1(x;\pi_1) = \int_{\Theta_1} f_1(x|\theta)\pi_1(\theta)\mathrm{d}\theta$$

and

$$m_2(x);\pi_2 = \int_{\Theta_2} f_2(x|\theta)\pi_2(\theta)\mathrm{d}\theta\,,$$

that is,

$$\pi_2 = \arg\min_{\pi_2} \int \log\left(\frac{m_1(x;\pi_1)}{m_2(x;\pi_2)}\right) m_1(x;\pi_1)\,\mathrm{d}x\,.$$

Obviously, this principle also has its limitations, in that for instance it cannot be applied to improper priors π_1 and, more importantly, ends up in a Dirac mass if no restriction is put on π_2. But it can be of use in setups where conjugate priors are relevant. For instance, if both $\mathfrak{M}_1$ and $\mathfrak{M}_2$ are two nested Gaussian linear regression models with the same variance $\sigma^2 \sim \pi(\sigma^2)$, a conjugate prior on $\mathfrak{M}_1$

$$y|\beta_1,\sigma^2 \sim \mathcal{N}(X_1\beta_1,\sigma^2)$$

where X_1 is a $(n \times k_1)$ matrix of rank $k_1 \leq n$, is Zellner's g-prior (8),

$$\beta_1|\sigma^2 \sim \mathcal{N}\left(s_1, \sigma^2 n_1 (X_1^{\mathrm{T}} X_1)^{-1}\right)\,.$$

If we restrict π_2 to be also a conjugate prior on the subvector β_2, i.e.

$$\beta_2|\sigma^2 \sim \mathcal{N}\left(s_2, \sigma^2 n_2 (X_2^{\mathrm{T}} X_2)^{-1}\right),$$

where X_2 is a $(n \times k_2)$ submatrix of X_1, the prior that minimize the Kullback-Leibler divergence between the two marginal distributions conditional on σ^2 is

$$\beta_2|X_2,\sigma^2 \sim \mathcal{N}\left(s_2^*, \sigma^2 n_2^* (X_2^{\mathrm{T}} X_2)^{-1}\right)$$

with

$$\begin{aligned} s_2^* &= (X_2^{\mathrm{T}} X_2)^{-1} X_2^{\mathrm{T}} X_1 s_1 \\ n_2^* &= n_1\,. \end{aligned}$$

In the particular case of variable selection, when dealing with a set $\{x_1,\ldots,x_p\}$ of p potential explanatory regressors (plus intercept), there are 2^p submodels $\mathfrak{M}_\gamma$, where $\gamma \in \Gamma = \{0,1\}^p$ indicates inclusion/exclusion of those variables by a binary representation. Then, if q_γ is the number of variables that are included in $\mathfrak{M}_\gamma$, if $t_1(\gamma) = \{t_{1,1}(\gamma),\ldots,t_{1,q_\gamma}(\gamma)\}$ are the indices of those variables and $t_0(\gamma)$ the indices of the variables not included, $\mathfrak{M}_\gamma$ can be represented as

$$y|\beta,\gamma,\sigma^2 \sim \mathcal{N}\left(X_{t_1(\gamma)}\beta_{t_1(\gamma)}, \sigma^2 I_n\right)\,.$$

Therefore, if we use Zellner's g-prior, i.e. a normal prior for β conditional on σ^2,

$$\beta|\sigma^2 \sim \mathcal{N}(\tilde{\beta}, c\sigma^2 (X^{\mathrm{T}} X)^{-1})$$

and a Jeffreys prior for σ^2,

$$\pi(\sigma^2) \propto \sigma^{-2}\,,$$

for the full model, the compatible prior is

$$\mathcal{N}\left(\left(X_{t_1(\gamma)}^{\mathrm{T}}X_{t_1(\gamma)}\right)^{-1}X_{t_1(\gamma)}^{\mathrm{T}}X\tilde{\beta}, c\sigma^2\left(X_{t_1(\gamma)}^{\mathrm{T}}X_{t_1(\gamma)}\right)^{-1}\right).$$

References

[1] J. Berger and L. Pericchi, *The intrinsic Bayes factor for model selection and prediction*, J. American Statist. Assoc. **91** (1996), 109–122.

[2] J. Berger and L. Pericchi, *Objective Bayesian methods for model selection: introduction and comparison*, In P. Lahiri, editor, Model Selection, Beachwood Ohio, Institute of Mathematical Stastistics, Lecture Notes - Monograph Series Volume **38** (2001), 135–207.

[3] S.D. Cano, J.A. Salmeron and C. Robert, *Technical Reports*.

[4] A. Dawid and S. Lauritzen, *Compatible prior distribution* In E. George, editor, Bayesian Methods with Application to Science Policy and Official Statistics, The sixth Wold meeting of the ISBA (2000), 109–118.

[5] J.M. Pérez and J. Berger, *Expected posterior prior distributions for model selection*, Biometrika **89** (2002), 491–512.

[6] C. Robert, *The Bayesian Choice*, Springer-Verlag, New York, second editon (2001).

[7] G.Casella and C. Robert, *Monte Carlo Statistical Methods*, Springer-Verlag, New York, second edition (2004).

[8] A. Zellner, *On assessing Prior Distributions and Bayesian Regression analysis with g-prior distribution regression using Bayesian variable selection*, Bayesian inference and decision techniques: Essays in Honor of Bruno de Finetti, North-Holland/Elsevier (1986), 233–243.

Local Parametric Methods in Nonparametric Regression

Vladimir Spokoiny

A new approach to nonparametric estimation is discussed. The method is based on the extension of the parametric maximum likelihood principle to the nonparametric situation. The method leads to the "oracle" estimation quality and includes the parametric situation as a special case.

Mirror averaging, aggregation and model selection

Alexandre Tsybakov
(joint work with Anatoli Juditsky, Philippe Rigollet)

Several problems in statistics and machine learning can be stated as follows: given a collection of M different estimators (classifiers), construct a new estimator (classifier) which is nearly as good as the best among them with respect to a given

risk criterion. This target is called model selection (MS) type aggregation, and it can be described in terms of the following stochastic optimization problem.

Let $(\mathcal{Z}, \mathfrak{F})$ be a measurable space and let Θ be the simplex

$$\Theta = \left\{ \theta \in \mathbb{R}^M : \sum_{j=1}^{M} \theta^{(j)} = 1,\ \theta^{(j)} \geq 0,\ j = 1, \ldots, M \right\}.$$

Here and throughout the paper we suppose that $M \geq 2$ and we denote by $z^{(j)}$ the jth component of a vector $z \in \mathbb{R}^M$. We denote by $\left[z^{(j)}\right]_{j=1}^{M}$ the vector $z = (z^{(1)}, \ldots, z^{(M)})^\top \in \mathbb{R}^M$.

Let Z be a random variable with values in $\mathcal{Z}$. The distribution of Z is denoted by P and the corresponding expectation by E. Suppose that P is unknown and that we observe n i.i.d. random variables $Z_1, \ldots, Z_n$ with values in $\mathcal{Z}$ having the same distribution as Z. The distribution (respectively, expectation) w.r.t. the sample $Z_1, \ldots, Z_n$ is denoted by P_n (respectively, by E_n).

Consider a measurable function $Q : \mathcal{Z} \times \Theta \to \mathbb{R}$ and the corresponding average risk function

$$A(\theta) = EQ(Z, \theta)\,,$$

assuming that this expectation exists for all $\theta \in \Theta$. Stochastic optimization problems that are usually studied in this context consist in minimization of A on some subsets of Θ, given the sample $Z_1, \ldots, Z_n$. Note that since the distribution of Z is unknown, direct (deterministic) minimization of A is not possible.

For $j \in \{1, \ldots, M\}$, denote by e_j the jth coordinate unit vector in $\mathbb{R}^M$: $e_j = (0, \ldots, 0, 1, 0, \ldots, 0) \in \mathbb{R}^M$, where 1 appears in jth position.

The stochastic optimization problem associated to MS aggregation is

$$\min_{\theta \in \{e_1, \ldots, e_M\}} A(\theta).$$

The aim of MS aggregation is to "mimic the oracle" $\min_{1 \leq j \leq M} A(e_j)$, i.e. , to construct an estimator $\tilde{\theta}_n$ measurable w.r.t. $Z_1, \ldots, Z_n$ and called aggregate, such that

$$E_n A(\tilde{\theta}_n) \leq \min_{1 \leq j \leq M} A(e_j) + \Delta_{n,M}\,, \tag{1}$$

where $\Delta_{n,M} > 0$ is a remainder term that should be as small as possible. Lower bounds can be established showing that, under some assumptions, the smallest possible value of $\Delta_{n,M}$ in a minimax sense has the form

$$\Delta_{n,M} = \frac{C \log M}{n}\,, \tag{2}$$

with some constant $C > 0$ [cf. Tsybakov (2003)].

Besides being in themselves precise finite sample results, oracle inequalities of the type (1) are very useful in adaptive nonparametric estimation. They allow one to prove that the aggregate estimator $\tilde{\theta}_n^\top H$ is adaptive in a minimax asymptotic sense (and even sharp minimax adaptive is several cases: for more discussion see, e.g., Nemirovski (2000)).

The aim of this paper is to obtain bounds of the form (1) – (2) under some general conditions on the loss function Q. For two special cases (density estimation with the Kullback-Leibler (KL) loss, and regression model with squared loss) such bounds has been proved earlier in the benchmark works of Catoni (2004) and Yang (2000). They independently obtained the bound for density estimation with the KL loss, and Catoni (2004) solved the problem for the regression model with squared loss. Bunea and Nobel (2005) suggested another proof of the regression result of Catoni (2004) improving it in the case of bounded response, and obtained some inequalities with suboptimal remainder terms under weaker conditions.

Here we study the recursive aggregate $\hat{\theta}_n$ which is defined in the following way. Set $\beta > 0$, define the vector

$$u_i \triangleq \Big(Q(Z_i, e_1), \ldots, Q(Z_i, e_M)\Big)^\top$$

and consider the iterations:

- Fix the initial values $\theta_0 \in \Theta$ and $\zeta_0 = 0 \in \mathbb{R}^M$.
- For $i = 1, \ldots, n-1$, do the recursive update

$$\begin{aligned} \zeta_i &= \zeta_{i-1} + u_i \,, \\ \theta_i &= \left[\frac{\mathrm{e}^{-\zeta_i^{(j)}/\beta}}{\sum_{k=1}^M \mathrm{e}^{-\zeta_i^{(k)}/\beta}}\right]_{j=1}^M . \end{aligned} \tag{3}$$

- Output at iteration n the average

$$\widehat{\theta}_n = \frac{1}{n}\sum_{i=1}^n \theta_{i-1} \,. \tag{4}$$

Note that $\hat{\theta}_n$ is measurable w.r.t. the subsample $(Z_1, \ldots, Z_{n-1})$. Recursions (3) – (4) constitute a special case of the mirror averaging algorithm of Juditsky, Nazin, Tsybakov and Vayatis (2005). For particular choices of β and Q, it yields the methods described by Catoni (2004) and Yang (2000). We prove the following results.

Theorem 1. *Let B be a measurable subset of $\mathcal{Z}$. Assume that $\beta > 0$ is such that the mapping $\theta \mapsto \exp(-Q(z,\theta)/\beta)$ is concave on the simplex Θ, for all $z \in B$. Assume also that there exists two functions $L_Q(\cdot)$ and $R_Q(\cdot)$ on $\mathcal{Z} \setminus B$, with values in $\mathbb{R}$ such that for all $z \in \mathcal{Z} \setminus B$ and all $\theta \in \Theta$ we have $L_Q(z) \le Q(z,\theta) \le R_Q(z)$. Then the aggregate $\hat{\theta}_n$ satisfies, for any $M \ge 2, n \ge 1$, the following oracle inequality*

$$E_{n-1} A(\hat{\theta}_n) \le \min_{1 \le j \le M} A(e_j) + \frac{\beta \log M}{n} + E\big[\big(R_Q(Z) - L_Q(Z)\big)\mathbb{1}_{\{Z \notin B\}}\big] \,,$$

where $\mathbb{1}_{\{\cdot\}}$ denotes the indicator function.

Theorem 2. *Let Q_1 be the function on $\mathcal{Z}\times\Theta\times\Theta$ defined by $Q_1(z,\theta,\theta') = Q(z,\theta) - Q(z,\theta')$ for all $z \in \mathcal{Z}$ and all $\theta,\theta' \in \Theta$. Assume that for some $\beta > 0$ there exists a Borel function $\Psi_\beta : \Theta \times \Theta \to \mathbb{R}_+$ such that the mapping $\theta \mapsto \Psi_\beta(\theta,\theta')$ is concave on the simplex Θ for any fixed $\theta' \in \Theta$, $\Psi_\beta(\theta,\theta) = 1$ and $E\exp(-Q_1(Z,\theta,\theta')/\beta) \le \Psi_\beta(\theta,\theta')$ for all $\theta,\theta' \in \Theta$. Then the aggregate $\hat{\theta}_n$ satisfies, for any $M \ge 2, n \ge 1$, the following oracle inequality*

$$E_{n-1}A(\hat{\theta}_n) \le \min_{1\le j\le M} A(e_j) + \frac{\beta \log M}{n}\,.$$

We show that the assumptions of Theorems 1 and 2 are fulfilled for several statistical models including regression, classification and density estimation. This allows one to construct in an easy way sharp adaptive nonparametric estimators for the above mentioned statistical problems.

References

[1] F. Bunea and A. Nobel, *Sequential procedures for aggregating arbitrary estimators of a conditional mean* (2005). Manuscript. `http://www.stat.fsu.edu/~flori`.

[2] O. Catoni, *Statistical Learning Theory and Stochastic Optimization. Ecole d'Eté de Probabilités de Saint-Flour XXXI - 2001*, Lecture Notes in Mathematics vol. **1851** (2004), Springer, New York.

[3] A. Juditsky, A. Nazin, A. Tsybakov and N. Vayatis, *Recursive aggregation of estimators via the mirror descent algorithm with averaging*, Problems of Information Transmission **41** (2005), n.4. `www.proba.jussieu.fr/pageperso/vayatis/publication.html`.

[4] A. Nemirovski, *Topics in Non-parametric Statistics*, Ecole d'Eté de Probabilités de Saint-Flour XXVIII - 1998, Lecture Notes in Mathematics vol. **1738** (2000), Springer, New York.

[5] A. Tsybakov, *Optimal rates of aggregation*, Computational Learning Theory and Kernel Machines(B.Schölkopf and M.Warmuth, eds.), Lecture Notes in Artificial Intelligence vol. **2777** (2003), Springer, Heidelberg, 303–313.

[6] Y. Yang, *Mixing strategies for density estimation*, Ann. Statist. **28** (2000), 75–87.

Higher Order Estimating Equations for Causal Inference

Aad van der Vaart

(joint work with Lingling Li, James Robins, Eric Tchetgen)

The general purpose of this talk is to introduce a new type of estimating equations that can cope with high-dimensional covariates. Based on a random sample $X_1,\ldots,X_n$ from a density p_η with respect to some measure on a sample space we wish to estimate a real-valued parameter which can be written as the value $\theta = \chi(\eta)$ of a function $\eta \mapsto \chi(\eta)$ on a given infinite-dimensional parameter space.

Such a problem has been considered in the 1980/90s within the context of semiparametric statistical models, where it was shown that for a variety of functions χ it is possible to estimate the parameter θ at rate of precision $n^{-1/2}$ with a normal limiting distribution for the scaled difference $\sqrt{n}(\hat{\theta}-\theta)$ of the estimator $\hat{\theta}$ and estimand θ. (See e.g. Van der Vaart (1998).) One method of estimator construction is through estimating equations. In the case that the parameter η can be partitioned as $\eta = (\theta, \gamma)$ into the parameter of interest and a nuisance parameter, such estimating equations take the form

$$\sum_{i=1}^{n} \psi_{\theta,\hat{\gamma}}(X_i) = 0,$$

for a preliminary estimator $\hat{\gamma}$ of the nuisance parameter, and suitable measurable functions $x \mapsto \psi_{\theta,\gamma}(x)$. The estimator $\hat{\theta}$ is defined as a (near) solution to this equation.

The preliminary estimator $\hat{\gamma}$ must satisfy certain conditions in order for the method to work, i.e. to ensure the asymptotic normality of the sequence $\sqrt{n}(\hat{\theta}-\theta)$. In some models the functions $\psi_{\theta,\gamma}$ can be chosen in such a way that some minimal stability of the estimator $\hat{\gamma}$, such as consistency, suffices. In this talk we focus on the different situations where a minimal rate of convergence for $\hat{\gamma}$ is crucial. In that case the quality of the estimator $\hat{\theta}$ may be improved by replacing the linear estimating equation by an equation of the type

$$\mathbb{U}_n \psi_{\theta,\hat{\gamma}} = 0,$$

where $\mathbb{U}_n$ denotes a U-statistic "operator" defined by

$$\mathbb{U}_n f = \frac{(n-k)!}{n!} \sum_{1 \le i_1 \ne \cdots \ne i_k \le n} \cdots \sum f(X_{i_1}, \ldots, X_{i_k}),$$

for each function $(x_1, \ldots, x_k) \mapsto f(x_1, \ldots, x_k)$.

Rather than solving an equation of this type it will be convenient to use a one-step estimator, defined by

$$\chi(\hat{\eta}) + \mathbb{U}_n \dot{\chi}_{\hat{\eta}},$$

for a given function $(x_1, \ldots, x_k) \mapsto \dot{\chi}_\eta(x_1, \ldots, x_k)$ called "influence function". This can be considered the zero of the linear approximation to the estimating equation at a given preliminary estimator.

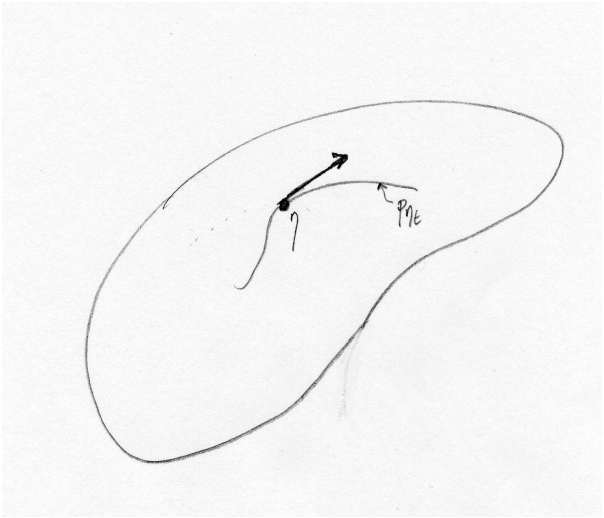
An illuminating example is the missing data problem. A typical observation is a triple $X = (YA, A, Z)$, which is a function of a treatment indicator A with value in $\{0, 1\}$ of an outcome Y and a covariate Z. The outcome is only observed if $A = 1$. Such a situation arises in the study of the causal effect of treatments. If the data are gathered in an observational study, then the outcome of the treatment Y and the treatment indicator A are typically dependent. By gathering information Z on all possible confounding variables it can be ensured that given Z treatment Y and treatment indicator A are independent. Because typically there is only little information about which variables could be confounding, the covariate vector Z must be chosen very high-dimensional. For simplicity we assume that Y also

takes its values in $\{0,1\}$. Then we can complete the description of the model by assuming that given Z the treatment outcome Y and treatment indicator A are independent; that given Z the variables Y and A have Bernoulli distributions with success probabilities $b(Z)$ and $p(Z)$, respectively; and that Z has a density f. The parameter is $\eta = (b, p, f)$ and the parameter of interest is $\chi(\eta) = \int bf = \mathrm{E}Y$. We are interested in the situation that Z is very high-dimensional and that the functions p, b, f are unknown.

We want influence functions $\dot{\chi}_\eta$ that work with general purpose $\hat{\eta}$. Good influence functions have "correct" inner products (covariances) with score functions of the model. For the linear case this idea has been developed in terms of a tangent space of a model. The *tangent space* (at η) is the linear span of all score functions $g = \frac{d}{dt}_{|t=0} \log p_{\eta_t}$ of suitable one-dimensional submodels $t \mapsto p_{\eta_t}$. See Figure 2. The *influence function* of $\eta \mapsto \chi(\eta)$ is a map $x \mapsto \dot{\chi}_\eta(x)$ such that for all submodels $t \mapsto p_{\eta_t}$

$$\frac{d}{dt}\chi(\eta_t)_{|t=0} = \mathrm{E}_\eta g(X_1)\dot{\chi}_\eta(X_1)$$

Such an influence function is unique up to the orthocomplement (in $L_2(\eta)$) of the tangent space, and it was shown in semiparametric theory that estimators for $\chi(\eta)$ of minimum asymptotic variance correspond to the unique influence function inside the closure of the tangent space.



In the missing data problem the likelihood is

$$p_\eta(X) = f(Z)p(Z)^A\big(1-p(Z)\big)^{1-A}b(Z)^{YA}\big(1-b(Z)\big)^{(1-Y)A}.$$

Submodels indexed by the perturbed parameters $p_t = p + t\pi$, $b_t = b + t\beta$ and $f_t = f(1 + t\phi)$ can easily be shown to lead to scores

$$\frac{A - p(Z)}{p(Z)(1-p)(Z)}\pi(Z) + \frac{A(Y - b(Z))}{b(Z)(1-b)(Z)}\beta(Z) + \phi(Z).$$

The influence function of the parameter of interest $\chi(\eta) = \int bf = \mathrm{E}Y$ is given by

$$\dot{\chi}_\eta(X) = \frac{A(Y - b(Z))}{p(Z)} + b(Z) - \chi(\eta).$$

In semiparametric models with a partitioned parameter $\eta = (\theta, \gamma)$ the influence function for $\chi(h) = \theta$ can be seen to be the scaled efficient score $\tilde{I}_{\theta,\gamma}^{-1}\tilde{\ell}_{\theta,\gamma}$, where $\tilde{\ell}_{\theta,\gamma}$ is the projection of the θ-score on the orthocomplement of the scores for γ and $\tilde{I}_{\theta,\gamma}$ is its variance.

The role of the influence function can be seen from the expansion of the one-step estimator $\hat{\theta} = \chi(\hat{\eta}) + \mathbb{U}_n\dot{\chi}_{\hat{\eta}}$

$$\begin{aligned}\hat{\theta} - \chi(\eta) &= (\mathbb{U}_n - \mathrm{E}_\eta)\dot{\chi}_{\hat{\eta}} + \big[\chi(\hat{\eta}) - \chi(\eta) - (\mathrm{E}_{\hat{\eta}} - \mathrm{E}_\eta)\dot{\chi}_{\hat{\eta}}(X_1)\big] \\ &= O_P\Big(\frac{1}{\sqrt{n}}\Big) + O_P\big(\|\hat{\eta} - \eta\|^2\big).\end{aligned}$$

The first equality is simple algebra. The second equality is typically true if the influence function $\eta \mapsto \dot{\chi}_\eta$ is continuous and the functional $\eta \mapsto \chi(\eta)$ is twice differentiable relative to an appropriate norm on the parameter space. The first term on the right times $\sqrt{n}$ is even typically asymptotically normally distributed. The use of the influence function ensures both that this term has minimal variance (if the influence function is chosen in the closure of the tangent space) and that the second term is second order rather than first order $O_P\big(\|\hat{\eta} - \eta\|\big)$. Because the first term gives the optimal behaviour we think of this term as a bias term.

In the missing data problem this second term is equal to

$$\int \Big(\frac{p}{\hat{p}} - 1\Big)(\hat{b} - b)\, dF = O_P\big(\|\hat{p} - p\|\,\|\hat{b} - b\|\big).$$

It follows that this bias term is negligible if the product of the convergence rates of the preliminary estimators of p and b is $o(n^{-1/2})$. This is only possible under a-priori assumptions on these parameters. For instance, if they are known to be α-smooth, then the optimal rate is $n^{-\alpha/(2\alpha+d)}$ for d the the dimension of the covariate, and it is needed that $\alpha > d/2$. Existence of at least five derivatives on a 10-dimensional covariate space would be necessary. If the true parameters have smaller regularity, then the bias term dominates. We can then improve the estimation and, in particular the induced confidence sets, by using higher order estimating equations.

In the higher-order case we look again for an influence function with "correct" inner products with score functions, this time the influence function being a function $(x_1, \ldots, x_k) \mapsto \dot{\chi}_\eta(x_1, \ldots, x_k)$ of k arguments, employing higher-order derivatives, and scores from a higher-order tangent space. The one-step estimator $\hat{\theta} = \chi(\hat{\eta}) + \mathbb{U}_n\dot{\chi}_{\hat{\eta}}$ ought now satisfy

$$\hat{\theta} - \theta = (\mathbb{U}_n - \mathrm{E}_\eta)\dot{\chi}_{\hat{\eta}} + O_P\big(\|\hat{\eta} - \eta\|^{k+1}\big),$$

as we have matched up more derivatives. Unfortunately, this is too good to be true. Existence of such influence functions would allow to reduce the bias term without increasing the variance term. Closer inspection shows that higher order

influence functions can be defined for finite-dimensional submodels, but not for the full model. Use of an influence for a submodel, say of dimension m, introduces an extra error, and the expansion for the one-step estimator takes the form

$$\hat\theta - \theta = (\mathbb{U}_n - \mathrm{E}_\eta)\dot\chi_{\hat\eta,m} + O_P\big(\|\hat\eta - \eta\|^{k+1}\big) + \text{approximation bias}_m.$$

Appropriate values of k and m (or a subspace) are obtained by balancing the three terms on the right.

In the missing data model the second order influence function takes the form

$$\begin{aligned}&\frac{A_1(Y_1 - b(Z_1))}{p(Z_1)} + b(Z_1) - \chi(\eta)\\ &+ \Big[-\frac{A_1(Y_1 - b(Z_1))}{p(Z_1)}(A_2 - p(Z_2)) - \frac{(A_1 - p(Z_1))}{p(Z_1)}\frac{A_2(Y_2 - b(Z_2))}{p(Z_2)}\Big]K_m^f(Z_1, Z_2),\end{aligned}$$

where K_m^f is the kernel of a projection $K_m^f : L_2(f) \to L$ onto an appropriate m-dimensional subspace. There are more complicated expressions for higher-order kernels. Use of this kernel leads to three terms in the expansion of $\hat\theta - \theta$ of the orders

$$O\Big(\frac{m^{k-1}}{n^k} \vee \frac{1}{n}\Big) + O\big(\|\hat b - b\|\|\hat p - p\|\|\hat f - f + \hat p - p\|^{k-1}\big) + O\big(\|b - K_m^f b\|\|p - K_m^f p\|\big).$$

Balancing these terms leads, for $d = 10$, to the table

α	m	k
≥ 5	n	1
$[2.5, 5)$	n	2
$[\cdots, 2.5)$	$n^{15\alpha/(2\alpha+10)}$	3

The first column gives a-priori smoothness of the parameters p, b and f, taken equal for simplicity. The third column gives the optimal value of k and the second the optimal dimension of the approximating subspace L.

Confidence intervals can be based on (conditional) asymptotic normality of $\hat\theta - \theta$. The U-statistics are asymptotically normal because the kernel shrinks to the diagonal. Monte Carlo experiments shows that this may work reasonably well.

In further work we hope to show that optimal values of k and m can be chosen data-dependent through cross-validation. This would lead to estimators that adapt to the unknown regularity of the parameters. Unfortunately, adaptation in the case of confidence intervals appears to be impossible. A true confidence interval must be based on a largest possible model. A statistician could report the results of the analysis as a sequence of conditional statements of the form: "if the truth can be assumed to be a-priori regular to this order, then the parameter of interest is in the interval $\cdots$", one statement for each reasonable a-priori assumption.

References

[1] A.W. van der Vaart, *Asymptotic statistics*, Cambridge University Press (1998).

Classification with reject option

Marten H. Wegkamp
(joint work with Radu Herbei)

Pattern recognition is about classifying an observation that takes values in some feature space $\mathcal{X}$ as coming from a fixed number of classes, say $0, 1, \ldots, M$. The simplest framework is that of binary classification ($M = 1$) with $\mathcal{X} = \mathbb{R}^k$. It is not assumed that an observation $X = x$ fully determines the label y; the same x may give rise to different labels. Based on a collection of labelled observations (x_i, y_i), the statistician's task is to form a classifier $f : \mathbb{R}^k \to \{0, 1\}$ which represents her guess of the label Y of a future observation X. This framework is known as supervised learning in the literature. The classifier

$$f(x) = \begin{cases} 0 & \text{if } \mathbb{P}\{Y = 0|X = x\} \geq \mathbb{P}\{Y = 1|X = x\} \\ 1 & \text{otherwise} \end{cases} \tag{1}$$

has the smallest probability of error, see, for example ((5), Theorem 2.1, page 10). We will allow for the classifiers to report "I don't know" expressing doubt, if the observation x is too hard to classify. This happens when the conditional probability $\eta(x) = \mathbb{P}\{Y = 1|X = x\}$ is close to $1/2$. Indeed, if $\mathbb{P}\{Y = 0|X = x\} = \mathbb{P}\{Y = 1|X = x\} = 1/2$, then we might just as well toss a coin to make a decision. The main purpose of supervised pattern recognition or machine learning is to classify the majority of future observations in an automatic way. However, allowing for the reject option ("I don't know") besides taking a hard decision (0 or 1) is of great importance in practice, for instance, in case of medical diagnoses. Nevertheless, this option is often ignored in the statistical literature. (13) and recently (6) are notable exceptions. Some references in the engineering literature are (4), (9), (7), (8), (11).

We follow the decision theoretic framework of (4), see (13) (Chapter 2) for a more general overview. Let $f : \mathbb{R}^k :\to \{0, 1, R\}$ be a classifier with a reject option, where the interpretation of the output R is of being in doubt and taking no decision. The misclassification probability is $\mathbb{P}\{f(X) \neq Y, f(X) \neq R\}$ and reject or doubt probability is $\mathbb{P}\{f(X) = R\}$. Assuming that the cost of making a wrong decision is 1 and that of utilizing the reject option is $d > 0$, the appropriate risk function to employ is

$$d\mathbb{P}\{f(X) = R\} + \mathbb{P}\{f(X) \neq Y, f(X) \neq R\}. \tag{2}$$

(4) shows that the optimal rule minimizing the risk (2) is

$$f^*(x) = \begin{cases} 0 & \text{if } 1 - \eta(x) > \eta(x) \text{ and } 1 - \eta(x) > 1 - d \\ 1 & \text{if } \eta(x) > 1 - \eta(x) \text{ and } \eta(x) > 1 - d \\ R & \text{if } \max(\eta(x), 1 - \eta(x)) \leq 1 - d \end{cases} \tag{3}$$

which we will refer to as the Bayes rule with reject option. According to this rule, we should never invoke the reject option if $d \geq 1/2$ and we should always reject if

$d = 0$. For this reason we can restrict ourselves to the cases $0 \leq d \leq 1/2$ and we denote the relevant risk function (2) by

$$L_d(f) = d\mathbb{P}\{f(X) = R\} + \mathbb{P}\{f(X) \neq Y, f(X) \neq R\}. \tag{4}$$

The Bayes rule (3) simplifies to

$$f^*(x) = \begin{cases} 0 & \text{if } \eta(x) < d \\ 1 & \text{if } \eta(x) > 1 - d \\ R & \text{otherwise,} \end{cases} \tag{5}$$

and we denote its risk by

$$L_d^* = L_d(f^*) = \min_{f:\mathbb{R}^k \to \{0,1,R\}} L_d(f). \tag{6}$$

The case $d = \frac{1}{2}$ reduces to the classical situation without the reject option and the Bayes classifier (5) reduces to (1). We will demonstrate that the behavior of $\eta(x)$ near the value $1/2$ and more generally in the interval $(d, 1-d)$ is of no real importance; the classification problem hinges on what happens outside this interval, especially at the values d and $1 - d$.

The talk is based on the technical report (10) and is organized as follows. We first discuss plug-in rules based on the Bayes rule (5). These rules are called this way since they replace the regression function $\eta(x)$ by an estimate $\widehat{\eta}(x)$ in formula (5). Besides introducing the reject option, we extend the existing theory for plug-in rules ((5), Theorem 2.2) since our bound depends explicitly on both the difference $|\widehat{\eta}(X) - \eta(X)|$ and the behavior of $\eta(X)$ near the values d and $1 - d$. We show that very fast rates are possible under reasonable margin conditions, extending a recent result by (1) to our more general framework. We illustrate the theory with an application to speech recognition.

Next we extend the existing theory of empirical risk minimizers by allowing for the reject option. Here an estimate is found by minimizing the empirical counterpart of the risk (4) over an entire class of classifiers $\mathcal{F}$. We demonstrate that the rates of the risk (4) of the resulting minimizers to the Bayes risk L_d^* depends on the metric entropy of (a transformed class of) $\mathcal{F}$ and on the behavior of $\eta(X)$ near the values d and $1 - d$. Again our results are in line with the recent developments of the theory for $d = 1/2$ (see, for example, (3), (12), (15), (16), (14)) and extend the theory to the general case $0 \leq d \leq 1/2$.

We push the theory even further since we differentiate between misclassification costs of the cases $Y = 1$ & $f(X) = 0$ and $Y = 0$ & $f(X) = 1$, a situation common in, for instance, medical studies where misclassifying a sick patient as healthy is worse than the opposite. The risk function (4) is changed to accommodate for this differentiation.

References

[1] J.Y. Audibert, J.-Y, Tsybakov, A. (2005). *Fast convergence rates for plug-in classifiers under margin conditions.* (personal communcation).

[2] P.L. Bartlett, O. Bousquet and S. Mendelson, *Local Rademacher complexities*, Annals of Statistics **33 (4)** (2005), 1497–1537.
[3] S. Boucheron, O. Bousquet and G. Lugosi, *Theory of Classification: a Survey of Recent Advances*, Manuscript (2004).
[4] C.K. Chow, *On optimum error and reject trade-off*, IEEE Transactions on Information Theory **16** (1970), 41–46.
[5] L. Devroye, L. Györfi and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*, Springer-Verlag, New York (1996).
[6] Y. Freund, Y. Mansour and R.E. Schapire, *Generalization bounds for averaged classifiers*, Annals of Statistics **32 (4)** (2004), 1698–1722.
[7] G. Fumera and F. Roli, *Analysis of error-reject trade-off in linearly combined multiple classifiers*, Pattern Recognition **37** (2004), 1245–1265.
[8] G. Fumera, F. Roli and G. Giacinto, *Reject option with multiple thresholds*, Pattern Recognition **33** (2000), 2099–2101.
[9] L. Györfi, Z. Györfi and I. Vajda, *Bayesian decision with rejection*, Problems of Control and Information Theory **8** (1978), 445–452.
[10] R. Herbei and M.H. Wegkamp, *Classification with reject option*, Department of Statistics, Florida State University **Preprint** (2005).
[11] L.K. Hansen, C. Lissberg and P. Salamon, *The error-reject tradeoff*, Open Systems & Information Dynamics **4** (1997), 159 – 184.
[12] P. Massart and E. Nédélec, *Risk bounds for statistical learning*, Université Paris Sud **Preprint**(2003).
[13] B.D. Ripley, *Pattern recognition and neural networks*, Cambridge University Press, Cambridge (1996).
[14] B. Tarigan and S.A. van de Geer, *Adaptivity of support vector machines with $l_1$penalty*, University of Leiden, Technical Report MI **2004-14** (2004).
[15] A.B. Tsybakov, *Optimal aggregation of classifiers in statistical learning*, Annals of Statistics **32 (1)** (2003), 135–166.
[16] A. B. Tsybakov and S.A. van de Geer, *Square root penalty: adaptation to the margin in classification and in edge estimation*, Prépublication PMA-820, Laboratoire de Probabilités et Modèles Aléatoires, Université Paris VII (2003).

Reporter: Gabriele Wieczorek

Participants

Prof. Dr. Peter L. Bartlett

Department of Statistics
University of California
367 Evans Hall
Berkeley, CA 94720-3860
USA

Prof. Dr. Maria-Jesus Bayarri

Departamento de Estadistica e I.O.
Universitat de Valencia
Av. Dr. Moliner 50
Burjasot
E-46100 Valencia

Prof. Dr. James O. Berger

Institute of Statistics and
Decision Sciences
Duke University
112B Old Chemistry Building
Durham NC 27708-0251
USA

Melanie Birke

Fakultät für Mathematik
Ruhr-Universität Bochum
44780 Bochum

Dr. Nicolai Bissantz

Institut f. Mathemat. Stochastik
Georg-August-Universität Göttingen
Maschmühlenweg 8-10
37073 Göttingen

Prof. Dr. Malgorzata Bogdan

Institute of Mathematics and
Computer Science
Wroclaw University of Technology
Wybrzeze Wyspianskiego 27
Wroclaw 50-370
POLAND

Prof. Dr. Stephane Boucheron

Institut Mathematiques de Jussieu
Universite Pierre et Marie Curie
175, Rue du Chevaleret
F-75013 Paris

Leif Boysen

Institut f. Mathemat. Stochastik
Georg-August-Universität Göttingen
Maschmühlenweg 8-10
37073 Göttingen

Prof. Dr. Florentina Bunea

Department of Statistics
Florida State University
Tallahassee Fl. 32306-4330
USA

Prof. Dr. Merlise Clyde

Institute of Statistics and
Decision Sciences
Duke University
Durham NC 27708-0251
USA

Prof. Dr. Rainer Dahlhaus

Institut für Angewandte Mathematik
Universität Heidelberg
Im Neuenheimer Feld 294
69120 Heidelberg

Prof. Dr. P. Laurie Davies

FB 6 - Mathematik
Universität Duisburg-Essen
45117 Essen

Prof. Dr. Holger Dette

Fakultät für Mathematik
Ruhr-Universität Bochum
44780 Bochum

Prof. Dr. Lutz Dümbgen

Mathematische Statistik
und Versicherungslehre
Universität Bern
Sidlerstraße 5
CH-3012 Bern

Prof. Dr. Ludwig Fahrmeir

Institut für Statistik
Universität München
Ludwigstr. 33
80539 München

Prof. Dr. Ursula Gather

Fachbereich Statistik
Universität Dortmund
44221 Dortmund

Prof. Dr. Sara van de Geer

Seminar für Statistik
ETH-Zentrum Zürich
LEO D12
Leonhardstr. 27
CH-8092 Zürich

Prof. Dr. Edward I. George

Department of Statistics
The Wharton School
University of Pennsylvania
3730 Walnut Street
Philadelphia, PA 19104-6340
USA

Prof. Dr. Jayanta K. Ghosh

Department of Statistics
Purdue University
Mathematical Sciences Building
West Lafayette, IN 47907
USA

Prof. Dr. Laszlo Györfi
Department of Mathematics
Technical University of Budapest
Stoczek u. 2
H-1111 Budapest XI

Prof. Dr. Wolfgang Härdle

Wirtschaftswissenschaftl. Fakultät
Lehrstuhl für Statistik
Humboldt-Universität Berlin
Spandauer Str. 1
10178 Berlin

Prof. Dr. Nils Lid Hjort

Department of Mathematics
University of Oslo
P. O. Box 1053 - Blindern
N-0316 Oslo

Dr. Hajo Holzmann

Institut f. Mathemat. Stochastik
Georg-August-Universität Göttingen
Maschmühlenweg 8-10
37073 Göttingen

Prof. Dr. Katja Ickstadt

Fachbereich Statistik
Universität Dortmund
44221 Dortmund

Prof. Dr. Arnold Janssen

Mathematisches Institut
Heinrich-Heine-Universität
Universitätsstr. 1
40225 Düsseldorf

Prof. Dr. Valen E. Johnson

Department of Biostatistics
Box 0447
U.T.M.D. Anderson Cancer Center
1515 Holcombe Boulevard
Houston TX 77030
USA

Prof. Dr. Vladimir Koltchinskii

Dept. of Mathematics and Statistics
University of New Mexico
Albuquerque, NM 87131-1141
USA

Prof. Dr. Jens-Peter Kreiß

Institut für Mathematische
Stochastik der TU Braunschweig
Pockelsstr. 14
38106 Braunschweig

Prof. Dr. Angelika van der Linde

Fachbereich 3
Mathematik und Informatik
Universität Bremen
Postfach 330440
28334 Bremen

Prof. Dr. Jean-Michel Loubes

Department of Mathematics
Univ. Paris-Sud
Bat. 425
F-91405 Orsay Cedex

Prof. Dr. Gabor Lugosi

Department of Economics
Pompeu Fabra University
Ramon Trias Fargas 25-27
E-08005 Barcelona

Rada Matic

Institut f. Mathemat. Stochastik
Georg-August-Universität Göttingen
Maschmühlenweg 8-10
37073 Göttingen

Prof. Dr. Robert E. McCulloch

Graduate School of Business
The University of Chicago
1101 E. 58th Street
Chicago IL 60637
USA

Prof. Dr. Peter Müller

Department of Biostatistics
Box 0447
U.T.M.D. Anderson Cancer Center
1515 Holcombe Boulevard
Houston TX 77030
USA

Prof. Dr. Klaus-Robert Müller

Fraunhofer Institut FIRST
Intelligent Data Analysis Group
(IDA)
Kekulestr. 7
12489 Berlin

Prof. Dr. Axel Munk

Institut f. Mathemat. Stochastik
Georg-August-Universität Göttingen
Maschmühlenweg 8-10
37073 Göttingen

Prof. Dr. Michael Helmut Neumann

Institut für Mathematische
Stochastik der TU Braunschweig
Pockelsstr. 14
38106 Braunschweig

Dr. Natalie Neumeyer

Fakultät für Mathematik
Ruhr-Universität Bochum
44780 Bochum

Prof. Dr. Andrew B. Nobel

Department of Statistics and
Operations Research
University of North Carolina
Chapel Hill, NC 27599-3260
USA

Vladimir Ostrovski

Mathematisches Institut
Heinrich-Heine-Universität
Universitätsstr. 1
40225 Düsseldorf

Dr. Rui Paulo

School of Mathematics
University of Bristol
University Walk
GB-Bristol BS8 1TW

Prof. Dr. Luis Raul Pericchi

Department of Mathematics
University of Puerto Rico
Rio Piedras Campus
P.O.Box 23355
San Juan 00931-3355
PUERTO RICO (U.S.A.)

Prof. Dr. Christian P. Robert

CEREMADE
Universite Paris Dauphine et CREST
Place du Marechal de Lattre de
Tassigny
F-75775 Paris Cedex 16

Regine Scheder

Fakultät für Mathematik
Ruhr-Universität Bochum
44780 Bochum

Prof. Dr. Vladimir Spokoiny

WIAS
Mohrenstr. 39
10117 Berlin

Prof. Dr. Ulrich Stadtmüller

Abteilung Zahlentheorie und
Wahrscheinlichkeitstheorie
Universität Ulm
Helmholtzstrasse 18
89069 Ulm

Dr. Ansgar Steland

Fakultät für Mathematik
Ruhr-Universität Bochum
Universitätsstr. 150
44801 Bochum

Prof. Dr. Alexander B. Tsybakov

Laboratoire de Probabilites
Universite Paris 6
4 place Jussieu
F-75252 Paris Cedex 05

Prof. Dr. Aad W. van der Vaart

Faculteit Wiskunde en Informatica
Vrije Universiteit Amsterdam
De Boelelaan 1081 a
NL-1081 HV Amsterdam

Janis Valeinis

Institut f. Mathemat. Stochastik
Georg-August-Universität Göttingen
Maschmühlenweg 8-10
37073 Göttingen

Prof. Dr. Ingrid Van Keilegom

Institut de Statistique
Universite Catholique de Louvain
Voie du Roman Pays 20
B-1348 Louvain-la-Neuve

Prof. Dr. Marten Wegkamp

Dept. of Statistics
Florida State University
Tallahassee, FL 32306-3033
USA

Gabriele Wieczorek

Fakultät für Mathematik
Ruhr-Universität Bochum
44780 Bochum