Mathematisches Forschungsinstitut Oberwolfach

Report No. 50/2007

Reassessing the Paradigms of Statistical Model-Building

Organised by
Ursula Gather, Dortmund
Peter Hall, Canberra
Hans-Rudolf Künsch, Zürich

October 21st – October 27th, 2007

Abstract. Statistical model-building is the science of constructing models from data and from information about the data-generation process, with the aim of analysing those data and drawing inference from that analysis. Many statistical tasks are undertaken during this analysis; they include classification, forecasting, prediction and testing. Model-building has assumed substantial importance, as new technologies enable data on highly complex phenomena to be gathered in very large quantities. This creates a demand for more complex models, and requires the model-building process itself to be adaptive.

The word "paradigm" refers to philosophies, frameworks and methodologies for developing and interpreting statistical models, in the context of data, and applying them for inference. In order to solve contemporary statistical problems it is often necessary to combine techniques from previously separate paradigms.

The workshop addressed model-building paradigms that are at the frontiers of modern statistical research. It tried to create synergies, by delineating the connections and collisions among different paradigms. It also endeavoured to shape the future evolution of paradigms.

Mathematics Subject Classification (2000): 62-06,62A01,62C05,62G99.

Introduction by the Organisers

The development of statistics during the last century has involved largely disjoint paradigms. Sometimes these have been complementary, for example in the case of Bayesian and frequentist methodologies. In other instances they have been overlapping, e.g. model-selection methods such as minimum description length methods and Akaike's information criterion; or evolutionary, e.g. parametric, semiparametric and nonparametric approaches; or developed along similar lines, e.g. parametric and nonparametric approaches to likelihood; or related in other ways,

e.g. dimension-reduction methods and techniques for analysing high-dimensional data. The mathematical theory behind these techniques is especially complex and difficult. One example is the understanding and harnessing of the geometry of likelihood, which is still a major task for theoretical statisticians.

Finding a statistical model may include graphical representation of the data, calculation of relevant statistics, checking of putative models against the data, and assessment of possible outliers or serial correlations. In other cases, such as in nonparametric regression, the family of models is so large that a major aspect of the problem is choosing a particular model from a very large class, for example in the context of sparsity.

It is often only at the final stage of model specification that such formalised strategies are employed. Examples include Akaike's information criterion (AIC), Bayes information criterion (BIC), minimum description length and stochastic complexity (MDL) and cross-validation.

When statistical model selection is framed in a mathematical setting, it often arises as an optimisation problem, and has many points of contact with applied mathematics. The constraints, and hence the objective function, are determined by the paradigm. In this context the type of topology (strong or weak), the methods for computation and other mathematical issues play major roles. Statisticians are forced to apply and also to develop mathematical theory in order to find the techniques and concepts they need to understand the complex, real-world problems that motivate advances in their subject.

In the past the paradigms of statistical model building were developed separately. In the future, multiple paradigms will have to be used simultaneously. Indeed, many frontier problems in statistics today already involve several different concepts simultaneously. For example, techniques for analysing complex, high-dimensional data sets often use methods for complexity measurement, dimension reduction and classification. The demand for multiple paradigms in statistical model building was a major motivator of the workshop.

Thus, the workshop drew together statisticians working on the development and application of statistical model building, with the aim of critiquing different approaches, assessing their usefulness, developing new techniques, and mapping future directions for research.

The workshop was well attended, with 45 participants from all over the world, among them many young researchers. We were able to bring together experts from different fields of statistics: Bayesian methods, machine learning, likelihood theory, minimum description length and others.

Each morning, especially towards the beginning of the workshop, somewhat longer lectures were presented by senior researchers. The opening lecture was given by L. D. Brown, followed by other review-type presentations during the first three days for example by A.P. Dawid, S. Fienberg, R. Beran, R. Shibata, J. Rissanen, P. L. Davies and A.B. Tsybakov.

These lectures gave rise to a lively floor discussion on Wednesday evening, on the very meaning of statistical paradigms and on the new tasks for statistics in

finding methods extracting important information from data. Particular attention was focused on challenging statistical problems emerging from new research areas, arising for example in the life sciences, physics, the social sciences, etc.

To characterise better these new challenges, which are often associated with new data structures, talks of more applied type were presented during the last two days, for example by G. Winkler, R. Carroll, A. Welsh, J. Ramsay and P. Hall. These yielded further discussion, and also new research cooperations among the participants.

Workshop: Reassessing the Paradigms of Statistical Model-Building

Table of Contents

Abstracts

V-fold penalization: an alternative to V-fold cross-validation

Sylvain Arlot

One of the most widely used model selection techniques is *V-fold cross-validation* (Geisser [7]). It estimates the prediction error of estimators built upon $n(V-1)V^{-1} < n$ data, which can be interpretated as overpenalization. From the asymptotical viewpoint, this can be suboptimal (when V is fixed) and it has to be corrected, for instance following Burman [4]. However, when the sample size is small, it may happen that $V = 2$ gives better results than $V = 10$, because overpenalization is benefic in some cases [1]. The choice of V in V-fold cross-validation can then be a difficult problem.

Following Efron's resampling heuristics [5], we propose to use a V-fold resampling scheme to define a new penalization procedure, called *V-fold penalization* ([2], Chap. 5). It generalizes Burman's bias correction, and produces a flexible procedure, where V is decoupled from the overpenalization factor.

In the framework of regression on a random design with heteroscedastic noise, we prove several non-asymptotic results about V-fold subsampling, and more general resampling schemes. In particular, V-fold penalization (with V fixed) satisfies a non-asymptotic oracle inequality with constant almost one, which implies its asymptotic optimality. Hence, it improves on V-fold cross-validation. Moreover, choosing a particular family of models, we obtain an estimator adaptive to the smoothness of the regression function and the heteroscedasticity of the noise. Thus, V-fold penalties are more robust that Mallows' C_p criterion.

The theoretical results concerning V-fold penalties stay valid for resampling penalties with general exchangeable weights ([2], Chap. 6). In particular, they can be applied to V-fold penalties with $V = n$, as well as bootstrap penalties (defined by Efron [6]). This extends an asymptotical result on bootstrap penalties in another framework (Shibata [11]). Using independent Rademacher weights, one obtain a localized version of Rademacher complexities (Koltchinskii [8] ; Bartlett, Boucheron and Lugosi [3]) that is much easier to compute than local Rademacher complexities (Lugosi and Wegkamp [10] ; Koltchinskii [9]).

Although we have to assume a particular structure for the models (*i.e.* they are all made of histograms), we believe that the same results hold in a much more general framework. We for instance have partial results for general bounded regression and binary classification ([2], Chap. 7).

A simulation study shows that V-fold penalties behave quite well in several cases. Moreover, they often outperform V-fold cross-validation and Mallows' C_p penalties, in particular in difficult heteroscedastic situations. Their flexibility allows to improve performances when the signal-to-noise ratio is small; this is obtained by taking V large enough, together with overpenalization.

The choice of V also appear to be quite easier: the performances of V-fold penalties are always better when V increases. Then, V has only to be chosen according to the computational complexity of the procedure, which is exactly the same as the one of V-fold cross-validation.

References

[1] Sylvain Arlot, *Model selection by resampling penalization*, Short version. Arxiv:math.ST/0701542, 2007.

[2] Sylvain Arlot, *Resampling and Model Selection*, PhD thesis, University Paris XI, December 2007. Available at `http://www.math.u-psud.fr/~arlot/publi.html.en`

[3] Peter L. Bartlett, Stéphane Boucheron, and Gábor Lugosi, *Model selection and error estimation*, Machine Learning **48** (2002), 85–113.

[4] Prabir Burman, *A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods*, Biometrika **76**(3) (1989), 503–514.

[5] Bradley Efron, *Bootstrap methods: another look at the jackknife*, Ann. Statist. **7**(1) (1979), 1–26.

[6] Bradley Efron, *Estimating the error rate of a prediction rule: improvement on cross-validation*, J. Amer. Statist. Assoc. **78**(382) (1983), 316–331.

[7] Seymour Geisser, *The predictive sample reuse method with applications*, Journal of the American Statistical Association **70** (1975), 320–328.

[8] Vladimir Koltchinskii, *Rademacher penalties and structural risk minimization*, IEEE Trans. Inform. Theory **47**(5) (2001), 1902–1914.

[9] Vladimir Koltchinskii, *2004 IMS Medallion Lecture: Local Rademacher Complexities and Oracle Inequalities in Risk Minimization*, Ann. Statist. **34**(6) (2006) 2593–2706.

[10] Gábor Lugosi and Marten Wegkamp, *Complexity regularization via localized random penalties*, Ann. Statist. **32**(4) (2004), 1679–1697.

[11] Ritei Shibata, *Bootstrap estimate of Kullback-Leibler information for model selection*, Statist. Sinica **7**(2) (1997), 375–394.

Statistical model versus fit versus data

Rudolf Beran

Until the mid-twentieth century, the main tools for most statisticians were probability models for data and mechanical calculators. In this pre-computer environment, grand and simple philosophical theories of statistics were inevitable. The highly influential monographs by Wald , Savage, and Fisher in the 1950's are cases in point. Such a phase has not been unusual in the development of a science. Within the intellectual frameworks of the pioneers, mathematical statistics has seen impressive developments as a methodology for analyzing noisy data. But have these ideas reached their limits?

Computer technologies have recalled to prominence the significant distinctions among data, probability models, pseudo-random numbers, procedures, and their numerical realizations. For instance, Tukey's book on exploratory data analysis in the mid-1970's made no use of probability models, focusing on statistical procedures and on non-probabilistic reasoning to support them. Yet, it also contains the sentence: "Today's understanding of how well elementary techniques work owes much to both mathematical analysis and experimentation by computer."

A major technology typically has evident effects and subconscious effects on human behavior. Once a major new technology establishes itself, the hidden effects of its predecessors become increasingly evident. Sharp perceptions of this phenomenon include, "The medium is the message" (H. M. McLuhan) and "To a man with a hammer, everything looks like a nail" (Mark Twain).

Statistics has succeeded remarkably in exporting ideas and formulations to the statistical sciences, fields with names often ending in "metrics" that express their basic concepts in statistical terms. However, data is rarely if ever certifiably random. Standard model assumptions in math stat theory are strong idealizations. Our usual math stat tools do not suffice to evaluate emerging complex data-analytic algorithms. *Statistics: Challenges for the Twenty-First Century*, a report written in 2004 by a distinguished panel for the U.S. National Science Foundation, has noted the intellectual gaps between current math stat theory and present challenges in data analysis.

Through use of computer technologies, statistics has the potential to become an experimentally supported information science, successor to its current formulation as a mathematical philosophy. The success of statistical procedures is, in fact, an empirical question. Statistics can speed its potential transition to an information science by being clear-headed about these matters. Subjects, such as physics, that made the transition from philosophy to science centuries ago (once pertinent technologies allowed) offer models for interplay between theory and experiment.

A technical case-study explores these topics further. The distinct means of a multi-way layout with one or more q-variate responses observed at each of p factor-level combinations can be arranged systematically into a $p \times q$ matrix M, each row specifying a mean response. The study develops practical regularized estimators of M that typically dominate, in asymptotic loss, the least squares fit to the model. The construction first devises a class of candidate estimators as the closure of a class of Bayes estimators for M; and then finds the candidate estimator with smallest estimated risk or loss. The candidate estimators rely on affine shrinkage of s-fold projection decompositions of the least squares estimator of M. As the number p of factor-level combinations in the multi-way layout tends to infinity, the loss of the regularized estimator is seen to converge asymptotically to that of the best candidate estimator. Adaptation over a class of s-fold projection decompositions as well as over affine shrinkage matrices is the main technical advance in this paper. Most importantly, the treatment is under a fixed effects statistical model that makes minimal assumptions.

A unified view of regression, shrinkage, empirical bayes, hierarchical bayes, and random effects

Lawrence D. Brown

A wide range of statistical problems involve estimation of means or conditional means of multidimensional normal distributions. There are many commonly employed classes of statistical models and related approaches to such problems. This talk surveys the interrelations among some of these approaches, and proposes some issues for further investigation.

The survey begins with a review of the background of shrinkage estimation. Stein (1956) surprised the statistical world with his discovery that the ordinary least squares estimator of a multivariate normal mean is not admissible in the usual setting. James and Stein (1961) then produced their classic estimator which often provides significant improvement over the ordinary estimator. 'Shrinkage' is a core feature of the estimator. An empirical Bayes interpretation of shrinkage was first proposed by Stein (1962) and Lindley (1962). The interpretation was effectively exploited by Efron and Morris (1972) and subsequently by many others. The empirical Bayes interpretation and its hierarchical fully Bayes first cousin, as first developed for this problem by Strawderman (1972), provide an important link to the manifestations of shrinkage in the various contemporary methodologies. The Bayesian viewpoint is also completely consistent with a random-effects view of the situation. These perspectives in turn allow for a shrinkage motivation of familiar ordinary linear regression.

Some analytic theory and data analyses illustrate the main points. The first of the data-based illustrations uses Galton's original data on adult heights. (See Hanley (2004) for the data.) The goal is to use heights of daughters within a family to predict the heights of the sons within that family. The second illustration sketches an analysis of US baseball batting averages, with the goal being to use each batters first half-season batting records in order to predict their second half-season performance. (See Brown (2007) for a thorough analysis of this data.) After preliminary manipulations both these examples involve estimation of means, and out-of-sample predictions, based on heteroscedastic Gaussian data. The data is moderately high dimensional (151 families and 567 batters, respectively).

It is (now) well-known that the observed sample means are themselves not desirable estimators in such contexts. For homoscedastic data shrinkage estimation ala James and Stein provides canonical frequentially motivated estimators that dominate the sample means. Shrinkage is intimately related to three other approaches to estimation (and other inference) for such data, which we termed the "three siblings". These are Empirical Bayes, Hierarchical Bayes, and Random Effects. The close connection among these three and their close relation to minimax shrinkage provides increased motivation for them. However, this does not provide much basis for choosing any one version from one among them as the version of choice. Indeed, in the canonical homoscedastic setting there is little practical difference in performance among them. There are, however, significant practical differences in heteroscedastic settings.

In the homoscedastic setting ordinary regression can also be viewed as a shrinkage estimator. The view here is the converse of that in Stigler (1990) in which shrinkage is interpreted as a version of ordinary regression. The interpretation of regression as a version of shrinkage augments the understanding of (any one of) the three siblings in heteroscedastic settings, and also further motivates their use. This shrinkage idea is encapsulated in rough form in the regression paradox that dates back to Galton's original treatments of his data. In heteroscedastic settings (as in the examples treated in our presentation) the general shrinkage idea behind regression seems appropriate, but its insistence on fitting a linear estimation/prediction form is not desirable.

For heteroscedastic problems, such as those considered here, there are significant numerical differences among different implementations of the different procedures. The most pronounced difference is that between the classical proposals for minimax shrinkage (as, eg in Berger (1985, Theorem 5.20)) and the various formulas for the three siblings. This difference has been noted by many researchers. See, eg Casella (1980). Roughly, the classical minimax proposals shrink the most on the dimensions where the variance is smallest. This type of behavior contrasts with all the other proposals here which shrink the least on those dimensions, and the classical minimax procedure is neither intuitively appealing nor numerically efficient in the examples.

To remedy this, a different type of risk function is proposed as a criterion for minimaxity (and admissibility) in problems such as ours that involve estimation of several means of qualitatively exchangeable importance, a-priori. Suppose it is desired to estimate the coordinate values $\{\theta_i : i = 1, .., p\}$ of the vector θ. The ordinary squared-error risk function for a procedure δ, is $R(\theta, \delta) = E_\theta\left(\|\delta(X) - \theta\|^2\right)$. We propose instead to judge a procedure by its ensemble risk. There are alternate versions of ensemble risk that can be motivated from different perspectives, and may lead to somewhat different results.

One version of this risk is

$$\overline{\overline{R}}\left(\gamma^2, \delta\right) = \int R(\theta, \delta)\, \phi_p\left(\theta; 0, \gamma^2\right) d\theta$$

where $\phi_p\left(\theta; 0, \gamma^2\right)$ denotes the p-dimensional normal density with iid coordinates having mean 0 and variance γ^2. (In this version the ensemble risk is a function of only one hyper-parameter, γ^2.)

Another version of ensemble risk can be defined as follows. Let $\theta_{(\bullet)}$ denote the p-dimensional vector whose coordinates are the increasingly ordered coordinate values of θ. Then define this version of ensemble risk as a function of the values of $\theta_{(\bullet)}$ by

$$\overline{R}\left(\theta_{(\bullet)}, \delta\right) = \frac{1}{p!} \sum_{\psi : \psi_{(\bullet)} = \theta_{(\bullet)}} R(\psi, \delta).$$

We conjecture that many of the standard shrinkage type estimators are minimax and nearly admissible for both $\overline{R}$ and $\overline{\overline{R}}$. (An appropriately chosen hierarchical Bayes estimator should be minimax and admissible.)

In the baseball batting example it is possible to provide an interesting comparison of the out-of-sample performance of several versions of empirical Bayes, hierarchical Bayes and ordinary shrinkage estimators. It turns out that a nonparametric empirical Bayes estimator suggested in Brown and Greenshtein (2007) performs best, with the ordinary shrinkage estimator and a method-of-moments parametric empirical Bayes estimator not far behind. Other versions of empirical Bayes and hierarchical Bayes perform less well, although - as anticipated - all of the methods dominate the ordinary, naive estimator. (Other numerical investigations we have performed suggest that the explanation for the weaker performance of some of the methods may be a robustness issue related to structural features of the baseball context that are not reflected in the motivation for these methods.)

Finally, it is noted that the general perspectives here extend considerably beyond the specific data structures of the examples. These perspectives apply to a much wider variety of settings in which shrinkage is also appropriate. These setting include multiple regression, longitudinal and panel data models, spatial models (especially those appropriate for "Kriging"), penalized likelihood methods ("regularization") involving quadratic penalty functions (especially smoothing splines), and various nonparametric regression and density estimation problems. Other setting involving varieties of shrinkage should be considered as being also related.

References

[1] J.O. Berger, *Statistical Decision Theory and Bayesian Analysis*, Springer-Verlag, NY 1985.

[2] L.D. Brown, *In-season prediction of batting averages: a field test of empirical Bayes and Bayes methodologies*, Ann. Appl. Statist., (2007), to appear.

[3] L.D. Brown and E. Greenshtein, *Nonparametric empirical Bayes and compound decision approaches to estimation of normal means*, manuscript (2007).

[4] G. Casella, *Minimax ridge regression estimation*, Ann. Statist. **8** (1980), 1036–1056

[5] B. Efron and C. Morris, *Stein's estimation rule and its competitors-an empirical Bayes approach*, Jour. Amer. Stat. Assoc., **68** (1973), 117–130.

[6] J.A. Hanley, *Transmuting women into men: Galton's family data on human stature*, Amer. Statist., **58** (2004), 237–243.

[7] W. James and C. Stein, *Estimation with quadratic loss*, Proc. Fourth Berk. Symp. Math. Statist. Prob., University of Calif. Press, **1** (1961), 311–319.

[8] D. Lindley, *Discussion on Professor Stein's paper*, Jour. Royal Statist. Soc, B, **24** (1962), 285–287.

[9] C. Stein, *Inadmissibility of the usual estimator for the mean of a multivariate normal distribution*, Proc. Third Berk. Symp. Math. Statist. Prob., University of Calif. Press, **1** (1956), 197–206

[10] C. Stein, *Confidence sets for the mean of a multivariate normal distribution (with discussion)*, Jour. Royal Statist. Soc, B, **24** (1962), 265–296

[11] S.M. Stigler, *The 1988 Neyman memorial lecture: a Galtonian perspective on shrinkage estimators*, Statist. Science, **5** (1990), 147–155

[12] W.E. Strawderman, *Proper Bayes estimators of the multivariate normal mean*, Ann. Math. Statist, **42** (1971), 385–388.

High-dimensional variable selection: faithfulness, strong associations and the PC-algorithm

Peter Bühlmann

We consider the problem of variable selection in high-dimensional linear models where the number of covariates greatly exceeds the sample size. In particular, we present the concept of partially faithful distributions and discuss their role for inferring associations between the response and the covariates. For partially faithful distributions, a simplified version of the PC-algorithm [9] which is computationally feasible even with thousands of covariates yields consistency for high-dimensional variable selection under clearly weaker conditions than penalty-based approaches; in fact, we prove that the PC-algorithm is consistent for very ill-posed design [1]. If partial faithfulness does not hold, we show that the PC-algorithm still consistently identifies some strong associations which are related to notions of causality [1].

The variable selection problem for high-dimensional models has recently gained a lot of attraction. A particular stream of research has focused on estimators and algorithms whose computation is feasible and provably correct [7, 13, 2, 3, 8, 11, 12]. As such, these methods distinguish themselves very clearly from heuristic optimization of an objective function or stochastic simulation or search, e.g. MCMC, which are often not really exploiting a high-dimensional search space. Prominent examples of computationally feasible and provably correct (w.r.t. computation) methods are penalty-based approaches, including the Lasso [10], the adaptive Lasso [13] or the Dantzig selector [3].

We propose here a method for linear models which is "diametrically opposed" to penalty-based schemes. Three reasons for another approach include the following: (i) it can be worthwhile to infer stronger concepts of associations than what is obtained from the usual regression coefficients, in particular when focusing on causal relations; (ii) from a theoretical perspective, we prove that in the framework of so-called partially faithful distributions, our method leads to consistent model selection for almost arbitrary designs and hence for much more general situations than what has been shown for the Dantzig selector, the Lasso or the adaptive Lasso; (iii) from a practical perspective, it can be very valuable to have a "diametrically opposed" method in the tool-kit for high-dimensional data analysis, raising the confidence for relevance of variables if they have been selected by say two very different methods. We will address all these reasons, without prioritizing one over the other.

Our method is a simplification of the PC-algorithm [9] which has been shown to be consistent for estimating high-dimensional directed acyclic graphs [6]. The simplification arises because selecting variables in a linear model is easier than assigning a directed association in a graphical model. In [1] we prove consistency for variable selection in high-dimensional linear models where the number of covariates can greatly exceed the sample size. For the ordinary problem of inferring the non-zero regression coefficients, we introduce and assume the framework of partially

faithful distributions. Partial faithfulness is novel and weaker than the faithfulness condition from graphical models [9, cf.]. Assuming such partial faithfulness in a linear model, which is arguably only a mild requirement, our simplified PC-algorithm is asymptotically consistent under almost arbitrarily ill-posed designs; essentially, we only need that the variables are identifiable in the population case and there are no conditions on the coherence or minimal sparse eigenvalues of the design. Furthermore, causal relations and stronger notions of associations than what is represented by the regression coefficients can be important. In particular, when faithfulness fails to hold, these concepts distinguish themselves very clearly from the regression-type associations. We also prove that for non-faithful distributions, the PC-algorithm is consistent for inferring some strong associations between the response variable and the covariates.

Moreover, the PC-algorithm is computationally feasible in high-dimensional problems: its computational complexity is crudely bounded by a polynomial in p, the dimension of the covariate space, and we illustrate that our implementation in `R` has about the same magnitude for computing time as the LARS-algorithm [4]. Our approach can also be adapted for preliminary reduction of the dimension of the covariate space: we call it "correlation screening" and the method bears some relations to "sure independence screening" [5].

Finally, we compare our PC-algorithm with the Lasso and the Elastic Net [14], and we demonstrate the usefulness of having "diametrically opposed" methods for analyzing a high-dimensional data-set on riboflavin production from bacillus subtilis.

References

[1] P. Bühlmann and M. Kalisch, *Variable selection for high-dimensional models: partial faithful distributions, strong associations and the PC-algorithm*, Preprint (2007).

[2] F. Bunea, A. Tsybakov and M. Wegkamp, *Sparsity oracle inequalities for the Lasso*, Electronic J. of Statistics **1** (2007), 155–168.

[3] E. Candes and T. Tao, *The Dantzig selector: statistical estimation when p is much larger than n*, Ann. Statist. **35** (2007), to appear.

[4] B. Efron, T. Hastie, I. Johnstone and R. Tibshirani, *Least Angle Regression (with discussion)*, Ann. Statist. **32** (2004), 407–451.

[5] J. Fan and J. Lv, *Sure independence screening for ultra-high dimensional feature space*, Preprint (2006).

[6] M. Kalisch and P. Bühlmann, *Estimating high-dimensional directed acyclic graphs with the PC-algorithm*, J. Machine Learning Research **8** (2007), 613–636.

[7] N. Meinshausen and P. Bühlmann, *High-dimensional graphs and variable selection with the Lasso*, Ann. Statist. **34** (2006), 1436–1462.

[8] N. Meinshausen and B. Yu, *Lasso-type recovery of sparse representations for high-dimensional data* (2006), Preprint.

[9] P. Spirtes, C. Glymour and R. Scheines, *Causation, Prediction, and Search*, The MIT Press (2nd ed.) (2000).

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

[11] S.A. van de Geer, *High-dimensional generalized linear models and the Lasso*, Ann. Statist. **36** (2008), to appear.

[12] C.-H. Zhang and J. Huang, *The sparsity and bias of the Lasso selection in high-dimensional linear regression*, Ann. Statist. **36** (2008), to appear.
[13] H. Zou, *The adaptive Lasso and its oracle properties*, J. Amer. Statist. Assoc., **101** (2006), 1418–1429.
[14] H. Zou and T. Hastie, *Regularization and variable selection via the ELastic Net*, J. Royal Statist. Soc., Series B, **67** (2005), 301–320.

Gene-environment interaction studies

Raymond J. Carroll
(joint work with Nilanjan Chartterjee)

Genetic epidemiologic studies often involve investigation of the association between a disease and a candidate genomic region of biologic interest. Typically, in such studies, genotype information is obtained on multiple loci that are known to harbor genetic variations within the region of interest. An increasingly popular approach for analysis of such multi-locus genetic data are haplotype-based regression methods where the effect of a genomic region on disease-risk is modelled through "haplotypes", the combinations of alleles (gene-variants) at multiple loci along individual homologous chromosomes. It is believed that association analysis based on haplotypes, which can efficiently capture inter-loci interactions as well as "indirect association" due to *linkage-disequilibrium* of the haplotypes with unobserved causal variant(s), can be more powerful than more traditional locus-by-locus methods.

A technical problem for haplotype-based regression analysis is that in traditional epidemiologic studies the haplotype information for the study subjects is not directly observable. Instead, locus-specific genotype data are observed, which contain information on the pair of alleles a subject carries on his/her pair of homologous chromosomes at each of the individual loci, but does not provide the "phase information", that is which combinations of alleles appear across multiple loci along the individual chromosomes. In general, the genotype data of a subject will be phase-ambiguous whenever the subject is heterozygous at two or more loci. Statistically, the lack of phase information can be viewed as a special missing data problem.

Recently, a variety of methods have been developed for haplotype-based analysis of case-control data using the logistic regression model [1]-[2]. Two classes of methods, namely "prospective" and "retrospective" have evolved. Prospective methods ignore the retrospective nature of the case-control design. In the classical setting, without any missing data, justification of prospective analysis of case-control data relies on the well known result about the equivalence of prospective and retrospective likelihoods under a semiparametric model that allows the distribution of the underlying covariates to remain completely nonparametric. Even with missing data, the equivalence of the prospective and retrospective likelihood may hold, provided the covariate distribution is allowed to remain unrestricted. For haplotype-based genetic analysis, however, complete nonparametric treatment

of the covariates, including haplotypes, may not be possible due to intrinsic identifiability issues for the phase ambiguous genotype data. Thus, in this setting, the proper retrospective analysis of case-control data requires special attention.

An attractive feature of the retrospective likelihood is that it can enhance efficiency of case-control analysis by directly incorporating certain type of covariate distributional constraints that are natural for genetic epidemiologic studies. The assumptions of Hardy-Weinberg-Equilibrium (HWE) and gene-environment independence are two prime examples of such constraints. The HWE model, which specifies simple relationships between *allele* and *genotype* frequencies at a given chromosomal locus, or between haplotype and diplotype (pair of haplotypes on homologous chromosomes) frequencies across multiple loci, is a natural law for a random mating large stable population. Often, it is also natural to assume that a subject's genetic susceptibility, a factor which is determined at birth, is independent of his/her subsequent environmental exposures. However, if these assumptions are violated in some situations, then retrospective methods can produce serious bias in odds ratio estimates. Thus, there is a need for alternative flexible models for specifying the joint distribution of genetic and environmental covariates that could be used to assess the sensitivity of the retrospective methods to underlying assumptions as well as to develop alternative robust methods.

If it is the underlying biologic units through which a mechanism of gene is determined, then it is natural to allow for direct association between haplotypes and environmental exposures. Moreover, if such association could exist, then quantifying the association between haplotypes and certain type of environmental exposures, such as lifestyle and behaviorial factors, would be of scientific interest.

In this article, we propose methods for retrospective analysis of case-control data using a novel model for the gene-environment distribution that can account for direct association between haplotypes and environmental exposures. We assume a standard logistic regression model to specify the disease risk conditional on diplotypes and environmental exposures. In addition, we assume a polytomous logistic regression model for specifying the population distribution of the diplotypes conditional on the environmental exposures, with the intercept parameters of the model specified in such a way that the *marginal* distribution of the diplotypes can follow certain population genetic constraints such as HWE. Moreover, by exploiting the equivalence of prospective and retrospective odds-ratios under the polytomous regression model, we further incorporate certain constraints on the diplotype-exposure odds-ratio parameters that could reflect specific "mode of effects" for the haplotypes. We allow the marginal distribution of the environmental exposure to remain completely nonparametric.

Under the proposed modelling framework, we describe a "semiparametric" estimating equation method for inference about the finite dimensional parameters of interest, namely the disease odds-ratios, haplotype frequencies and haplotype-exposure odds-ratios. We develop a suitable expectation-maximization (EM) algorithm to account for the phase-ambiguity problem. We study asymptotic theory of the proposed estimator under the underlying semiparametric setting.

We assess the finite sample performance of the proposed estimator based on case-control data that were simulated utilizing haplotype patterns and frequencies obtained from a real study. We also apply the proposed methodology to a case-control study of colorectal adenoma to investigate whether certain haplotypes in the smoking metabolism gene, *NAT2*, could modify smoking-related risk of colorectal adenoma and whether the same haplotypes could influence an individual's susceptibility to smoking as well. A SAS macro is available from the first author to implement the methodology.

References

[1] N. Chatterjee and R. J. Carroll, *Semiparametric maximum likelihood estimation in case-control studies of gene-environment interactions* Biometrika **92** (2005), 399-418.

[2] C. Spinka, R. J. Carroll and N. Chatterjee, *Analysis of case-control studies of genetic and environmental factors with missing genetic information and haplotype-phase ambiguity* Genetic Epidemiology **29** (2005), 108-127.

Order selection in inverse regression models

Gerda Claeskens

(joint work with N. Bissantz, H. Holzmann, A. Munk)

In inverse regression models, $Y = K\mu + \epsilon$, the unknown regression function $\mu(.)$, is not observed directly but only after application of an operator K, which cannot be continuously inverted. For simplicity we assume K to be known. Thus, only noisy, indirect observations Y for the function μ are available. We propose two omnibus test statistics for use in inverse regression problems. Eubank and Hart (1992) first studied lack-of-fit testing based on selecting an appropriate order of an orthogonal series expansion. They named this order selection testing, since indeed one form of the test statistic can be viewed as a test on the selected (or estimated) order of the series. The selected order in such a test is obtained via a modified version of Akaike's (1973) information criterion. A similar type of test, though originally introduced for testing the distribution function in a goodness-of-fit setting, uses instead the Bayesian information criterion (Schwarz, 1978). These tests build on the idea of a Neyman smooth test and were introduced by Ledwina (1994). Both the order selection test and the Neyman smooth test extend naturally to inverse regression modeling, where the orthogonal series expansion is canonically given by the singular value expansion, and the ordering of the singular functions is determined by the magnitude of the corresponding singular values.

We also introduce two model selection criteria which extend the classical AIC and BIC to inverse regression problems. In a simulation study we show that the 'inverse' order selection and Neyman smooth tests outperform their 'direct' counterparts in many cases. The theory is motivated by data arising in confocal fluorescence microscopy. Here, images are observed with blurring, modeled as convolution, and stochastic error at subsequent times. The aim is to reduce the signal to noise ratio by averaging over the distinct images. In this context it

is relevant to decide and test the hypothesis whether the images are still equal, or have changed by outside influences such as moving of the object table. The proposed tests are used for this purpose.

References

[1] H. Akaike, *Information theory and an extension of the maximum likelihood principle*, Second International Symposium on Information Theory (1973), B. Petrov and F. Csáki (editors), 267–281, Akadémiai Kiadó, Budapest.
[2] R. L. Eubank and J. D. Hart, *Testing goodness-of-fit in regression via order selection criteria*, Ann. Statist. **20** (1992), 1412–1425.
[3] T. Ledwina, *Data-driven version of Neyman's smooth test of fit*, J. Amer. Statist. Assoc. **89** (1994), 1000–1005.
[4] G. Schwarz, *Estimating the dimension of a model*, Ann. Statist. **6** (1978), 461–464.

Formal frameworks for causal modelling

A. Philip Dawid

There is currently a wide variety of statistical approaches to causal inference, of greater or less formality, including structural equation models, Rubin's potential response model, and Pearl's graphical causal networks. These are based on a variety of distinct foundations, ingredients, assumptions and methods, and involve a variety of conceptions of the effects of interventions, or of stable relationships across regimes; a variety of views as to the role hypothetical and counterfactual reasoning; and a variety of semantics and uses for algebraic, graphical and other representations. But the different approaches are all in agreement that causal inference requires significant modifications and extensions to standard statistical machinery.

The foundational issues underlying these varying views, and their implications for data analysis, deserve deeper examination than they typically receive. I survey the field from a somewhat idiosyncratic philosophical viewpoint, and argue for a number of heretical views, including the following:

The relationship between a causal model and the empirical world is subtle and often misunderstood. Causal models are often used for purposes that they can not support, such as extraction of causal conclusions from observational data. Important distinctions, such as between prospective and retrospective assignation of cause, are frequently overlooked. Some commonly accepted properties, such as the deterministic nature of variables and the relationships between them, are at variance with traditional statistical approaches and insights. Others are untestable even in principle, but can make important differences to the inferences drawn. As a counterbalance to these shortcomings of the various complex modern approaches, I demonstrate the power of traditional statistical and decision-theoretic tools to address causal issues simply and cleanly

Non-crossing quantile curves

Holger Dette
(joint work with S. Volgushev)

The problem of crossing quantile estimates in quantile regression has been mentioned by numerous authors [see e.g. He (1997) or Koenker (2005) among many others]. In this paper an estimate of conditional quantiles is proposed, that avoids the problem of crossing quantile curves [calculated for various $p \in (0,1)$]. The method uses an initial estimate of the conditional distribution function in a first step and solves the problem of inversion and monotonization with respect to $p \in (0,1)$ simultaneously. For a given initial estimate $\widehat{F}_x(y)$ of the conditional distribution function $F(y|x)$ the (non-crossing) quantile curves are defined by

$$\widehat{F}^{-1}_{x,G}(p) \quad = \quad (G^{-1} \circ \widehat{H}^{-1}_x)(p) \tag{1}$$

where

$$\widehat{H}^{-1}_x(p) = \frac{1}{Nh_d} \sum_{i=1}^{N} \int_{-\infty}^{p} K_d \left(\frac{\widehat{F}_x(G^{-1}(i/N)) - u}{h_d} \right) du \tag{2}$$

G is a given distribution function with $\mathrm{supp}(G) = \mathbb{R}$, K_d is a nonnegative kernel and h_d a corresponding bandwidth. Under some assumptions of regularity it is shown that the weak convergence of

$$a_n(\widehat{F}_x(F_x^{-1}(p)) - F(F_x^{-1}(p)|x) - b_n(F_x^{-1}(p)|x)) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2(F_x^{-1}(p)|x))$$

implies that the corresponding quantile estimator converges also in law, i.e.

$$a_n \left(\widehat{F}^{-1}_{x,G}(p) - F_x^{-1}(p) + \frac{b_n(F_x^{-1}(p)|x)}{F'_x(F_x^{-1}(p))} \right) \xrightarrow{\mathcal{D}} \mathcal{N} \left(0, \frac{\sigma^2(F_x^{-1}(p)|x)}{(F'_x(F_x^{-1}(p)))^2} \right)$$

The performance of the new procedure is illustrated by means of a simulation study and some comparisons with the currently available procedures which are similar in spirit with the proposed method, are presented.

References

[1] H. Dette, N. Neumeyer, K.F. Pilz, *A simple nonparametric estimator of a strictly monotone regression function*, Bernoulli **12** (2006), 469-490.
[2] X. He , *Quantile curves without crossing*, The American Statistician **51** (1997), 186–191.
[3] R. Koenker, *Quantile Regression*, Cambridge University Press (2005).
[4] K. Yu, M. C. Jones, *Local linear quantile regression*, J. Am. Stat. Assoc. **93** (1998), 228-237.

Approximating data

Laurie Davies
(joint work with Arne Kovac, Monika Meise)

The main paradigms of statistics, the frequentist and Bayesian ones, can be described using the simple urn model. An urn contains an unknown number of red and white balls and a finite sample of size n with replacement is taken. The canonical model is the binomial $b(n, \theta)$ where the parameter θ is identified with the proportion p of red balls in the urn. Using this model the Bayesian can introduce concepts such as betting odds, coherence, utility, prior distributions for parameters, the likelihood principle, stopping rules, sufficient statistics and posterior distributions. The frequentist can introduce the concepts of estimation, biased and unbiased, loss functions, statistical tests, optimality, the Neyman-Pearson lemma, likelihood, maximum likelihood, asymptotics and asymptotic optimality. The advantage of this simple model is that it is simple. The model seems perfectly reasonable and there is no problem in identifying the parameter θ with the proportion p. However the very simplicity of the situation is also its weakness as the problem of the relationship between the model and the real world does not arise. The situation changes if we take what is perhaps the next simplest situation, that of a location model. Given data, say measurements of the amount of copper (milligrams per litre) in a sample of drinking water, there is no longer a canonical model. Indeed many different location models $F(\cdot - \theta)$ are consistent with the data and the concepts used for the urn are of no help in deciding which different choices of F are appropriate. One reason for the lack of applicability of the concepts is that they operate in a density (likelihood) based strong topology whereas the adequateness of a model is decided at the level of distribution functions using a weak topology. For this reason a concept of approximation in statistics cannot be likelihood based. The problem of the choice of the model F in the example of the location parameter can only be solved by some form of regularization. The function F should not offer precision for free and this leads to the choice of the least informative model consistent with the data. It the data are consistent with the normal model then this can be chosen as it minimizes the Fisher information and is, in Tukey's sense, bland or hornless. We turn now to non-parametric regression. Suppose we have data $(t_i, y(t_i)), i = 1, \dots n$ and wish to apply the model

$$Y(t) = f(t) + \sigma Z(t), \quad 0 \le t \le 1$$

where $Z(t)$ is standard Gaussian white noise. Based on the behaviour of the partial sums $\sum_{i=j}^{k} Z(t_i)/\sqrt{k-j+1}$, $1 \le i \le j \le n$ we define an approximation region $\mathcal{A}_n$ by

$$\mathcal{A}_n = \left\{ \widehat{f}_n \,:\, \max_I |w_n(\widehat{f}_n, I)| \le \sigma \sqrt{\tau_n(\alpha) \log n} \right\}$$

where

$$w_n(\widehat{f}_n, I) = \frac{1}{\sqrt{|I|}} \sum_{t_i \in I} (y(t_i) - \widehat{f}_n(t_i))$$

$|I|$ is the number of points t_i in the interval I and $\tau_n(\alpha)$ is defined by

$$\mathbf{P}\left(\max_I \frac{1}{\sqrt{|I|}}\Big|\sum_{t_i\in I} Z(t_i)\Big| \le \sigma\sqrt{\tau_n(\alpha)\log n}\right) = \alpha$$

It is easy to show that for data$(t_i, Y(t_i)), i = 1, \dots n$ generated under the model the region $\mathcal{A}_n$ is a universal, exact and non-asymptotic α-confidence region for f. On replacing σ by

$$\sigma_n = 1.048358\text{median}(|y(t_2) - y(t_1)|, \dots, |y(t_n) - y(t_{n-1})|)$$

the region becomes an honest rather an exact α-confidence region for f. Given $\mathcal{A}_n$ all questions relating to adequate models for the data can be answered by some form of regularization within $\mathcal{A}_n$. These can be some form of shape regularization such as minimizing the number of intervals of increase or decrease of all functions in $\mathcal{A}_n$, or they can be some form of smoothness regularization such as minimizing the total variation of a derivative of all functions in $\mathcal{A}_n$, or a combination of both. Rates of convergence and honest confidence bounds can be obtained for any form of regularization. The ideas can be extended to nonparametric density estimation.

References

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

[2] P. L. Davies and A. Kovac, *Densities, Spectral Densities and Modality*, Annals of Statistics **32** (2004), 1093–1136.

[3] P. L. Davies, A. Kovac and M. Meise, *Confidence Regions, Regularization and Non-Parametric Regression*, Technical Report 12, SFB 475, University of Dortmund, 2007.

Bayesian mixed membership models for soft clustering and network analysis

Stephen E. Fienberg

Many applications of statistics involving very large data sets utilize ideas on clustering and classification where units can conceivably belong to multiple groups. Bayesian mixed membership models provide a natural way to address such "soft" clustering and classification problems. These models typically rely on four levels of assumptions: population, sub ject, latent variable, and sampling scheme. Population level assumptions describe a general structure of the population that is common to all subjects. Subject level assumptions specify the distribution of observable responses given the population structure and individual membership scores. Membership scores are usually unknown and hence can also be viewed as latent variables which can be treated as fixed or random in the model. Finally, the last level of assumptions specifies the number of distinct observed characteristics (attributes) and the number of replications for each characteristic.

We describe three applications of mixed membership modeling: (i) to disability indicators from the National Long Term Care Survey, (ii) abstracts and bibliographies of research reports in The Proceedings of the National Academy of Sciences, (iii) protein-protein interactions in yeast. The last application involves extensions to mixed-membership methods that incorporate stochastic block-modeling for network analysis. Our methods include the computation of full posterior distributions for application (i), as well as various forms of variational approximations for applications (ii) and (iii). In the examples, we also discuss issues of model assessment and specification.

References

[1] E. M. Airoldi, S. E. Fienberg, C. Joutard, and T. Love, *Discovering latent patterns with hierarchical Bayesian mixed-membership models*, In P. Poncelet F. Masseglia and M. Teisseire, editors, Data Mining Patterns: New Methods and Applications (2007), page (in press). Idea Group Inc., Hershey, PA.

[2] E. M. Airoldi, D. M. Blei, S, E. Fienberg, and E. P. Xing, *Combining stochastic block models and mixed membership for statistical network analysis*. In Airoldi, E.M., Blei, D.M., Fienberg, S.E., Goldenberg, A., Xing, E.P., and Zheng, A.X., Eds. Statistical Network Analysis: Models, Issues & New Directions (ICML 2006). Lecture Notes in Computer Science, Springer-Verlag, **4503** (2007), 57–74.

[3] E. A. Erosheva, S. E. Fienberg, and C. Joutard, *Describing disability through individual-level mixture models for multivariate binary data*, Annals of Applied Statistics **1**(2) (2007), in press.

[4] E. A. Erosheva, S. E. Fienberg, and J. Lafferty, *Mixed-membership models of scientific publications*, Proceedings of the National Academy of Sciences **101** (2004) (Suppl.1):5220–5227.

Pre-modeling via BART

Edward I. George

(joint work with Hugh A. Chipman, Robert E. McCulloch)

Consider the canonical regression setup where one wants to learn about the relationship between y, a variable of interest, and $x_1, \ldots, x_p$, p potential predictor variables. For the general purposes of discovering the form of $f(x_1, \ldots, x_p) \equiv E(Y \mid x_1, \ldots, x_p)$ and making predictive inference about a future y, we propose an approach called BART (Bayesian Additive Regression Trees). BART approximates f by a Bayesian "sum-of-trees" model where fitting and inference are accomplished via an iterative backfitting MCMC algorithm. By using a large number of trees, which yields a redundant 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. As in boosting, each tree is constrained to be a weak learner that contributes only a small amount to the fit. However, in contrast to boosting, BART is based on a fully Bayes statistical model: a prior and a likelihood. This approach enables a full and accurate assessment of uncertainty in model predictions, while remaining highly competitive in terms of predictive accuracy.

BART can also be viewed in the context of Bayesian nonparametrics. The key idea is to use a model 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.

Lastly, BART can also be used to screen for relevant predictors, thereby providing an essentially nonparametric approach to variable selection in the sense that it does not rely on an initial parametric model assumption for selection. As the BART algorithm moves along, different potential predictors enter the sum-of-trees 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. By varying the size of the sum-of-trees model, BART can identify those subsets of $x_1, \ldots, x_p$ which contain the strongest predictive information, subsets which then may used to obtain a parametric model. BART also provides an omnibus test: the absence of any relationship between y and any subset of $x_1, \ldots, x_p$ is indicated when BART posterior intervals for f reveal no signal.

The catch-up phenomenon

Peter Grünwald
(joint work with Steven de Rooij, Tim van Erven)

We consider inference based on a countable set of models (sets of probability distributions), focusing on two tasks: model selection and model averaging. In model selection tasks, the goal is to select the model that best explains the given data. In model averaging, the goal is to find the weighted combination of models that leads to the best prediction of future data from the same source.

An attractive property of some criteria for model selection is that they are consistent under weak conditions, i.e. if the true distribution P^* is in one of the models, then the P^*-probability that this model is selected goes to one as the sample size increases. BIC [11], Bayes factor model selection, Minimum Description Length (MDL) model selection [2] and prequential model validation [3] are examples of widely used model selection criteria that are usually consistent. However, other model selection criteria such as AIC [1] and leave-one-out cross-validation (LOO) [13], while often inconsistent, do typically yield better predictions. This is especially the case in nonparametric settings, where P^* can be arbitrarily well-approximated by a sequence of distributions in the (parametric) models under consideration, but is not itself contained in any of these. In many such cases, the predictive distribution converges to the true distribution at the optimal rate for AIC and LOO [12, 8], whereas in general BIC, the Bayes factor method and

prequential validation only achieve the optimal rate to within an $O(\log n)$ factor [10, 15, 5]. Here we reconcile these seemingly conflicting approaches [14] by improving the rate of convergence achieved in Bayesian model selection without losing its convergence properties. In this abstract we merely provide an example that gives a novel analysis of the reason why Bayes sometimes converges too slowly; this analysis then leads to a new approach, essentially an extension of a Bayes/MDL-approach, which both achieves consistency and optimal convergence rates. This extension is discussed in the conference paper [4].

Given priors on models and parameters therein, Bayesian inference is based on the posterior distribution that is obtained by conditioning on observed outcomes. In model selection the preferred model is the one with maximum a posteriori probability. In prediction the marginal distributions $p_1, p_2, \ldots$ (defined as $p_k(x^n) = \int_{\theta\in\Theta_k} p_\theta(x^n)w(\theta)\ d\theta$) are weighted according to the posterior, a process called Bayesian Model Averaging (BMA). We denote the resulting distribution $p_{\rm bma}$.

In a sequential setting, the probability of a data sequence $x^n := x_1, \ldots, x_n$ under a distribution p typically decreases exponentially fast in n. It is therefore common to consider $-\log p(x^n)$, which we call the *codelength* of x^n achieved by p. This name refers to the correspondence between codelength functions and probability distributions based on the Kraft inequality, but one may also think of the codelength as the accumulated log loss that is incurred if we sequentially predict the x_i by conditioning on the past, i.e. using $p(\cdot \mid x^{i-1})$ [2, 5, 3, 9]. All logarithms are taken to base 2, allowing us to measure codelength in *bits*.

Prediction using $p_{\rm bma}$ has the advantage that the codelength it achieves on x^n is close to the codelength of $p_{\widehat{k}}$, where $\widehat{k}$ is the index of best of the marginals $p_1, p_2, \ldots$ Namely, given a prior w on model indices, the difference between $-\log p_{\rm bma}(x^n) = -\log(\sum_k p_k(x^n)w(k))$ and $-\log p_{\widehat{k}}(x^n)$ must be in the range $[0, -\log w(\widehat{k})]$, whatever data x^n are observed. Thus, using BMA for prediction is sensible if we are satisfied with doing essentially as well as the best model under consideration. However, it is often possible to combine $p_1, p_2, \ldots$ into a distribution that achieves smaller codelength than $p_{\widehat{k}}$! This is possible if the index $\widehat{k}$ of the best distribution *changes with the sample size in a predictable way.* This is common in model selection, for example with nested models, say $\mathcal{M}_1 \subset \mathcal{M}_2$. In this case p_1 typically predicts better at small sample sizes (roughly, because $\mathcal{M}_2$ has more parameters that need to be learned than $\mathcal{M}_1$), while p_2 predicts better eventually. Figure 1 illustrates this phenomenon. It shows the accumulated codelength difference $-\log p_2(x^n) - (-\log p_1(x^n))$ on "The Picture of Dorian Gray" by Oscar Wilde, where p_1 and p_2 are the Bayesian marginal distributions for the first-order and second-order Markov chains, respectively, and each character in the book is an outcome. Note that the example models $\mathcal{M}_1$ and $\mathcal{M}_2$ are very crude; for this particular application much better models are available. In more complicated, more realistic model selection scenarios, the models may still be wrong, but it may not be known how to improve them. Thus, $\mathcal{M}_1$ and $\mathcal{M}_2$ serve as a simple illustration only. We used uniform priors on the model parameters, but for other common priors similar behaviour can be expected. Clearly p_1 is better for about the first

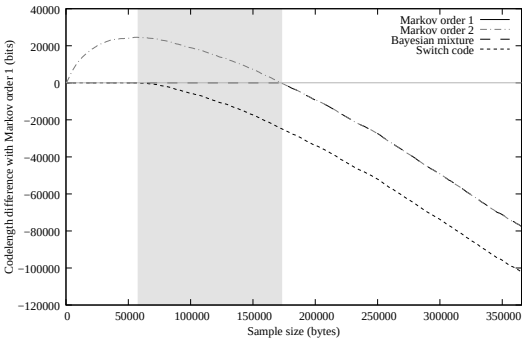

FIGURE 1. The Catch-Up Phenomenon

100 000 outcomes, gaining a head start of approximately 40 000 bits. Ideally we should predict the initial 100 000 outcomes using p_1 and the rest using p_2. However, p_{bma} only starts to behave like p_2 when it *catches up* with p_1 at a sample size of about 310 000, when the codelength of p_2 drops below that of p_1. Thus, in the shaded area p_{bma} behaves like p_1 while p_2 gives higher probability to, and better predictions of, those outcomes: since at $n = 100\,000$, p_2 is 40 000 bits behind, and at $n = 310\,000$, it has caught up, in between it must have outperformed p_1 by 40 000 bits! The general pattern that first one model is better and then another occurs widely, both on real-world data and in theoretical settings. We argue that failure to take this effect into account leads to the suboptimal rate of convergence achieved by Bayes factor model selection and related methods. We have developed an alternative method to combine distributions p_1 and p_2 into a single distribution p_{sw}, which we call the *switch-distribution*. Figure 1 shows that p_{sw} behaves like p_1 initially, but in contrast to p_{bma} it starts to mimic p_2 *almost immediately* after p_2 starts making better predictions; it essentially does this *no matter what sequence* x^n *is actually observed*. p_{sw} differs from p_{bma} in that it is based on a prior distribution on *sequences of models* rather than simply a prior distribution on models. This allows us to avoid the implicit assumption that there is one model which is best at all sample sizes. After conditioning on past observations, the posterior we obtain gives a better indication of which model performs best *at the current sample size*, thereby achieving a faster rate of convergence. Indeed, the switch-distribution is related to earlier algorithms for *tracking the best expert* developed in the universal prediction literature [6]; however, the applications we have in mind and the theorems we prove are completely different. In the conference paper [4] we show that model selection based on the switch-distribution is consistent (Theorem 1), but unlike standard Bayes factor model selection achieves optimal rates of convergence (Theorem 2). We also give a practical algorithm that computes the switch-distribution for K (rather than 2) predictors in $\Theta(n \cdot K)$ time.

References

[1] H. Akaike, *A new look at statistical model identification*, IEEE T. Automat. Contr. **19**(6) (1974), 716–723.
[2] A. Barron, J. Rissanen, and B. Yu, *The minimum description length principle in coding and modeling*, IEEE T. Inform. Theory **44**(6) (1998), 2743–2760.
[3] A. P. Dawid, *Statistical theory: The prequential approach*, J. Roy. Stat. Soc. A, **147**(2) (1984), 278–292.
[4] T. van Erven and P.D. Grünwald and S. de Rooij, *Catching up Faster in Bayesian Model Selection and Model Averaging*, Proceedings NIPS **20** (2007).
[5] P. D. Grünwald, *The Minimum Description Length Principle*, The MIT Press 2007.
[6] M. Herbster and M. K. Warmuth, *Tracking the best expert*, Machine Learning **32** (1998).
[8] K. Li, *Asymptotic optimality of c_p , c_l, cross-validation and generalized cross-validation: Discrete index set*, Ann. Stat. **15** (1987), 958–975.
[9] J. Rissanen, *Universal coding, information, prediction, and estimation*, IEEE T. Inform. Theory **IT-30**(4) (1984), 629–636.
[10] J. Rissanen, T. P. Speed, and B. Yu, *Density estimation by stochastic complexity*, IEEE T. Inform. Theory **38**(2) (1992), 315–323.
[11] G. Schwarz, *Estimating the dimension of a model*, Ann. Stat. **6**(2) (1978), 461–464.
[12] R. Shibata, *Asymptotic mean efficiency of a selection of regression variables*, Ann. I. Stat. Math. **35** (1983), 415–423.
[13] M. Stone, *An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion*, J. Roy. Stat. Soc. B **39** (1977), 44–47.
[14] Y. Yang, *Can the strengths of AIC and BIC be shared?*, Biometrica **92**(4) (2005), 937–950.
[15] Y. Yang, *Model selection for nonparametric regression*, Statistica Sinica **9** (1999), 475–499.

Labour market modelling and hypothesis testing for functional data

Peter Hall

(joint work with Federico A. Bugni, Joel L. Horowitz and George R. Neumann)

Models for employment and wages can be based on economic theories of supply and demand, leading to so-called "equilibrium job search models" that describe the manner in which people move from job to job, or from employment to unemployment and back again. Some of these models are explicitly parametric in nature, and prescribe stochastic processes that could conceivably be good approximations to the real processes that generate observed data. In particular, certain equilibrium search models specify the entire wage process up to a finite-dimensional parameter. In these and other cases it is natural to test the theoretical model against the data.

However, while the likelihood for such data can often be written down exactly, and maximum likelihood estimators derived, it is far from clear how to determine whether the model is adequate. In a wide range of related settings the approach that is taken is to try to simplify the problem, for example by reducing the number of degrees of freedom or the number of dimensions. However, sometimes greater insight can be gained by representing the data in a way that is arguably more complex than the data, for example by representing data vectors via graphs of random functions, in place of plots of data points. The former is the approach we take. Rather than directly address the goodness of fit problem for vectors of data,

we focus on the random wage paths that are produced by graphing functionals of those data.

A theoretical model for a stochastic process explicitly or implicitly specifies the probability distribution of the random functions (or sample paths) that represent realisations of the process. If the model depends on an unknown, finite-dimensional parameter then the specification is up to the value of this parameter. Functional data can be used to construct an empirical analog of the probability distribution of the random functions (the empirical distribution of the data). Therefore, a test of the hypothesis that the postulated model generated the data can be implemented by comparing the empirical and theoretical distributions of the sample paths. This amounts to testing a finite-dimensional parametric model of a probability distribution, against a nonparametric alternative.

When the random variable of interest is finite-dimensional, the Cramér-von Mises and Kolmogorov-Smirnov tests, among many others, can be used for this purpose. We generalise the Cramér-von Mises test to distributions of random functions, or infinite-dimensional random variables, that depend on an unknown finite-dimensional parameter. Novel aspects of this approach include the introduction of functional data methods for specification testing in econometrics, and the development of parametric bootstrap methods that facilitate the use of techniques based on integration over function spaces. The functional data view offers new ways of conceptualising specification testing problems in econometrics, and suggests new approaches to testing continuous-time models, such as models of financial data, that are quite different from the equilibrium search model that motivates the present work.

Specifically, suppose that the distribution of a random function Y depends on an unknown, finite-dimensional parameter θ, and that we have a random sample $\{X_1, \ldots, X_n\}$ of n realisations of a random function X, that may be distributed as Y for some value of θ. We develop a Cramér-von Mises type test of the null hypothesis, H_0, that the distribution of X is identical to that of Y for some unspecified value of θ. We present the test statistic and explain how to compute it; we study the test statistic's asymptotic distribution under fixed and local alternative hypotheses; and we introduce a bootstrap procedure for computing critical values for the test. Properties and performance of the method are illustrated by applying it to the equilibrium job search model introduced and developed by Mortensen [4], Burdett and Mortensen [2], Bowlus, Kiefer, and Neumann [1], and Christensen, Lentz, Mortensen, Neumann and Werwatz [3]. This model aims to explain the frequencies and durations of spells of unemployment, as well as the distribution of wages among employed individuals. In particular, it provides an explanation for why seemingly identical individuals have different wages. One of the model's outputs is a random function, Y say, that gives an individual's wage as a function of time up to an unknown, vector-valued parameter. The model is tested in the context of wage curves computed from data from the National Longitudinal Survey of Youth, and shown not to provide a convincing fit.

References

[1] A.J. Bowlus, N.M. Kiefer and G.R. Neumann, *Equilibrium search models and the transition from school to work*, International Economic Review **42** (2001), 317–343.
[2] K. Burdett and D.T. Mortensen, *Wage differentials, employer size, and unemployment*, International Economic Review **39** (1998), 257–273.
[3] B.J. Christensen, R. Lentz, D.T. Mortensen, G.R. Neumann and A. Werwatz, *On-the-job search and the wage distribution*, Journal of Labor Economics **23** (2005), 31–58.
[4] D.T. Mortensen, *Equilibrium wage distributions: a synthesis*, in: Panel Data and Labor Market Studies, J. Hartog, G. Ridder, and J. Theeuwe, eds., pp. 279–296. North-Holland: Amsterdam.

Adaptive choice of time varying copulae

Wolfgang Härdle
(joint work with Enzo Giacomini, Vladimir Spokoiny)

1. Introduction

Time series of financial data are high dimensional and have typically a non-Gaussian behavior. The standard modelling approach based on properties of the multivariate normal distribution therefore often fails to reproduce the stylized facts (i.e. fat tails, asymmetry) observed in returns from financial assets.

Modelling distributions with copulae avoids the "procrustean bed" of normality assumptions, producing better fits of the empirical characteristics of financial returns. A natural extension is to apply copulae in a dynamic framework with conditional distributions modelled by copulae with time varying parameters. The question though is how to steer the time varying copulae parameters.

In this paper we follow a semiparametric approach, *locally* selecting the time varynig copula parameter. The choice is performed via an *adaptive estimation* under the assumption of local homogeneity: for every time point there exists an interval of time homogeneity in which the copula parameter can be well approximated by a constant. This interval is recovered from the data using local change point analysis.

The obtained time varying dependence structure can be used in financial engineering applications. Using copulae with adaptively estimated dependence parameters we estimate the Value-at-Risk (VaR) from DAX portfolios over time. As benchmark procedure we choose *RiskMetrics*, a methodology based on conditional normal distributions with a GARCH specification for the covariance matrix. Backtesting underlines the improved performance of the proposed *adaptive time varying copulae fitting*.

2. Value-at-Risk and Copulae

The VaR at level α from a portfolio $w \in \mathbb{R}^d$ is defined as the $\alpha-$ quantile from the distribution of L_t (the P&L function) and depends on the specification of the

d-dimensional distribution of its risk factors increments, here log-returns X_t. In the copulae based approach the log-returns are modelled as:

$$X_{t,j} = \mu_{t,j} + \sigma_{t,j}\varepsilon_{t,j} \tag{1}$$

where $\mu_{t,j} = E[X_{t,j} \mid \mathcal{F}_{t-1}]$ and $\sigma^2_{t,j} = E[X^2_{t,j} \mid \mathcal{F}_{t-1}]$, $j = 1, \ldots, d$. The standardised innovations $\varepsilon_t = (\varepsilon_{t,1}, \ldots, \varepsilon_{t,d})^\top$ have joint cdf F_{ε_t} given by

$$F_{\varepsilon_t}(x_1, \ldots, x_d) = C_\theta\{F_{t,1}(x_1), \ldots, F_{t,d}(x_d)\} \tag{2}$$

where $F_{t,j}$ is the cdf of $\varepsilon_{t,j}$ and C_θ is a *copula* belonging to a parametric family $\mathcal{C} = \{C_\theta, \theta \in \Theta\}$. For details on the above model specification see [2] and [3].

To obtain the Value-at-Risk in this set up, the copula dependence is estimated from a sample of log-returns and used to generate P&L samples. Their quantiles at different levels are the estimators for the Value-at-Risk, see [4].

3. Modelling with Time Varying Copulae

In fact, the cdf F_{ε_t} from (2) is modelled as $F_{t,\varepsilon_t} = C_{\theta_t}\{F_{t,1}(\cdot), \ldots, F_{t,d}(\cdot)\}$ with probability measure P_{θ_t}. In order to estimate the copula parameter we choose an interval of homogeneity employing a local parametric fitting approach as introduced by [1], [5] and [6].

The idea is to select for each time point t_0 an interval $I_{t_0} = [t_0 - m_{t_0}, t_0]$ such that θ_t can be well approximated by a constant value θ. The aim is to obtain I_{t_0} as close as possible to the "oracle" interval, defined as the largest interval $I = [t_0 - m^*_{t_0}, t_0]$ for which the *small modelling bias condition (SMB)*:

$$\Delta_I(\theta) = \sum_{t \in I} \mathcal{K}(P_{\theta_t}, P_\theta) \leq \Delta \tag{3}$$

for some $\Delta \geq 0$ holds. Here θ is constant and $\mathcal{K}(\cdot, \cdot)$ denotes the *Kullback-Leibler divergence*. In the oracle interval, the parameter $\theta_{t_0} = \theta_t|_{t=t_0}$ can be "optimally" estimated from $I = [t_0 - m^*_{t_0}, t_0]$.

The adaptive Local Change Point (LCP) detection procedure "mimics" the "oracle" in the sense that it delivers the same accuracy of estimation as the "oracle" one. For a given point t_0 the LCP starts with a family of nested intervals $I_0 \subset I_1 \subset \ldots \subset I_K = I_{K+1}$ of the form $I_k = [t_0 - m_k, t_0]$. Every interval I_k leads to an estimate $\widetilde{\theta}_k$ of the copula parameter θ_{t_0}. The LCP sequentially tests the homogeneity hypothesis for the copula parameter (i.e. $\theta_t = \theta$) against change point alternative in each interval I_k using critical values $\mathfrak{z}_k$. In case of rejection or if the largest possible interval is reached the procedure stops and $I_{\widehat{k}}$ denotes the latest accepted interval.

The critical values $\mathfrak{z}_k$ are sequentially selected by Monte Carlo simulation as the minimal values providing:

$$\boldsymbol{E}_{\theta^*}\big|L_{I_k}(\widetilde{\theta}_k, \widehat{\theta}_k)\big|^{1/2} \leq \rho\mathfrak{R}(\theta^*), \qquad k = 1, \ldots, K, \qquad \theta^* \in \Theta. \tag{4}$$

	RM		MW		LCP	
			α			
	0.05	0.01	0.05	0.01	0.05	0.01
$A_{\mathcal{W}}$	0.23	0.45	0.11	−0.49	0.11	−0.36
$D_{\mathcal{W}}$	0.04	0.14	0.06	0.08	0.06	0.10
$A_{\mathcal{W}}$	0.16	0.57	−0.10	−0.65	−0.09	−0.65
$D_{\mathcal{W}}$	0.04	0.16	0.06	0.09	0.06	0.08

TABLE 1. Average relative exceedance error over portfolios $A_{\mathcal{W}}$ and corresponding standard deviation $D_{\mathcal{W}}$ for 2 groups of DAX stocks across levels α and methods

where L_I is the log-likelihood corresponding to interval I, $\widehat{\theta}_k = \widetilde{\theta}_k \mathbf{1}_{\{k \le \widehat{k}\}} + \widetilde{\theta}_{\widehat{k}} \mathbf{1}_{\{k > \widehat{k}\}}$ and $\mathfrak{R}(\theta^*)$ is the risk of the non-adaptive estimate $\widetilde{\theta}_k$:

$$\mathfrak{R}(\theta^*) = \max_{k \ge 1} \boldsymbol{E}_{\theta^*} \left| L_{I_k}(\widetilde{\theta}_k, \theta^*) \right|^{1/2}.$$

For details, see [5]. The theoretical results from [6] indicate that the LCP procedure provides the rate optimal estimation of the underlying parameter when this smoothly varies with time. It has also been shown that the procedure is very sensitive to structural breaks, providing the minimal possible delay in detection of changes.

4. Empirical Results

The VaR from 2 groups of portfolios composed of 6 German stock is estimated based on time varying Clayton copulae and *RiskMetrics* (RM) approaches. The time varying copula parameters are selected by Local Change Point (LCP) and moving window (MW) procedures. At each time t the estimated Value-at-Risk at level α for a portfolio w is compared with l_t, the *exceedance ratio* is given by

$$\widehat{\alpha}_w(\alpha) = \frac{1}{T} \sum_{t=1}^{T} \mathbf{1}_{\{l_t < \widehat{VaR}_t(\alpha)\}}$$

and the *relative exceedance error* by

$$e_w = \frac{\widehat{\alpha}_w - \alpha}{\alpha}.$$

We compute e_w for a set $\mathcal{W}$ of random portfolios. The average relative exceedance error over portfolios $A_{\mathcal{W}}$ and the corresponding standard deviation $D_{\mathcal{W}}$ (table 1) are used to evaluate the performances of the time varying copulae and *RiskMetrics* methods in VaR estimation.

References

[1] D. Belomestny and V. Spokoiny, *Spatial Aggregation of Local Likelihood Estimates with Applications to Classification*, The Annals of Statistics, to appear.

[2] X. Chen and Y. Fan, *Estimation of Copula-Based Semiparametric Time Series Models*, Journal of Econometrics, **130**(2) (2006), 307–335.

[3] X. Chen, Y. Fan and V. Tsyrennikov, *Efficient Estimation of Semiparametric Multivariate Copula Models*, Journal of the American Statistical Association, **101** (2006), 1228–1240.

[4] E. Giacomini and W Härdle, *Value-at-Risk Calculations with Time Varying Copulae*, Bulletin of the International Statistical Institute, Proceedings of the 55th Session, (2005).

[5] J. Polzehl and V. Spokoiny, *Aggregation-Separation Approach for Likelihood Estimation*, Probability Theory and Related Fields, **135** (2006), 335–362.

[6] V. Spokoiny, *Local Parametric Methods in Nonparametric Estimation*, Springer-Verlag, Berlin, Heidelberg, 2007, to appear.

Model selection for cube root asymptotics

Nils Lid Hjort

1. Background: standard likelihood theory for regular models

It is convenient to first summarise two central results from the large-sample theory for likelihoods in regular parametric families, pertaining respectively to the behaviour of maximum likelihood (ML) estimators and the derivation of the Akaike information criterion AIC. The purpose of this paper is to investigate what similar (but more difficult) arguments lead to for the more complicated types of models that are associated with cube root asymptotics, cf. Kim and Pollard (1990).

For i.i.d. observations $X_1, \ldots, X_n$ from a data generating density f, for which we employ a parametric family of approximations $f(x, \theta)$, introduce

$$A_n(\theta) = n^{-1}\ell_n(\theta) = n^{-1}\sum_{i=1}^{n} \log f(X_i, \theta),$$

writing ℓ_n for the log-likelihood function. It converges pointwise a.s. to $A(\theta) = \int f \log f_\theta \, \mathrm{d}y$. Under regularity conditions, the ML estimator $\widehat{\theta} = \mathrm{argmax}(A_n)$ tends a.s. to $\theta_0 = \mathrm{argmax}(A)$, which is also the least false parameter value under Kullback–Leibler divergence.

A good model is one for which the attained Kullback–Leibler distance from g to $f(\cdot, \widehat{\theta})$ is small, which is tantamount to a large value of $A(\widehat{\theta})$. The AIC emerges as an attempt to estimate this quantity unbiasedly. For this purpose, consider the initial estimator $A_n(\widehat{\theta}) = n^{-1}\ell_n(\widehat{\theta})$. To see the amount with which $A_n(\widehat{\theta})$ tends to overestimate $A(\widehat{\theta})$, consider the random function

$$\begin{aligned} H_n(s) &= n\{A_n(\theta_0 + s/\sqrt{n}) - A_n(\theta_0)\} \\ &= \sum_{i=1}^{n} \{\log f(X_i, \theta_0 + s/\sqrt{n}) - \log f(X_i, \theta_0)\} \\ &= s^{\mathrm{t}}\sqrt{nU}_n - \tfrac{1}{2}s^{\mathrm{t}}J_n s + o_p(1), \end{aligned}$$

where $\overline{U}_n = n^{-1}\sum_{i=1}^n u(X_i,\theta_0)$ is the average of the score functions at θ_0, and J_n converges in probability to

$$J = -A''(\theta_0) = -\mathrm{E}_f \partial^2 \log f(X,\theta_0)/\partial\theta\partial\theta^{\mathrm{t}}.$$

Under weak regularity conditions, H_n tends in distribution to the random process $H(s) = s^{\mathrm{t}}U - \frac{1}{2}s^{\mathrm{t}}Js$, where $U \sim \mathrm{N}_p(0,K)$, with $K = \mathrm{Var}_f u(X,\theta_0)$ and p the parameter length.

We may derive two basic results from this. The first is that

$$\begin{aligned} M_n = \sqrt{n}(\widehat{\theta}-\theta_0) = \mathrm{argmax}(H_n) &\to_d M \\ &= \mathrm{argmax}(H) = J^{-1}U \sim \mathrm{N}_p(0, J^{-1}KJ^{-1}), \end{aligned}$$

a classic result about ML behaviour for large n. The second, with some modest efforts, is that

$$A_n(\widehat{\theta}) - A(\widehat{\theta}) = \overline{\varepsilon}_n + n^{-1}W_n,$$

where $\overline{\varepsilon}_n$ is the average of zero-mean variables and

$$\begin{aligned} W_n &= H_n(M_n) - n\{A(\widehat{\theta}) - A(\theta_0)\} \\ &\to_d H(M) - \tfrac{1}{2}M^{\mathrm{t}}A''(\theta_0)M = H(M) + \tfrac{1}{2}M^{\mathrm{t}}JM = U^{\mathrm{t}}J^{-1}U. \end{aligned}$$

Thus $A_n(\widehat{\theta})$ tends to overshoot its target $A(\widehat{\theta})$ with a random amount having mean value close to p^*/n, where $p^* = \mathrm{Tr}(J^{-1}K)$. This is the rationale behind using

$$\mathrm{AIC} = \ell_n(\widehat{\theta}) - \widehat{p}^*,$$

with $\widehat{p}^*$ any sensible estimator of p^*. See Claeskens and Hjort (2008, Ch. 2).

2. A quantile-based histogram model

Suppose inference is to be carried out for some density f with cdf F on the real line. Consider its quantiles $q_j = F^{-1}(j/k)$ for $j = 1,\ldots,k-1$. Treating these as unknown parameters, a model emerges by taking $Q = F^{-1}$ as a linear interpolator between the $Q(j/k) = q_j$. The corresponding F is also linear between quantile points q_j, with density

$$f_k(x) = \frac{1}{k}\frac{1}{q_j - q_{j-1}} \quad \text{for } x \in (q_{j-1}, q_j).$$

This is one of the pyramid models worked with in Hjort and Walker (2008), where the fine-ness parameter k may also increase with sample size.

The point is now that this family of densities is outside the standard regularity scope associated with the theory summarised in Section 1, and is in fact instead in the realm of cube root asymptotics. To indicate how this comes about, consider

$$A_n(q) = \sum_{j=1}^k \{F_n(q_j) - F_n(q_{j-1})\}\log(q_j - q_{j-1}),$$

with F_n the empirical cdf. The ML estimator $\widehat{q} = (\widehat{q}_1, \ldots, \widehat{q}_{k-1})$ is the vector that minimises A_n, and is consistent for the least false parameter vector q^0 that minimises

$$A(q) = \sum_{j=1}^{k} \{F(q_j) - F(q_{j-1})\} \log(q_j - q_{j-1}),$$

the limit of A_n. The key process to work with is now

$$H_n(s) = n^{2/3}\{A_n(q^0 + s/n^{1/3}) - A_n(q^0)\}, \quad \text{with } s = (s_1, \ldots, s_{k-1})^{\mathrm{t}}.$$

Working with densities on $[0,1]$, for convenience, and writing $q_0 = 0$ and $q_k = 1$, one may prove convergence in distribution

$$H_n(s) \to_d H(s) = \tfrac{1}{2} s^{\mathrm{t}} A''(q^0) s + \sum_{j=1}^{k-1} d_j f(q_j^0)^{1/2} W_j^*(s_j),$$

where $d_j = a_j(q^0) - a_{j-1}(q^0)$ and $a_j(q) = \log(q_j - q_{j-1})$. Also, the W_j^* are independent two-sided Brownian motions.

As a corollary to this key result one obtains

$$n^{1/3}(\widehat{q} - q^0) \to_d (M_1, \ldots, M_{k-1})^{\mathrm{t}} = \mathrm{argmin}(H).$$

The limit distribution is complicated, and generalises the so-called Chernoff distribution that Groeneboom (1989) and co-authors have worked with, to higher dimensions. The distribution depends on quantities related to the unknown density f, but may all be estimated consistently, making in principle inference and confidence intervals etc. possible, via simulations from the estimated distribution.

3. Selecting the fine-ness of the quantile histogram

The next challenge is to construct a mechanism for selecting the degree k of fineness for the quantile-based histogram estimator f_k. The AIC and BIC theories do not work as such, since the model family is not smooth enough. One may attempt to follow the line of arguments for the classic case, as per Section 1. With appropriate additional efforts, exploiting the H_n process of Section 2, this may be seen to lead to the following procedure. Let for each k

$$\mathrm{CIC}(k) = n \log k + \sum_{j=1}^{k} N_j(\widehat{q}_{j-1}, \widehat{q}_j) \log(\widehat{q}_j - \widehat{q}_{j-1}) + n^{1/3} \widehat{k}^*,$$

with $N_j(a,b) = F_n(b) - F_n(a)$ the number of data points falling inside (a,b). Also, $\widehat{k}^*$ is any consistent estimator of the quantity k^*, the mean value of

$$W = \tfrac{1}{2} M^{\mathrm{t}} A''(q^0) M - H(M) = -\sum_{j=1}^{k-1} d_j f(q_j^0)^{1/2} W_j^*(M_j).$$

This Cube-root Information Criterion is then arguably the natural parallel to the model-robust AIC method outlined at the end of Section 1.

The CIC method as given here is of course constructed specifically for the quantile-based histogram model of Section 2. Its scope is considerably broader,

however, since there are many important statistical models that exhibit the same cube root asymptotics aspects; cf. again Kim and Pollard (1990). A case in point is that of locally constant regression curves with unknown split points, where methods of Banerjee and McKeague (2007) may be generalised and combined with those of the present contribution to provide an instrument for selecting the right amount of jumps in the best approximating model.

References

[1] M. Banerjee and I.W. McKeague, *Confidence sets for split points in decision trees*, Annals of Statistics **35**(2007), 543–574.

[2] G. Claeskens and N.L. Hjort, *Model Selection and Model Averaging*, Cambridge University Press (2008).

[3] P. Groeneboom, *Brownian motion with a parabolic drift and Airy functions*, Probability Theory and Related Fields **81**(1989), 79–110.

[4] N.L. Hjort and S.G. Walker, *Quantile pyramids for Bayesian nonparametrics*, Annals of Statistics (2008), to appear.

[5] K. Kim and D. Pollard, *Cube root asymptotics*, Annals of Statistics **18** (1990), 191–219.

Multiresolution and model choice

Arne Kovac

We consider various settings of the nonparametric regression problem where for given data $y_1, \ldots, y_n$ at time points $t_1, \ldots, t_n$ we require an approximation f that is simple and close to the data. Most approaches develop first an algorithm that takes the data and some additional parameters like bandwidth and kernel function for kernel estimators. In a second step another method is developed for choosing the additional parameters, very often based on minimizing error criteria on test beds like cross-validation. Typically these methods do not produce simple approximations for complex data sets.

In this talk we study approaches that work the other way round and define first a criterion for approximation, giving rise to a set of functions each being an adequate model for the data. In a second step we aim to find a particular simple function among them and try to minimize measures such as the number of local extreme values.

The multiresolution criterion has turned out to be useful for defining approximation. Given noisy data $y_1, \ldots, y_n$ we require a function f to satisfy

$$\left| \sum_{i \in I} (y_i - f_i) \right| < w_I \cdot \sigma \tag{1}$$

with $w_I = \sqrt{|I| \cdot 2 \log(n)}$ for all intervals I of some family $\mathcal{I}$ of subintervals of $\{1, \ldots, n\}$ (Davies and Kovac, 2001; Davies, Kovac and Meise, 2007). This criterion is very strict in the sense that approximations from most popular smoothing methods like smoothing splines with cross validation, adaptive weights smoothing or kernel estimators using local plug-in bandwidths do not usually satisfy this criterion for complex data sets. Wavelet thresholding equipped with the universal

$\tau = \sqrt{2\log(n)}$ threshold (Donoho et al, 1995) have residuals that satisfy similar multiresolution conditions, but usually still hurt some of the multiresolution conditions in (1).

By replacing $y_i - f_i$ with terms such as $\text{sign}(y_i - f_i)$ (Kovac, 2002) or, more generally, $R_i'(\widehat{f}_i)$ with data-dependent functions R_i (Dümbgen and Kovac, 2005) the multiresolution criterion can be adapted to situations with outliers, quantile regression or Poisson regression. An extension to inverse problems is also straightforward: Assume that K is some linear operator and that we want to use Kf instead of f to approximate the data. Then we require a function to satisfy

$$\left|\sum_{i\in I}(y_i - (Kf)_i)\right| < w_I \cdot \sigma \quad \text{for all } I \in \mathcal{I}.$$

The multiresolution criterion can also be used in the context of estimating parameters of an ordinary differential equation. Here we model the data y as noisy observations from an ODE

$$\dot{x}(t) = f(x, u, t|\theta)$$

and want to estimate θ. Again it makes sense to only allow values for θ such that the residuals of x satisfy the multiresolution criterion.

Extending the multiresolution criterion to two or more dimensions is not straightforward, one possibility is to use a decomposition of the residuals using wedgelets (Polzehl and Spokoiny, 2003).

There are several possible ways for maximizing simplicity among all adequate functions. One way consists in minimising total variation (Davies, Kovac and Meise, 2007):

$$\sum_{i=1}^{n-1} |f_{i+1} - f_i| = \min \quad \text{s.t. } f \text{ satisfies (1)}.$$

This leads to a linear program which can be computationally relatively expensive for some data sets. The computational complexity of problems like

$$\sum_{i=1}^{n} R_i(f_i) + \sum_{i=1}^{n-1} \lambda_i |f_{i+1} - f_i|$$

is considerably smaller and is for common choices of R_i not larger than $O(n\log(n))$ using a generalization of the taut string algorithm (Dümbgen and Kovac, 2005). The local penalty parameters can be chosen by the local squeezing technique (Davies and Kovac, 2001) to make sure that the solution satisfies the multiresolution criterion. Finally by using quick update steps it is possible to calculate the solution for the first n data from the solution for the first $n-1$ data without recalculating most of the solution. This allows an extension to online processing (Kovac and Wei, 2007).

References

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

[2] P. L. Davies, A. Kovac and M. Meise, *Confidence Regions, Regularization and Non-Parametric Regression.*, Technical report (2007).

[3] L. Dümbgen and A. Kovac, *Extensions of Smoothing via Taut Strings*, Technical report (2005).

[4] D. L. Donoho, I. .M. Johnstone, G. Kerkyacharian and D. Picard, *Wavelet shrinkage: asymptopia?*, Journal of the Royal Statistical Society, Ser. B **57** (1995), 371–394.

[5] A. Kovac, *Robust nonparametric regression and modality.* in: Developments in Robust Statistics, R. Dutter, P. Filzmoser, U. Gather, P. Rousseeuw (eds.), Physica, Heidelberg, 218–227.

[6] A. Kovac and Y. Wei, *A taut string method for online data*, Technical report (2007).

[7] J. Polzehl and V. Spokoiny, *Image denoising: Pointwise adaptive approach*, Annals of Statistics **31** (2003), 30–57.

Least squares type estimation of the transition density of a particular hidden Markov chain

Claire Lacour

We consider the following additive hidden Markov model:

$$Y_i = X_i + \varepsilon_i \qquad i = 1, \dots, n+1$$

with $(X_i)_{i\geq 1}$ a real-valued Markov chain, $(\varepsilon_i)_{i\geq 1}$ a sequence of independent and identically distributed variables and $(X_i)_{i\geq 1}$ and $(\varepsilon_i)_{i\geq 1}$ independent. Only the variables $Y_1, ..., Y_{n+1}$ are observed. We assume that the transition of the Markov chain , i.e. the distribution of X_{i+1} knowing X_i, has a density Π, defined by $\Pi(x,y)dy = P(X_{i+1} \in dy | X_i = x)$. The aim is to estimate this transition density Π on a compact set $A_1 \times A_2$.

This model belongs to the class of hidden Markov models, and is also similar to the so-called convolution model. As proceeded for this model, we use extensively the Fourier transform. The restrictions on the error distribution and the rate of convergence obtained for our estimator are also of the same kind.

The distribution of the noise $(\varepsilon_i)_{i\geq 1}$ is assumed to be entirely known. Moreover, we assume that its characteristic function q_ε^* verifies the assumption

$$\text{There exist } \gamma > 0 \text{ and } k_0 > 0 \text{ such that } \forall x \in \mathbb{R} \ |q_\varepsilon^*(x)| \geq k_0 (x^2+1)^{-\gamma/2}.$$

Among the so-called ordinary smooth noises, we can cite the Laplace distribution, the exponential distribution and all the Gamma or symmetric Gamma distributions.

We also assume that the Markov chain is irreducible, positive recurrent and stationary with unknown density f. This stationary density is assumed to be bounded and verifies the condition

$$\forall x \in A_1 \quad f(x) \geq f_0 > 0$$

The process (X_i) is assumed to be geometrically β-mixing. Many examples of Markov chains satisfying these assumptions are given in [4]

The estimation of the transition density of this kind of hidden Markov chain is studied in [2]. His estimator is based on thresholding of a wavelet-vaguelette decomposition. The drawback of this estimator is that it does not achieve the minimax rate because of a logarithmic loss. Clémençon[5] describes an estimation procedure by quotient of an estimator of the joint density and an estimator of the stationary density f. The minimax rate is reached by this estimator if we assume that f and Πf have the regularity α. But f can be much less regular than Π. Our aim is then to find an estimator of the transition density which does not have these disadvantages.

To estimate Π, we use an original contrast inspired by the least squares contrast. The first idea is to connect our problem with the regression model. For any function G, we can write

$$G(X_{i+1}) = \left(\int \Pi(.,y)G(y)dy \right)(X_i) + \eta_{i+1}$$

where $\eta_{i+1} = G(X_{i+1}) - \mathbb{E}[G(X_{i+1})|X_i]$. Then, for all function G, we can consider $\int \Pi G$ as a regression function. The least squares contrast to estimate this regression function, if the X_i were known, should be $(1/n)\sum_{i=1}^n [t^2(X_i) - 2t(X_i)G(X_{i+1})]$. If $\int G^2 = 1$, this contrast can be written $(1/n)\sum_{i=1}^n [\int T^2(X_i,y)dy - 2T(X_i,X_{i+1})]$ by setting $T(x,y) = t(x)G(y)$ i.e. T such that $\int T(x,y)G(y)dy = t(x)$. It is this contrast that is used in [4] but in our case, only the $Y_1,\dots,Y_{n+1}$ are known. So we introduce two operators V and Q such that $\mathbb{E}[V_T(Y_i,Y_{i+1})|X_i,X_{i+1}] = T(X_i,X_{i+1})$ and $\mathbb{E}[Q_{T^2}(Y_i)|X_i] = \int T^2(X_i,y)dy$. It leads to the following contrast:

$$\gamma_n(T) = \frac{1}{n}\sum_{i=1}^n [Q_{T^2}(Y_i) - 2V_T(Y_i,Y_{i+1})]. \tag{1}$$

The operators Q and V are found by a fast computation using the Fourier transform which yields $V_T^*(u,v) = T^*(u,v)/(q_\varepsilon^*(-u)q_\varepsilon^*(-v))$ and $Q_T^*(u) = V_T^*(u,0)$.

A collection of estimators is then defined by minimization of this contrast on wavelet spaces S_m. We use compactly supported wavelets on the interval described in [3], so that the functions in S_m are all supported in the compact $A_1 \times A_2$.

The minimization of the contrast on S_m needs further computation. Indeed, by denoting (ψ_λ) a basis of S_m, a function $\widehat{\Pi}(x,y) = \sum \widehat{a}_\lambda \psi_\lambda(x,y)$ minimizes the contrast (1) if and only if $G\widehat{A} = Z$ with

$$G = \left(\frac{1}{n}\sum_{i=1}^n Q_{\psi_\lambda\psi_\mu}(Y_i) \right)_{\lambda,\mu}; \quad Z = \left(\frac{1}{n}\sum_{i=1}^n V_{\psi_\lambda}(Y_i,Y_{i+1}) \right)_\lambda; \quad \widehat{A} = (\widehat{a}_\lambda)_\lambda$$

But G is not necessarily invertible, so we introduce the set $\Gamma = \left\{ \min \mathrm{Sp}(G) \geq \frac{2}{3}f_0 \right\}$ where Sp denotes the spectrum, i.e. the set of the eigenvalues of the matrix and f_0 is the lower bound of f on A_1. On Γ, G is invertible and γ_n is convex so that the minimization of γ_n admits the solution $A = G^{-1}Z$. Thus we define $\widehat{\Pi}_m = \arg\min_{T\in S_m} \gamma_n(T)\mathbb{1}_\Gamma$.

A method of model selection inspired by [1] and based on contrast (1) is used to build an adaptive estimator. A data driven choice of model is performed via the minimization of a penalized criterion. The chosen model is the one which minimizes the empirical risk added to a penalty function. Our definitive estimator is

$$\widetilde{\Pi} = \begin{cases} \widehat{\Pi}_{\widehat{m}} & \text{if } \|\widehat{\Pi}_{\widehat{m}}\|_2 \leq n^{1/2}, \\ 0 & \text{else.} \end{cases}$$

with

$$\widehat{m} = \arg \min_{m \in \mathcal{M}_n} \{\gamma_n(\widehat{\Pi}_m) + \text{pen}(m)\}.$$

In most cases in estimation of mixing processes, a unknown term, reflecting the dependence between the variables, appears in the penalty. Here a conditioning argument allows to lead us back to independent variables and thus to avoid such a mixing term in the penalty. For an ordinary smooth noise with regularity γ, whe choose the following penalty

$$\text{pen}(m) \geq \frac{K_0}{f_0} \frac{D_m^{4\gamma+2}}{n}$$

and we obtain the rate of convergence $n^{-\alpha/(2\alpha+4\gamma+2)}$ if the transition Π is supposed to belong to a Besov space with regularity α. Our estimator is then better than the one of [2] which achieves only the rate $(\ln(n)/n)^{\alpha/(2\alpha+4\gamma+2)}$. Moreover this rate is obtained without supposing known the regularity of f, our estimator is then adaptive.

References

[1] A. Barron, L. Birgé and P. Massart, *Risk bounds for model selection via penalization*, Probab. Theory Related Fields **13**(3) (1999), 301–413.

[2] S. Clémençon, *Nonparametric estimation for some specific classes of hidden Markov models*, Preprint Modal'X n° 03-9 (2003). `http://www.u-paris10.fr/65897276/0/fiche__pagelibre/`

[3] A. Cohen, I. Daubechies and P. Vial, *Wavelets on the interval and fast wavelet transforms*, Appl. Comput. Harmon. Anal. **1**(1) (1993), 54–81.

[4] C. Lacour, *Adaptive estimation of the transition density of a Markov chain*, Ann. Inst. H. Poincaré Probab. Statist. **43**(5) (2007), 571–597.

[5] C. Lacour, *Adaptive estimation of the transition density of a particular hidden Markov chain*, J. Multivariate Anal. (2008), to appear.

Conditional predictive inference post model selection

Hannes Leeb

This talk is about inference on future observations based on a model that has been selected on the basis of the data and then has been fitted to the same data. I focus in particular on situations where the number of candidate models is large, and where the number of explanatory variables in a 'good' model can be large as well, in relation to sample size. Such a situation is faced, for example, by Stenbakken and Souders [3] who predict the performance of analog/digital converters from

partial measurements by selecting 64 explanatory variables (measurements) from a total of 8,192 based on a sample of size 88; further examples include [1, 2, 4]. Note that in these studies, the model that is selected, on the basis of the data, is often quite complex in relation to sample size, in the sense that the number of explanatory variables in the selected model and the sample size are of the same order of magnitude. Also note that the total number of candidate models in these studies exceeds sample size by several orders of magnitude. In such situations, inferential tools that assess the predictors' accuracy like, e.g., the mean-squared error of the predictor, or prediction intervals, are needed.

I consider a Gaussian regression model with random design, where the number of explanatory variables can be infinite, and where no regularity conditions are imposed on the unknown parameters. I use a variant of generalized cross-validation to evaluate the performance of candidate models for prediction out-of-sample,[1] to select a 'good' model, and to conduct predictive inference based on the selected model. The performance of the resulting model selector and the quality of predictive inference procedures are evaluated conditional on the training sample. I describe the performance of these methods by explicit finite-sample performance bounds. For example, I show that the proposed prediction interval is approximately valid and short with high probability, even in statistically challenging situations where the number of explanatory variables in a 'good' model is of the same order as sample size, and where the total number of candidate models is of a larger order than sample size. Here, approximately valid means that the prediction interval's actual coverage probability is close to the nominal level, and approximately short means that its length is close to the length of a certain infeasible 'prediction interval' that is based on actually knowing the 'best' candidate model. These results hold uniformly over all data-generating processes under consideration.

References

[1] B.-L. Adam, Y. Qu, J. W. Davis, M. D. Ward, M. A. Clements, L. H. Cazares, O. J. Semmes, P. F. Schellmanner, Y. Yasui, Z. Feng, and G. L. Jr. Wright, *Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men*, Cancer Research **62** (2002), 3609–3614.

[2] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, D. C. Bloomfield, and E. S. Lander, *Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring*, **286** (1999), 531–537.

[3] G. N. Stenbakken and T. M. Souders, *Test point selection and testability measures via QR factorization of linear models*, IEEE Transactions on Instrumentation and Measurement, **36** (1987), 406–410.

[4] M. J. van de Vijver, Y. D. He, L. J. van't Veer, H. Dai, A. A. M. Hart, D. W. Voskuil, G. J. Schreiber, J. L. Peterse, C. Roberts, M. J. Marton, M. Parrish, D. Atsma,

[1]Here, prediction 'out-of-sample' means prediction of new responses given hitherto unobserved explanatory variables, whereas 'in-sample' prediction means prediction of new responses for the same explanatory variables as observed in the training data.

A. Witteveen, A. Glas, L. Delahaye, T. van der Velde, H. Bartelink, S. Rodenhuis, E. T. Rutgers, S. H. Friend, and R. Bernards, *A gene-expression signature as a predictor of survival in breast cancer*, The New England Journal of Medicine **347** (2002), 1999–2009.

Robust model selection in generalized linear models

Samuel Müller
(joint work with Alan H. Welsh)

Model selection is fundamental to the practical application of statistics and there is a substantial literature on the selection of linear regression models. A growing part of this literature is concerned with robust approaches to selecting linear regression models: see Müller & Welsh (2005) for references. The literature on the selection of generalized linear models and the related marginal models fitted by generalized estimating equations — though both are widely used — is much smaller and has only recently incorporated robustness considerations: see Müller & Welsh (2007) and Cantoni et. al. (2007) for references.

Our perspective on model selection is that a useful model should (i) parsimoniously describe the relationship between the sample data y and X and (ii) be able to predict independent new observations. The ability to parsimoniously describe the relationship between the sample data can be measured by applying a penalised loss function to the observed residuals and we use the expected variance-weighted prediction loss to measure the ability to predict new observations. In addition, we encourage the consideration of different types of estimator of each of the models. We intend to identify useful models whether or not a true model exists and our interest is not restricted to a single best model but to the identification of useful models (which make the selection criterion small). In this context, if a true model exists and it is part of the full model, then consistency in the sense that a procedure identifies the true model with probability tending to one is a desirable property. We show that the concept of a true model is useful in order to establish asymptotic results and to build model selection criteria such that they have the potential to consistently select the true model. We present a generalization of the robust bootstrap model selection criterion of Müller & Welsh (2005) to generalized linear models. Under the 'true model' paradigm we show that the extension of the methodology of Müller & Welsh (2005) from linear regression to generalized linear models is less straightforward than expected but we are still able to improve on the methodology of Mueller & Welsh (2005). For example the stratified bias-adjusted m-out-of-n bootstrap estimator $\widehat{\beta}^{c*}_{\alpha,m} - \mathrm{E}_*(\widehat{\beta}^{c*}_{\alpha,m} - \widehat{\beta}^{c}_{\alpha})$ rather than the stratified m-out-of-n bootstrap estimator $\widehat{\beta}^{c*}_{\alpha,m}$ is used in estimating the expected prediction loss. This achieves the same purpose but avoids the centering of the explanatory variables and the requirement that we include an intercept in every model used in Müller & Welsh (2005). Simulation results show that our procedure can be more efficient than AIC or BIC even for non-robust simulation settings.

References

[1] E. Cantoni, C. Field, J. Mills Flemming, E. Ronchetti *Longitudinal variable selection by cross-validation in the case of many covariates*, Statistics in Medicine **26** (2007), 919–930.

[2] S. Müller, A.H. Welsh *Outlier robust model selection in linear regression*, Journal of the American Statistical Association **100** (2005), 1297–1310.

[3] S. Müller, A.H. Welsh *Robust model selection in generalized linear models*, preprint.

Jumps

Axel Munk

(joint work with Leif Boysen, Volkmar Liebscher, Olaf Wittich)

We study the asymptotics for jump-penalized least squares regression aiming at approximating a regression function by piecewise constant functions. Besides conventional consistency and convergence rates of the estimates in $L^2([0,1])$ our results cover other metrics like Skorokhod metric on the space of càdlàg functions and uniform metrics on $C([0,1])$ as well as convergence of the scale spaces, the family of estimates under varying smoothing parameter. We will show that the estimates used are in an adaptive sense rate optimal over a scale of approximation spaces, including the class of functions of bounded variation, (piecewise) Hölder continuous functions of order $1 \geq \alpha > 0$ and the class of step functions. In the latter setting, we will also deduce the rates known from changepoint analysis for detecting the jumps. Our penalty is an l_0 penalty which typically leads to an optimization problem which cannot be computed in polynomial time. The present situation is different, however, and we discuss a dynamic program which allows to computes the LSE in $O(n^2)$ steps. It turns out, that for data of size $\approx 10^4$ this is sufficent on a PC, for data sets of larger size this is still a severe burden. To overcome this problem we combine this with a forward search algorithm, which allows to handle several millions observations. Our method is illustrated with the reconstruction of ion channel activity measured from impedance tomography. Finally this is combined with a statistical multiscale analysis in order to estimate the number of jumps.

References

[1] L. Boysen, V. Liebscher, A. Munk and O. Wittich, *Scale Space Consistency of Piecewise Constant Least Squares Estimators - Another Look at the Regressogram*, IMS Lecture Notes - Monograph Series **55**: Asymptotics, Particle Processes and Inverse Problems, ed. E. Cator, G. Joengbloed, C. Kraaikamp, R. Lopuhaä, J. Wellner, IMS Beachwood, Ohio (2007), 65-84.

[2] L. Boysen, A. Kempe, V. Liebscher, A. Munk and O. Wittich, *Consistencies and rates of convergence of jump penalized least squares estimators* Ann. Statist. (2006), to appear.

[3] Römer et al., *Channel activity of a viral transmembrane peptide in micro-blms*, J. Amer. Chem. Soc. **49** (2004), 16267-74.

[4] E. Schmitt, M. Vrouernraets and C. Steinem, *Channel activity of ompf monitored in nano blms*, Biophys. J. **91** (2006), 2163-71.

Testing independence in nonparametric regression

Natalie Neumeyer

We consider independent and identically distributed data $(X_1, Y_1), \ldots, (X_n, Y_n)$, where X_i is d-dimensional and Y_i one-dimensional. Under the purpose of modelling the data via a homoscedastic regression model

$$Y_i = m(X_i) + \varepsilon_i$$

with regression function $m(x) = E[Y_i \mid X_i = x]$, where the error $\varepsilon_i = Y_i - E[Y_i \mid X_i]$ is independent of the covariate X_i, we propose a new test for the hypothesis

$$H_0 : X_i \text{ and } \varepsilon_i \text{ are independent.}$$

We suggest a simple kernel based test statistic, i. e.

$$T_n = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} \frac{1}{h_n} K\Big(\frac{\widehat{\varepsilon}_i - \widehat{\varepsilon}_j}{h_n}\Big) \times \int (I\{X_i \leq x\} - F_{X,n}(x))(I\{X_j \leq x\} - F_{X,n}(x)) w(x)\, dx,$$

where h_n is a sequence of positive bandwidths, K a kernel function, $I\{\cdot\}$ the indicator function, w a weight function, and $F_{X,n}$ denotes the empirical distribution function of the covariates $X_1, \ldots, X_n$. Note that the errors ε_i are not observable and, hence, are nonparametrically estimated by residuals $\widehat{\varepsilon}_i = Y_i - \widehat{m}(X_i)$, where $\widehat{m}$ denotes, for instance, a local polynomial estimator for the regression function.

T_n estimates an L_2-distance of the conditional (given the errors) and unconditional distribution of the covariates, i. e.

$$\int \int \Big(P(X_1 \leq x \mid \varepsilon_1 = y) - F_X(x)\Big)^2 f_\varepsilon^2(y) w(x)\, dy\, dx,$$

where F_X denotes the covariate distribution function and f_ε the error density.

Were errors observable and residuals replaced by true errors the test statistic T_n would coincide with the test for independence proposed by Zheng [5], which was further investigated by Dette and Neumeyer [1]. It turns out that the replacement of errors by residuals has no influence on the asymptotic distribution under the null hypothesis, but it has under the alternative. Under typical regularity assumptions the test statistic (suitably standardized) has an asymptotic normal law both under the null hypothesis and under fixed alternatives. Because the asymptotic null distribution depends on unknown features of the data-generating process, we recommend to apply resampling procedures. It can be shown in asymptotic theory and it is supported by simulations that the classical residual bootstrap is applicable.

Please note that the test statistic is asymmetric and that interchanging the roles of covariates X_i and residuals $\widehat{\varepsilon}_i$ might on first glance seem to be a more canonical way to consider the problem (considering the L_2-distance between the conditional error distribution given the covariate and the unconditional error distribution). However, interchanging those roles has a substantial influence on the asymptotic

behaviour of the test statistic. It causes unwanted bias and asymptotic theory is so far only available for one-dimensional covariates, where it results in a normal distribution with a rather complicated variance.

The proposed test statistic T_n can be adjusted to justify a regression model with heteroscedastic variance, i. e. $Y_i = m(X_i) + \sigma(X_i)\varepsilon_i$, where the covariates X_i are independent of the errors $\varepsilon_i = (Y_i - E[Y_i \mid X_i])/(var(Y_i \mid X_i))^{1/2}$.

Although the independence of error and covariate is a common assumption, to the present author's knowledge so far there are only two tests available in literature. In the homoscedastic model Einmahl and Van Keilegom [2] consider a very innovative procedure based on a stochastic process of differences of the observations Y_i, which converges weakly to a bivariate Gaussian process. The test avoids estimating the regression function. In the heteroscedastic model Einmahl and Van Keilegom [3] propose tests based on the difference of the empirical distribution function of $(X_i, \widehat{\varepsilon}_i)$ and the product of the marginal empirical distribution functions. The considered process converges weakly to a bivariate Gaussian process. Both procedures are presented for one-dimensional covariates only and cannot easily be extended to the important multivariate case. In contrast the new procedure is valid for multivariate covariates.

References

[1] H. Dette, N. Neumeyer, *A note on a specification test of independence*, Metrika **51** (2000), 133–144.

[2] J. Einmahl, I. Van Keilegom, *Tests for independence in nonparametric regression*, Statist. Sinica (2007a), to appear. `http://www.stat.ucl.ac.be/ISpersonnel/vankeile/pub.html`

[3] J. Einmahl, I. Van Keilegom, *Specification tests in nonparametric regression*, J. Econometrics (2007b), to appear. `http://www.stat.ucl.ac.be/ISpersonnel/vankeile/pub.html`

[4] N. Neumeyer, *Testing independence in nonparametric regression*, preprint (2007). `http://www.math.uni-hamburg.de/research/ims.html`

[5] J.X. Zheng, *A consistent specification test of independence*, J. Nonparametr. Statist. **7** (1997), 297–306.

On the distribution of penalized maximum likelihood estimators

Benedikt M. Pötscher

(joint work with Hannes Leeb, Ulrike Schneider)

Penalized maximum likelihood estimators have been studied intensively in the last few years. A prominent example is the least absolute selection and shrinkage (LASSO) estimator of Tibshirani (1996). Related variants of the LASSO include the Bridge estimators studied by Frank and Friedman (1993), least angle regression (LARS) of Efron, Hastie, Johnston, Tibshirani (2004), the smoothly clipped absolute deviation (SCAD) estimator of Fan and Li (2001), or the adaptive LASSO of Zou (2006). Other estimators that fit into this framework are hard- and soft-thresholding estimators. While many properties of penalized maximum likelihood estimators are now well understood, the understanding of their distributional properties, such as finite-sample and large-sample limit distributions,

is still incomplete. The probably most important contribution in this respect is Knight and Fu (2000) who study the asymptotic distribution of the LASSO estimator (and of Bridge estimators more generally) when the tuning parameter governing the influence of the penalty term is chosen so that the LASSO acts as a conservative model selection procedure (that is, a procedure that does not select underparameterized models asymptotically, but selects overparameterized models with positive probability asymptotically). In that paper, the asymptotic distribution is obtained in a fixed-parameter as well as in a standard local alternatives setup. Knight and Fu (2000) is complemented by a result in Zou (2006) who considers the fixed-parameter asymptotic distribution of the LASSO when tuned to act as a consistent model selection procedure. Zou (2006) also studies the fixed-parameter asymptotic distribution of the adaptive LASSO. Another contribution is Fan and Li (2001) who derive the asymptotic distribution of the SCAD estimator when the tuning parameter is chosen so that the SCAD estimator performs consistent model selection. The results in the latter paper are also fixed-parameter asymptotic results. It is well-known that fixed-parameter (i.e., pointwise) asymptotic results can give a wrong picture of the estimators' actual behavior, especially when the estimator performs model selection; see, e.g., Kabaila (1995), or Leeb and Pötscher (2005, 2007). Therefore, it is interesting to take a closer look at the actual distributional properties of such estimators.

In this talk, which is based on Pötscher and Leeb (2007) and Pötscher and Schneider (2007), we consider the finite-sample as well as the asymptotic distributions of the hard-thresholding, the LASSO (which coincides with soft-thresholding in our context), the adaptive LASSO, and of the SCAD estimator in a simple Gaussian model. We study both the cases where the estimators are tuned to perform conservative model selection as well as where the tuning is such that the estimators perform consistent model selection. We find that the finite-sample distributions can be decisively non-normal (e.g., multimodal). Moreover, we find that a fixed-parameter asymptotic analysis gives highly misleading results. In particular, the so-called "oracle property" which has been established for some of the estimators considered above is shown to be highly misleading. Therefore, we also discuss the asymptotic distributions of the estimators mentioned before in a general 'moving parameter' asymptotic framework, which better captures essential features of the finite-sample distribution and shows that the large-sample limits retain the non-normality present in finite samples.

We also show that the finite-sample distribution of the estimators considered can not be estimated in any reasonable sense, complementing results of this sort in the literature (Leeb and Pötscher (2006a,b, 2008), Pötscher (2006)).

We note that penalized maximum likelihood estimators are intimately related to more classical post-model-selection estimators. The distributional properties of the latter estimators have been studied by Sen (1979), Pötscher (1991), and Leeb and Pötscher (2003, 2005, 2006a,b, 2008).

References

[1] B. Efron, T. Hastie, I. Johnstone and R. Tibshirani,*Least angle regression*, Annals of Statistics **32** (2004), 407-499.

[2] J. Fan and R. Li, *Variable selection via nonconcave penalized likelihood and its oracle properties*, Journal of the American Statistical Association **96** (2001), 1348-1360.

[3] I. E. Frank and J. H. Friedman, *A statistical view of some chemometrics regression tools (with discussion)*, Technometrics **35** (1993), 109-148.

[4] P. Kabaila, *The effect of model selection on confidence regions and prediction regions*, Econometric Theory **11** (1995), 537–549.

[5] K. Knight and W. Fu, *Asymptotics for lasso-type estimators*, Annals of Statistics **28** (2000), 1356-1378.

[6] H. Leeb and B. M. Pötscher, *The finite-sample distribution of post-model-selection estimators and uniform versus nonuniform approximations*, Econometric Theory **19** (2003), 100–142.

[7] H. Leeb and B. M. Pötscher, *Model selection and inference: Facts and fiction,* Econometric Theory **21** (2005), 21–59.

[8] H. Leeb and B. M. Pötscher *Performance limits for estimators of the risk or distribution of shrinkage-type estimators, and some general lower risk-bound results*, Econometric Theory **22** (2006a), 21–59. (Corrections: ibidem, forthcoming).

[9] H. Leeb and B. M. Pötscher *Can one estimate the conditional distribution of post-model-selection estimators?*, Annals of Statistics **34** (2006b), 2554-2591.

[10] H. Leeb and B. M. Pötscher *Sparse estimators and the oracle property, or the return of Hodges' estimator*, Journal of Econometrics (2007), doi:10.1016/j.jeconom.2007.05.017.

[11] H. Leeb and B. M. Pötscher *Can one estimate the unconditional distribution of post-model-selection estimators?*, Econometric Theory **24** (2008), forthcoming.

[12] B. M. Pötscher, *Effects of model selection on inference*, Econometric Theory **7** (1991), 163–185.

[13] B. M. Pötscher, *The distribution of model averaging estimators and an impossibility result regarding its estimation*, IMS Lecture Notes–Monograph Series **52** (2006), 113-129.

[14] B. M. Pötscher and H. Leeb *On the distribution of penalized maximum likelihood estimators: The LASSO, SCAD, and Thresholding*, Manuscript (2007).

[15] B. M. Pötscher and U. Schneider, *On the distribution of the adaptive LASSO estimator*, Manuscript (2007).

[16] P. K. Sen, *Asymptotic properties of maximum likelihood estimators based on conditional specification*, Annals of Statistics **7** (1979), 1019-1033.

[17] R. Tibshirani, *Regression shrinkage and selection via the lasso*, Journal of the Royal Statistical Society Series B **58** (1996), 267-288.

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

Parameter cascading for high dimensional models

James Ramsay

(joint work with David Campbell, Jiguo Cao, Giles Hooker)

High dimensional models often involve three classes of parameters. Nuisance parameters c are required to fit the data, are large in number, their number tends to depend on how much data is available, often define localized effects on the fit, and their values are seldom of direct interest. Structural parameters θ are the conventional kind; a small fixed number and their values are of interpretive importance.

Above these are the complexity parameters γ that define the overall complexity of the solution.

This talk defines a general framework for parameter estimation that synthesizes a variety of common approaches and brings some important new advantages. The *parameter cascade* defines nuisance parameters as functions $c(\theta, \gamma)$ of structural and complexity parameters, and in turn defines structural parameters as functions $\theta(\gamma)$ of complexity parameters. These functional relationships are often defined by choosing three different optimization criteria corresponding to each level.

It is common to define the lowest level or inner criterion $L(c|\theta, \gamma)$ as a regularized loss function with the penalty controlled by γ, as in

$$J(c|\theta,\gamma) = \sum_i^N [y_i - \beta' z_i - c'\phi(t_i)]^2 + e^{-\gamma} c' [\int \| \frac{d^2\phi}{dt^2} - \alpha_0 \phi(t) - \alpha_1 \frac{d\phi}{dt} \|^2 dt] c$$

where $x(t) = c'\phi(t)$, z_i is a p-vector of covariate values, and $\phi(t)$ is a vector of K basis functions. There are three groups of parameters to estimate:

- The K coefficients in c defining the basis function expansion of $x(t)$.
- The $p + 2$ model parameters α and β defining the data fitting model and the roughness penalty, respectively. For simplicity, we use θ to collect these two vectors together; $\theta = (\alpha', \beta')'$.
- The single smoothing parameter γ.

The regularization assures that $c(\theta, \gamma)$ is smooth in a specified sense, and effectively controls the degrees of freedom allocated to the nuisance parameters. But $c(\theta, \gamma)$ may also be defined explicitly, or by an algorithm whose result depends on θ and γ, as in kernel smoothing. This functional relationship between nuisance and other parameters is a generalization of the familiar *profiling* procedure often used in nonlinear regression, where the three optimization criteria are the same.

The middle level optimization is usually an unregularized measure of fit, such as

$$H(\theta|\gamma) = \sum_i^N [y_i - \beta' z_i - c(\theta, \gamma)'\phi(t_i)]^2,$$

and the fact that the status of c as a parameter as been eliminated by replacing it by a function of the other two classes implicitly ensures regularization. Of course we need the derivative of $c(\theta, \gamma)$, and this is, by the *Implicit Function Theorem*,

$$\frac{dc}{d\theta} = -(\frac{\partial^2 F}{\partial \theta^2})^{-1} (\frac{\partial^2 F}{\partial \theta \partial c}).$$

Finally, the top level optimization is a measure of model complexity such as the generalized cross-validation measure of predictive complexity

$$G(\gamma) \sim \frac{\|[I - A(\gamma)]y\|^2}{\|[I - A(\gamma)\|^2},$$

where $A(\gamma)$ is the smoothing operator, is effectively a *Raleigh coefficient* showing the size of the residual vector $[I-A(\gamma)]y$ relative to the size of the *residual operator* $I - A(\gamma)$. The Implicit Function Theorem again gives us $\frac{d\theta}{d\gamma}$.

Estimation of confidence intervals and other inferential methods can proceed at this point by classical methods such as the delta method. The application will typically require further use of the Implicit Function Theorem to compute the required derivatives.

This general framework can be seen to include a number of specific parameter estimation strategies in common use, such as the process of removing nuisance parameters by marginalizing a likelihood. Since the marginal likelihood

$$L^*(\theta|y) = \int L(\theta, c|y)p(c)dc$$

is a linear operation, it is necessarily the optimum of a functional quadratic optimization problem, and in fact minimizes

$$J(c|\theta, y) = \int [L(\theta, c|y) - L^*(\theta|y)]^2 e^{\ln p(c)+C} dc$$

for any constant C. We see here a *functional regression problem* in which function $L(\theta, c|y)$ is approximated by a marginal function $L^*(\theta|y)$ conditional on specific values of structural parameter θ and data y. What is missing in marginalization, however, is any counterpart of smoothing parameter γ that permits a continuum of regularization. But it seems perfectly feasible to remove this difficulty by appending a continuously controlled penalty to this definition of $J(c|\theta, y)$.

The parameter cascade procedure brings important advantages to parameter estimation in the presence of nuisance parameters.

- Gradients and hessians at any level can be analytically computed using the Implicit Function Theorem.
- Interval estimation methods are readily at hand.
- Compared to marginalizing out the nuisance parameters employed in Bayesian approaches using MCMC, generalized profiling is
  - much faster,
  - much more stable,
  - much easier to program,
  - permits an adaptive control of the contribution of c to the fit,
  - requires no "tuning" by an MCMC expert, and
  - can be deployed to the user community much more conveniently.

References

[1] J. Cao, J. O. Ramsay, *Parameter cascades and profiling in functional data analysis*, Computational Statistics, **22** (2007), 335-351.

[2] J. O. Ramsay, G. Hooker, J. Cao. and D. Campbell, *Parameter estimation for differential equations: A generalized smoothing approach (with discussion)*, Journal of the Royal Statistical Society, Series B **69** (2007), 741-796.

Sequential normalization and optimally distinguishable models

Jorma Rissanen

This talk is about two distinct topics: The first describes a universal model for regression problems and time series which is strictly better than the plug-in prequential model or the predictive MDL model. The second introduces a sense of optimality to hypothesis testing by reducing the uncountable set of composite hypotheses to, in effect, a few 'optimally distinguishable' ones.

Consider a class of parametric models

$$\mathcal{M}_k = \{f(\cdot;\theta) : \theta \in \Omega \subseteq R^k\},$$

where $\cdot$ represents a sequence of data points $y^n = y_1, \ldots, y_n$ or $(y^n|x^n) = \{(y_i|x_i)\}$ of any type, and $\theta = \theta_1, \ldots, \theta_k$ denotes the parameters. The model class could also be taken as the union $\mathcal{M} = \bigcup_k \mathcal{M}_k$, to handle 'nonparametric' model classes like histograms.

Consider the following nonpredictive representation of data

$$y_t = \widehat{a}_1(y^t)y_{t-1} + \ldots + \widehat{a}_k(y^t)y_{t-k} + \widehat{e}_t$$

where the coefficients are the least squares estimates, their number initially such that they can be uniquely calculated. This is an example of a more general regression problem. Define $\widehat{s}_t = \sum_{i \le t} \widehat{e}_i^2$. Because the parameter estimates depend on y_t, $\widehat{e}_t$ is not a prediction error, and the density function induced is not gaussian but

$$\widehat{f}(y^n) = p(y_1) \prod_t K_{t-1}^{-1} \frac{\widehat{s}_t^{-t/2}}{\widehat{s}_{t-1}^{-(t-1)/2}},$$

where $p(y_1)$ is a suitably selected initial density function, and

$$\begin{aligned} K_{t-1} &= \frac{\sqrt{\pi}}{1-d_t}\Gamma\left(\frac{t-1}{2}\right)/\Gamma(t/2) \\ d_t &= \overline{x}_t' V_t \overline{x}_t \\ \overline{x}_t &= col\{y_{t-1}, \ldots, y_{t-k}\}. \end{aligned}$$

Further, $V_t = (X_t X_t')^{-1}$, where X_t is the regressor matrix defined by the columns $\overline{x}_i$.

One can show that the density function is *universal*, capable of imitating or estimating any normal one $f(y^n; \overline{a}, \sigma^2)$ induced by a k'th order AR model with gaussian iid input ϵ_t, in the sense that $n^{-1}\ln(\widehat{f}(y^n)/f(y^n;\overline{a},\sigma^2)) \to 0$, either in the mean or almost surely or both. In the mean case the estimation error is measured by the Kullback-Leibler distance. Moreover, the universal model is *optimal* since the convergence rate is the fastest possible, or the distance measure smallest possible.

As to the second part of the talk, a compact parameter space Ω can be partitioned into equivalence classes, defined by the largest curvilinear rectangles $B_{d/n}(\theta^i)$ within the ellipsoids $\delta_t' J(\theta^i)\delta_t = d/n$, where $\delta_t = \theta - \theta^i$ and $J(\theta)$ is the Fisher information matrix. The centers define a finite number of models $f(y^n;\theta^i)$.

There are two desired properties of a well separated family: The density functions $f(y^n;\theta)$ for $\theta \in B_{d/n}(\theta^i)$ in each equivalence class should be close to its representative $f(y^n;\theta^i)$ so that they could be collapsed to it, and each representative should assign a large probability mass to its equivalence class to make the adjacent models different. If the CLT holds, these are conflicting properties for the family constructed above, but ideally satisfied by the family

$$\widehat{f}(y^n|\theta^i) = \begin{cases} f(y^n;\widehat{\theta}(y^n))/Q_{d/n}(\theta^i) & \text{if } \widehat{\theta}(y^n) \in B_{d/n}(\theta^i) \\ 0 & \text{otherwise} \end{cases}$$

where

$$Q_{d/n}(\theta^i) = \int_{\widehat{\theta} \in B_{d/n}(\theta^i)} g(\widehat{\theta};\widehat{\theta}) d\widehat{\theta}$$

and $g(\widehat{\theta};\theta)$ is the density function on the ML estimates.

We may ask for the value of the parameter d for which the desired real models are as close as possible to the perfectly distinguishable models in terms of the KL distance; i.e.

$$\min_d D(\widehat{f}(Y^n|\theta^i)\|f(Y^n;\theta^i)) = \min_d \int \widehat{f}(y^n|\theta^i) \log \frac{\widehat{f}(y^n|\theta^i)}{f(y^n;\theta^i)} dy^n.$$

If the family of models $\mathcal{M}_k$ satisfies the CLT the optimal value $\widehat{d}_n \to 3k$ as n grows for all $\theta = \theta_i$.

The same value defines the quantization of the parameters providing a minimal sufficient statistics decomposition of the models in the sense of Kolmogorov; see [1], and optimal separation of noise from the learnable information in the data.

Testing a null hypothesis, say $f(y^n;\theta^0)$, against a composite hypothesis amounts simply to accept it iff the ML estimate falls within its equivalence class. The error probabilities, which do not require the untenable assumption of 'true' hypotheses, of such a decision is easy to evaluate.

References

[1] J. Rissanen, *Information and Complexity in Statistical Modeling*, Springer Verlag (2007), 142 pages.

Some issues on variable selection with applications to longitudinal data

Elvezio Ronchetti

Variable selection is an important issue in statistical modelling. In this talk we address three aspects related to the performance of standard model selection criteria when (1) the signal-to-noise ratio is low, when (2) the number of variables p is (much) larger than the number of observations n, and when (3) there are deviations from the stochastic assumptions of the model (robustness issue).

1. Low signal-to-noise ratio

This situation is common across economics and social sciences. We illustrate the behavior of standard model selection criteria (such as AIC, BIC, etc.) in predictability studies in finance, in the case of regression models based on financial and macroeconomic factors for the prediction of stock and bonds returns. We reproduce in a simulation setting two benchmark studies, namely those by Pesaran and Timmermann(1995) and Bossaerts and Hillion(1999) and we find that the limited out-of-sample forecasting power is mainly due to the low discrimination power of standard model selection criteria. In particular a very large class of indistinguishable models (from the model selected by a given criterion) appears. Details can be found in Dell'Aquila and Ronchetti (2006).

2. $p >> n$

This situation becomes more and more important in a variety of applications, including microarray data (p genes, n patients), financial data (p stocks, n observations in time), and data mining. For a regression model, Donoho and Stodden (2006) give the relative error of LASSO in the estimation of the regression coefficients in the (δ, ρ) plane, where $\delta = n/p$ (degree of underdetermination) and $\rho = k/n$ (degree of sparsity) with k the number of non-zero coefficients in the regression model. They show that there is a transition phase in the (δ, ρ) plane and determine the breakdown point of LASSO.

3. The robustness issue

In the presence of small deviations from the assumed model, standard variable selection criteria break down and fail to capture the important variables. In this context it is important to develop variable selection procedures which are insensitive to such small deviations and pick models which fit well the *majority* of the data. We illustrate this point in the framework of longitudinal models which are commonly used for studying data collected on individuals repeatedly through time. While there are now a variety of such models available (Marginal Models, Mixed Effects Models, etc.), the important issue of variable selection has been somewhat neglected in this context. We discuss some recent proposals based on a generalized version of Mallows' C_p suitable for use with both parametric and nonparametric models. We examine their performance and their robustness properties with popular marginal longitudinal models (fitted using GEE) and contrast results with what is typically done in practice: variable selection based on Wald-type or score-type tests. Details can be found in Ronchetti and Staudte (1994) and Cantoni, Flemming, Ronchetti (2005).

References

[1] P. Bossaerts, P. Hillion, *Implementing statistical criteria to select return forecasting models: what do we learn?*, Review of Financial Studies **12** (1999), 405-428.

[2] E. Cantoni, J. Flemming, E. Ronchetti, *Variable selection for marginal longitudinal generalized linear models*, Biometrics **61** (2005), 507-514.

[3] R. Dell'Aquila, E. Ronchetti, *Stock and bond return predictability: the discrimination power of model selection criteria*, Computational Statistics & Data Analysis **50** (2006), 1478-1495.
[4] D. Donoho, V. Stodden, *Breakdown point of model selection when the number of variables exceeds the number of observations*, Proceedings of the International Joint Conference on Neural Networks (2006), forthcoming.
[5] M.H. Pesaran, A. Timmermann, *Predictability of stock returns: robustness and economic significance*, Journal of Finance **50** (1995), 1201-1228.
[6] E. Ronchetti, R. Staudte, *A robust version of Mallows' C_p*, Journal of the American Statistical Association **89** (1994), 550-559.

Practices of model building

Ritei Shibata

Model building plays a central role in the stream of data heading in the collection toward decision making but it is not straightforward. Creating a model from data is really a tough job, a kind of art, since phenomena behind the data and the generating process have to be both modelled in a balance. It is clear that such a goal can not be easily achieved by a simple data smoothing or by averaging several possible models. Although a rough idea can be obtained by smoothing or averaging, but a model really appreciated by scientists or people in industries can be obtained only by continuous efforts, generous idea and constant interaction with them. Also, it is worthy of note that application of a formal model selection procedure like AIC is only powerful when the underlying models are all polished well. I have reported several practices of model building, including stochastic neural network modelling and clustered marked point process modelling. I hope that accumulation of such practices will open the door to the new paradigm of model building in the frame work of data science.

References

[1] S. Kamitsuji and R. Shibata, *Effectiveness of Stochastic Neural Network for Prediction of Fall or Rise of TOPIX*, Asia-Pacific Financial Markets **10** (2003), 187–204.
[2] R. Shibata, *Modelling FX new bid prices as a clustered marked point process*, Proceedings in Computational Statistics 2006, Ed. Alfredo Rizzi and Maurizio Vichi, 1565–1572, 2006, Physica-Verlag, Heidelberg.

Adaptive estimation in a linear inverse problem

Vladimir Spokoiny
(joint work with Céline Vial)

Consider a general set-up of a linear inverse problem when the observed data Y from a Hilbert space $\mathcal{H}_Y$ are modelled by a linear operator equation

$$Y = AX + \varepsilon \tag{1}$$

where X is the unknown parameter vector from some Hilbert space $\mathcal{H}_X$, $A : \mathcal{H}_X \to \mathcal{H}_Y$ is a linear operator, and ε is a random Gaussian noise in $\mathcal{H}_Y$ with the known correlation structure given by the covariance operator Σ. The goal is to estimate

a linear functional $\theta = \theta(X)$ which can be represented in the form $\langle \vartheta, X \rangle$ for some known element $\vartheta \in \mathcal{H}_X$.

A naive estimation approach is based on the explicit least square solution of the problem (1):

$$\widetilde{\theta} = \langle \vartheta, (A^*A)^- A^*Y \rangle = \langle A(A^*A)^- \vartheta, Y \rangle = \langle \phi, Y \rangle$$

where A^* is the conjugate operator to A, C^- means a pseudo-inverse of C and $\phi = A(A^*A)^- \vartheta$. However, this approach cannot be efficiently applied if A is a compact operator because the inverse of A^*A does not exists or is an unbounded operator. One can regularize the problem if some additional information about smoothness of the element X is available. This allows to replace $(A^*A)^-$ by its regularization $g_\alpha(A^*A)$ where g_α means some regularized inversion and α is the corresponding parameters. See, e.g., Goldenshluger and Pereversev (1999) for typical examples. The quality of estimation heavily depends on the choice of the regularization parameter α and its choice is a challenging problem. Usually one fixes a finite ordered set of values $\alpha_1 < \alpha_2 < \ldots < \alpha_K$ and considers the corresponding estimates

$$\widetilde{\theta}_k = \langle \phi_k, Y \rangle, \qquad \phi_k = A g_{\alpha_k}(A^*A)\vartheta.$$

Now the original problem can be reformulated as follows: given a set of estimates $\widetilde{\theta}_k$ for known vectors ϕ_k, build an estimate $\widehat{\theta}$ of the functional θ which performs nearly as good as the best in this family.

For a given sequence of estimates $\widetilde{\theta}_k = \langle \phi_k, X \rangle$ consider the sequence of nested hypothesis $H_k : \theta_1 = \ldots = \theta_k = \theta$. The proposed selection procedure is sequential: we start with $k = 2$ and at every step k the hypothesis H_k is tested against $H_1, \ldots, H_{k-1}$. If H_k is not rejected then we continue with the next larger k. The final estimate corresponds to the latest accepted hypothesis. For testing H_k against H_l with $l < k$, we check that the new estimate $\widetilde{\theta}_k$ belongs to the confidence intervals built on the base of $\widetilde{\theta}_l$. More precisely, we apply the test statistics:

$$T_{lk} = \bigl(\widetilde{\theta}_l - \widetilde{\theta}_k\bigr)^2 / v_l \,, \qquad l < k,$$

where v_l is the variance of $\widetilde{\theta}_l$. Big values of T_{lk} indicate a significant difference between the estimates $\widetilde{\theta}_l$ and $\widetilde{\theta}_k$. The estimate $\widetilde{\theta}_k$ (or the hypothesis H_k) is accepted if H_{k-1} was accepted and $T_{lk} \le \mathfrak{z}_l$ for all $l < k$, that is, the new estimate $\widetilde{\theta}_k$ belongs to the intersection of all the confidence intervals $\mathcal{E}_l(\mathfrak{z}_l)$ built on the previous steps of the procedure. The formal definition is given by

$$\widehat{k} = \max\{k \le K : \quad T^*_{lk} \le \mathfrak{z}_l \quad l = 1, \ldots, k-1\}, \qquad T^*_{lk} = \max_{l<j\le k} T_{lj} \,.$$

Here the "critical values" $\mathfrak{z}_1, \ldots, \mathfrak{z}_{K-1}$ are the parameters of the procedure which are selected by the reasoning similar to the standard approach of hypothesis testing theory: to provide the prescribed performance of the procedure under the simplest (null) hypothesis. In the considered set-up, the null means $X \equiv 0$. In this case it is natural to expect that the estimate $\widehat{\theta}_k$ coming out of the first steps of the procedure until the index k is close to the nonadaptive counterpart $\widetilde{\theta}_k$. This

particularly means that the probability of rejecting one of the estimates $\widetilde{\theta}_2, \ldots, \widetilde{\theta}_k$ under the null hypothesis should be very small.

Suppose that the risk of estimation for an estimate $\widehat{\theta}$ of θ is measured by $\boldsymbol{E}\big|\widehat{\theta} - \theta\big|^{2r}$ for some $r > 0$. Under the null hypothesis $X \equiv 0$, every estimate $\widetilde{\theta}_k$ fulfills $\widetilde{\theta}_k = \langle \phi_k, \varepsilon \rangle$ and hence, it is a zero mean normal variable with the variance v_k. Therefore,

$$\boldsymbol{E}_0 \big| v_k^{-1} \big(\widetilde{\theta}_k - \theta\big)^2 \big|^r = \mathfrak{c}_r$$

where $\mathfrak{c}_r = E|\xi|^{2r}$ and ξ is standard normal. We require that the parameters $\mathfrak{z}_1, \ldots, \mathfrak{z}_{K-1}$ of the procedure are selected in such a way that

$$\boldsymbol{E}_0 \big| v_k^{-1} \big(\widehat{\theta}_k - \widetilde{\theta}_k\big)^2 \big|^r \le \alpha \mathfrak{c}_r \,, \qquad k = 2, \ldots, K. \tag{2}$$

Here α is the preselected constant which is similar to the confidence level of a testing procedure. This gives us $K-1$ conditions to fix $K-1$ parameters. As in the testing problem, we are interested to select the critical values as small as possible under the constraint (2).

The theoretical results about the quality of the adaptive estimate $\widehat{\theta}$ are established under the following condition on the variances v_k.

(*MD*): for some constants $\mathfrak{u}_0, \mathfrak{u}$ with $1 < \mathfrak{u}_0 \le \mathfrak{u}$, the variances v_k satisfy

$$v_{k-1} \le \mathfrak{u} v_k, \qquad \mathfrak{u}_0 v_k \le v_{k-1}, \qquad 2 \le k \le K.$$

Theorem. Assume (*MD*). Let $\theta_k = \theta$ for all $k \ge 1$. Then there are three constants a_0, a_1 and a_2 depending on r and $\mathfrak{u}_0$, $\mathfrak{u}$ only such that the choice

$$\mathfrak{z}_k = a_0 + a_1 \log \alpha^{-1} + a_2 r \log(v_k / v_K)$$

ensures (2) for all $k \le K$. Particularly, $\boldsymbol{E}_0 \big| v_K^{-1} \big(\widetilde{\theta}_K - \widehat{\theta}\big)^2 \big|^r \le \alpha \mathfrak{c}_r$.

Let B_k means the covariance matrix of the vector $\widetilde{\theta}(k)$. For $k \ge 1$, define also $b(k) = (b_1, \ldots, b_k)^\top$ with $b_k = \theta_k - \theta$ and

$$\Delta_k \stackrel{\mathrm{def}}{=} b^\top(k) B_k^{-1} b(k).$$

This quantity measures the "modeling bias" and allows to define the "oracle" choice k^* as the maximal index for which $\Delta_k \le \Delta$. Now we present the following "oracle" inequality which claims that the adaptive estimate $\widehat{\theta}$ achieves essentially the same accuracy as the "oracle" $\widetilde{\theta}_{k^*}$.

Theorem. Let k^* be the maximal value k such that $\Delta_k \le \Delta$. Then

$$\boldsymbol{E} \big| v_{k^*}^{-1} \big(\widetilde{\theta}_{k^*} - \widehat{\theta}\big)^2 \big|^{r/2} \le \sqrt{\alpha \mathfrak{c}_r e^{\Delta}} + \mathfrak{z}_{k^*}^{r/2} \,.$$

References

[1] T.T. Cai and P. Hall, *Prediction in functional linear regression*, Ann. Statist. **34**(5) (2006), 2159–2179.

[2] L. Cavalier, *On the problem of local adaptive estimation in tomography*, Bernoulli **7** (2001), 63–78.

[3] Cavalier L., Golubev, G.K., *Risk hull method and regularization by projections of ill-posed inverse problems*, Annals of Statistics **34**(4) (2006), 1653–1677.

[4] L. Cavalier , G.K. Golubev, D. Picard and A. Tsybakov, *Oracle inequalities for inverse problems*, Annals of Statistics **30**(3) (2002), 843–874.
[5] L. Cavalier and N.W. Hengartner, *Adaptive estimation for inverse problems with noisy operators,*Inverse Problems **21** (2005), 1345–1361.
[6] L. Cavalier and A. Tsybakov, *Penalized blockwise Stein's method, monotone oracles and sharp adaptive estimation,*Mathematical Methods of Statistics **10** (2001), 247-282.
[7] L. Cavalier and A. Tsybakov, *Sharp adaptation for inverse problems with random noise*, Probability Theory and Related Fields **123** (2002), 323-354.
[8] A. Goldenshluger, *On pointwise adaptive nonparametric deconvolution*, Bernoulli **5** (1999), 907–925.
[9] A. Goldenshluger and S. Pereverzev, *On adaptive inverse estimation of linear functionals in Hilbert scales*, Bernoulli **9**(5) (2003), 783–807.
[10] A. Goldenshluger and S. Pereverzev *Adaptive estimation of linear functionals in Hilbert scales from indirect white noise observations*, Probab. Theory and Related Fields **118** (2000), 169-186.
[11] G.K. Golubev, *The Method of Risk Envelope in Estimation of Linear Functionals*, Problems of Information Transmission **40** (1) (2004), 53–65. Translated from Problemy Peredachi Informatsii **1**(2004), 58–72.
[12] O.V. Lepskii, *A problem of adaptive estimation in Gaussian white noise*, Theory Probab. Appl. **35**(3) (1990), 454–466. Translated from Teor. Veroyatnost. i Primenen. **35** (3)(1990), 459–470.
[13] P. Mathé and S. V. Pereverzev, *Optimal discretization of inverse problems in Hilbert scales. Regularization and self-regularization of projection methods*, SIAM J. Numer. Anal. **38**(6) (2001), 1999–2021.
[14] A.B. Tsybakov, *Adaptive estimation for inverse problems: a logarithmic effect in L_2*, C.R. Acad. Sci. Paris, ser. 1 **330** (2000), 835–840.

Sparsity oracle inequalities

Alexander B. Tsybakov

This talk gives an overview of recent results on sparsity oracle inequalities (SOI), mainly based on [1, 2, 3, 5]. Consider the regression model: assume that we observe the pairs $(X_1, Y_1), \dots, (X_n, Y_n) \in \mathbb{R}^d \times \mathbb{R}$ where

$$Y_i = f(X_i) + \xi_i, \quad i = 1, \dots, n.$$

Here the regression function $f : \mathbb{R}^d \to \mathbb{R}$ is unknown, the errors ξ_i are independent Gaussian $\mathcal{N}(0, \sigma^2)$ random variables and $X_i \in \mathbb{R}^d$ are arbitrary fixed design points. We study estimation of f based on the data $(X_1, Y_1), \dots, (X_n, Y_n)$.

Let $\{f_1, \dots, f_M\}$ be a given dictionary of functions, $f_j : \mathbb{R}^d \to \mathbb{R}$. We approximate the regression function f by a linear combination $\mathsf{f}_\lambda(x) = \sum_{j=1}^M \lambda_j f_j(x)$ with weights $\lambda = (\lambda_1, \dots, \lambda_M)$, where possibly $M \gg n$. The number of non-zero coordinates of a vector $\lambda \in \mathbb{R}^M$ denoted by $M(\lambda) = \sum_{j=1}^M \mathbb{I}_{\{\lambda_j \neq 0\}}$ characterizes the *sparsity* of λ: the smaller $M(\lambda)$, the "sparser" λ.

Consider the norm $\|f\| = \sqrt{\frac{1}{n}\sum_{i=1}^n f^2(X_i)}$. Denote by $\langle \cdot, \cdot \rangle$ the corresponding scalar product, introduce the Gram matrix associated to the dictionary: $\Psi = (\langle f_j, f_{j'} \rangle)_{1 \le j, j' \le M}$, and denote by $\mathrm{tr}(\Psi)$ the trace of Ψ.

Our target is to construct an estimator $\widetilde{f}_n$ satisfying *sparsity oracle inequality* (SOI), i.e., an inequality of the form

$$\mathbb{E}\|\widetilde{f}_n - f\|^2 \le \inf_{\lambda\in\mathbb{R}^M} \left\{ \|f - \mathsf{f}_\lambda\|^2 + \frac{M(\lambda) l_{n,M}}{n} \right\} \tag{1}$$

with a factor $l_{n,M}$ which is at most logarithmic in M and n. We show that SOI simultaneously implies several optimality properties: adaptivity to the sparsity pattern in high-dimensional linear models if the true f is linear; minimax adaptivity to the smoothness of f if $\widetilde{f}_n$ is used in classical nonparametric regression framework; optimality of aggregation rates [6] if $\widetilde{f}_n$ is viewed as an aggregate and the functions f_j as preliminary estimators based on a training sample considered as frozen. We argue that proof of optimality of estimators for these seemingly different problems can be done in a unified way via SOI.

One of the ways of constructing $\widetilde{f}_n$ satisfying SOI is the following [5]. Consider a "phantom" model $Y_i = \mathsf{f}_\lambda(X_i) + \xi_i'$ where ξ_i' are i.i.d. normally distributed random variables with mean 0 and variance $2\sigma^2$. Let π be a prior distribution on λ such that the components λ_j are i.i.d. with density $\tau^{-1} q_0(\cdot/\tau)$ where q_0 is the Student t_3 density, so that $q_0(t) \sim |t|^{-4}$, for $|t| \to \infty$, and $\tau > 0$ is a tuning parameter. Define now the estimator $\widetilde{f}_n^B$ as the Bayes posterior mean of f_λ under the "phantom" model and the prior π. Then the following result holds.

Theorem 1. *Let $\widetilde{f}_n^B$ be defined as above with $\tau = \sigma/\sqrt{n\,\mathrm{tr}(\Psi)}$. Then*

$$\mathbb{E}\|\widetilde{f}_n^B - f\|^2 \le \inf_{\lambda\in\mathbb{R}^M} \left\{ \|\mathsf{f}_\lambda - f\|^2 + C\sigma^2 \frac{M(\lambda)}{n} \Big(1 + \log_+ \Big\{ \frac{\sqrt{n\,\mathrm{tr}(\Psi)}}{\sigma} \|\lambda\|_\infty \Big\} \Big) \right\} + \frac{3\sigma^2}{n}$$

where $\|\lambda\|_\infty = \max_{j=1,\dots,M} |\lambda_j|$, $C > 0$ is an absolute constant and $\log_+ x = \max(\log x, 0)$.

Consider now a particular setting. Assume that we deal with high-dimensional linear regression model, i.e., $f = \mathsf{f}_{\lambda^*}$ with some $\lambda^* \in \mathbb{R}^M$, and possibly $M \gg n$. This is the framework considered in [4] among others who assumed in addition that the functions f_j are normalized: $\|f_j\| = 1$, $j = 1, \dots, n$. For $\lambda \in \mathbb{R}^M$ we set $\|\lambda\|_1 = \sum_{j=1}^M |\lambda_j|$. Then we get the following corollary of Theorem 1.

Corollary 1. *Let $\widetilde{f}_n^B$ be defined as in Theorem 1. If there exists $\lambda^* \in \mathbb{R}^M$ such that $f = \mathsf{f}_{\lambda^*}$ and $\|f_j\| \le 1$, $j = 1, \dots, M$, we have*

$$\mathbb{E}\|\widetilde{f}_n^B - \mathsf{f}_{\lambda^*}\|^2 \le C\sigma^2 \min\left(\frac{M(\lambda^*)\, l_{n,M}^*}{n}\,,\ \|\lambda^*\|_1 \sqrt{\frac{l_{n,M}^*}{n}} \right) + \frac{3\sigma^2}{n} \tag{2}$$

where $C > 0$ is an absolute constant and $l_{n,M}^*$ is a logarithmic factor:

$$l_{n,M}^* = 1 + \log_+ \left(\frac{\sqrt{nM}}{\sigma} \|\lambda^*\|_1 \right).$$

This corollary reveals an interesting effect: up to log-factors, the rate of convergence of $\widetilde{f}_n^B$ has two regimes: $M(\lambda^*)/n$ and $\|\lambda^*\|_1/\sqrt{n}$, with the change point at $M(\lambda^*) \sim \|\lambda^*\|_1\sqrt{n}$. The two characteristics of sparsity, $M(\lambda^*)$ and $\|\lambda^*\|_1$, are involved. For $M(\lambda^*) \ll \|\lambda^*\|_1\sqrt{n}$ the rate is determined by $M(\lambda^*)$, and for $M(\lambda^*) \gg \|\lambda^*\|_1\sqrt{n}$ by $\|\lambda^*\|_1$. Note that previously known bounds for the high-dimensional linear regression, cf. [1–5], are all of the order $M(\lambda^*)/n$, up to log-factors. The bound (2) gives an improvement, as compared to those results, in the situations when the vector λ^* contains relatively many non-zero components, for example, $M(\lambda^*) > n$ but the norm $\|\lambda^*\|_1$ is still small. Furthermore, (2) holds with no assumption on the dictionary $\{f_1, \dots, f_M\}$, except for the mere normalization. This stays in contrast with very restrictive assumptions on the dictionary needed to get bounds on the risks of the Lasso and Dantzig estimators [1–4].

References

[1] P.J. Bickel, Y. Ritov, and A.B. Tsybakov, *Simultaneous analysis of Lasso and Dantzig selector*, (2007), submitted. `www.proba.jussieu.fr/pageperso/tsybakov/BRT_LassoDan.pdf`

[2] F. Bunea, A.B. Tsybakov, and M.H. Wegkamp, *Aggregation for Gaussian regression*, Ann. Statist., **35** (2007), 1674–1697.

[3] F. Bunea, A.B. Tsybakov, and M.H. Wegkamp, *Sparsity oracle inequalities for the Lasso*, Electronic J. Statist., **1** (2007), 169–194.

[4] E. Candes, and T. Tao, *The Dantzig selector: statistical estimation when p is much larger than n*, Annals of Statistics, to appear.

[5] A. Dalalyan, and A.B. Tsybakov, *Aggregation by exponential weighting and sharp oracle inequalities*, Proc. 20th Annual Conf. Learning Theory (COLT-2007), Lecture Notes in Artificial Intelligence, **4539** (2007), 97–111, Springer, Berlin-Heidelberg.

[6] A.B. Tsybakov, *Optimal rates of aggregation*, Computational Learning Theory and Kernel Machines, (B. Schölkopf and M. Warmuth, eds.), Lecture Notes in Artificial Intelligence, **2777** (2003), 303–313, Springer, Berlin-Heidelberg.

Flexible modelling based on copulas in nonparametric regression

Ingrid Van Keilegom

(joint work with Roel Braekers)

Consider the model $Y = m(X) + \varepsilon$, where $m(\cdot) = \text{med}(Y|\cdot)$ is an unknown but smooth median regression function. It is often assumed that ε and X are independent. However, in many applications this assumption is violated. In this paper we propose to model the dependence between ε and X by means of a copula model, i.e.

$$(\varepsilon, X) \sim C_\theta(F_\varepsilon(\cdot), F_X(\cdot)),$$

where C_θ is a copula function depending on an unknown parameter θ. Since many copula families contain the independent copula as a special case, the so-obtained regression model is more flexible than the 'classical' regression model.

We estimate the parameter θ via a pseudo-likelihood method and prove the asymptotic normality of the estimator, based on delicate empirical process theory.

The procedure is illustrated by means of a simulation study, and the method is applied on data on food expenditures in households.

References

[1] R. Braekers, I. Van Keilegom, *Flexible modelling based on copulas in nonparametric regression*, submitted.

Regression models for survey data

Alan H. Welsh
(joint work with R.L. Chambers, D. Steel, S. Wang)

Most statisticians would agree that in modelling and inference, the structure of the population is important and that the relationship of the sample to the population is important. We consider the import of these statements in the context of trying to model relationships between variables using survey data. We consider a simple, abstract framework in which the information available on all the units in the population is represented by auxiliary variables and the information available on the in sample units alone is represented by survey variables. For simplicity, we assume non-informative sampling given the auxiliary variables and full response to the survey. We can consider two types of regression models. In disaggregated analysis, interest is in relating one of the survey variables (the response) to the other survey variables and the auxiliary variables while in aggregated analysis, interest is in relating the response to the other survey variables, marginally to the auxiliary variables. When the data on each unit in the population is independent, the standard approach is to formulate regression models (usually linear) at both levels and fit these by ordinary least squares using the sample data. By modelling the joint distribution of the auxiliary and survey variables, we can explore the relationship between the two analyses and derive the likelihood to see what the (asymptotically efficient) maximum likelihood estimators look like. We find that the standard approach is appropriate for disaggregated analysis: disaggregated analysis is achieved by modelling the relationship between the response, the other survey variables and the auxiliary variables, using the data from the in sample units alone. On the other hand, we also show that the distribution of the auxiliary variables affects the form of the aggregated model: we obtain linear models at both levels when the joint distribution is Gaussian but, in general, the two models are different and it does not make sense to impose the same form at both levels. In particular, with nonlinear models (such as logistic and log-linear models), the models incompatible if they are assumed to have the same form. Thus, the structure of the population (reflected in the auxiliary variables) matters. Also, even when the joint distribution of the variables is Gaussian, the maximum likelihood estimators of the parameters in the linear aggregated model are not the ordinary least squares estimators based on the sample data. In fact, they turn out to be the Pearson adjusted estimators which involve the auxiliary variables from all the units in the population, not just those in the sample. Intuitively, the relationship

between the auxiliary variables for the in sample units and the auxiliary variables for the population is used to adjust for the effects of sampling. Thus, the relationship between the sample and the population matters when we fit the aggregated model.

Extraction of primitive features from time series by complexity penalized M-estimation

Gerhard Winkler

(joint work with Felix Friedrich, Angela Kempe, Volkmar Liebscher, Darina Roeske, Olaf Wittich)

We discuss the topic of the workshop along a particularly simple example. We adopt a variational approach to the interpretation or explanation of time series, starting from the most primitive instance of complexity penalized functionals on the space of signals. Minimal points of such a functional can formally be interpreted as penalized M-estimators. They should exhibit a proper balance between fidelity to data and some well-defined regularity condition.

To be more precise, let $\{1, \dots, n\}$ be the discrete set of time points. Elements of $\mathbb{R}^n$ will be interpreted as time series or signals $x = (x_1, \dots, x_n)$. The set of its jumps is $J(x) = \{i = 1, \dots, n-1 : x_i \neq x_{i+1}\}$. The symbol $|A|$ denotes the cardinality of a set A. The functional we consider is of the form

$$P_\gamma : \mathbb{R}^n \times \mathbb{R}^n \longrightarrow \mathbb{R}, \quad (x, y) \longmapsto \gamma \cdot |J(x)| + \sum_{i=1}^{n} (y_i - x_i)^2, \ \gamma \geq 0,$$

with a control parameter γ. Vectors y represent data, and each signal x is a candidate for their representation.

The functional is a sum of two terms; one of them rates fidelity to data and the other one serves as a penalty or regularization term. The latter is designed in order to drive the estimate towards step functions with as few plateaus - or equivalently, as few jumps - as possible under the fidelity requirement. Hence it extracts primitive morphological features - in the present case plateaus or jumps - from data, and thus acts like a morphological filter. Emphasis is on the *interpretation of data*; there is no effort to restore some underlying true signal, in particular there is nothing like denoising.

Therefore this approach may justifiably be called *parsimonious* since it is an attempt to explain data in the sense of Ockham's razor: *Entia non sunt multiplicanda sine necessitate*. This is similar in spirit to the work of P.L. Davies and A. Kovac, see for example [3], who aim at an explanation of data by a minimal number of modes, using L^1-based approaches. For a deeper discussion of the underlying ideas see P.L. Davies [2].

The above method is also *sparse* in the sense of D.L. Donoho, M.Elad and V.N. Temlyakov, see [4]. In fact, O. Wittich et al. define and discuss *dictionaries* of appropriate indicator functions, see [22].

The circle of these ideas is systematically and (in a narrow sense) completely discussed in the series [19, 20, 7, 1, 21] of papers by the above authors, where [1] is joint work with L. Boysen and A. Munck. It consists of an introduction and overview [19], rigorous analytical results [20], fast algorithms [7], a statistical analysis [1], and a synapsis in the context of Mumford-Shah functionals [21]. The theory is exploited in [9] for the identification of biological noise and subsequently in [11] for the quality control of microarrays from molecular biology.

A final remark might contribute to some discussions during the present conference. Formally, there is no probability involved. Therefore, we have an instance of 'statistics without probability'. On the other hand, one may - somewhat artificially - interpret the functional as a negative log-posterior, despite a proper prior does not exist (because of invariance of the penalty w.r.t. to the addition of constants). Others, in turn, might consider the penalty as a regularization in the sense of numerical analysis. What an individual prefers seems to be a question of taste, force of habit, or provenience. This is an instance of a 'babel confusion of tongues' in the 'brave new world of statistics'.

References

[1] L. Boysen, V. Liebscher, A. Munk, and O. Wittich. *Jump-penalized least squares: Consistencies and rates of convergence*, The Ann. of Statist. (2008), in press.

[2] P. L. Davies, *Data features*, J. of the Netherlands Society for Statistics and Operations Research **49**(2) (1995), 185–245.

[3] P.L. Davies and A. Kovac, *Local extremes, runs, strings and multiresolution*, Ann. Stat., **29**(1) (2001), 1–65.

[4] D.L. Donoho, M. Elad, and V.N. Temlyakov, *Stable reconstruction of sparse overcomplete representations in the presence of noise*, IEEE Trans. Information Theory **52**(1) (2006), 6–18.

[5] F. Friedrich, *AntsInFields: Stochastic simulation and Bayesian inference for Gibbs fields*, (2003).

[6] F. Friedrich, *Complexity Penalized Segmentations in 2D - Efficient Algorithms and Approximation Properties*, PhD thesis, Munich University of Technology, Institute of Biomathematics and Biometry, National Research Center for Environment and Health, Munich, Germany (2005).

[7] F. Friedrich, A. Kempe, V. Liebscher, and G. Winkler, *Complexity penalized M-estimation: Fast computation*, JCGS (2007), in print.

[8] A. Kempe. *Statistical analysis of discontinuous phenomena with Potts functionals*, PhD thesis, Institute of Biomathematics and Biometry, National Research Center for Environment and Health, Munich, Germany (2004).

[9] V. Liebscher, A. Kempe, and G.Winkler, *Testing for noise by complexity penalised M-estimation*, (2008) in preparation.

[10] V. Liebscher and G. Winkler, *A Potts model for segmentation and jump-detection*, In V. Benes, J. Janacek, and I. Saxl, editors, Proceedings S4G International Conference on Stereology, Spatial Statistics and Stochastic Geometry, Prague June 21 to 24 1999, Union of Czech Mathematicians and Physicists, Prague (1999), 185–190.

[11] D. Roeske, V. Liebscher, and G. Winkler, *Parameter choice for penalized loglikelihood estimation and classification of signals*, (2008).

[12] G. Winkler, *Image Analysis, Random Fields and Markov Chain Monte Carlo Methods. A Mathematical Introduction*, Stochastic Modelling and Applied Probability **27**, Springer

Verlag, Berlin, Heidelberg, New York, second edition, 1995, 2003. Completely rewritten and revised, Corrected 3rd printing 2006.
[13] G. Winkler, *Image Analysis, Random Fields and Dynamic Monte Carlo Methods*, Beijing World Publishing Corporation, Beijing Peoples Republic of China (1999). Reprint of 'Image Analysis, Random Fields and Dynamic Monte Carlo Methods', Springer Verlag, 1995.
[14] G. Winkler, *Image Analysis, Random Fields and Dynamic Monte Carlo Methods*, GEO Publishing House, Novosibirsk, Russia (2002), 343 pages. In Russian.
[15] G. Winkler, A. Kempe, V. Liebscher, and O. Wittich, *Parsimonious segmentation of time series' by Potts models*, In D. Baier and K.-D. Wernecke, editors, Innovations in Classification, Data Science, and Information Systems. Proc. 27th Annual GfKl Conference, University of Cottbus, March 12 - 14, 2003., Studies in Classification, Data Analysis, and Knowledge Organization, Heidelberg-Berlin, Gesellschaft für Klassifikation, Springer-Verlag (2004), 295–302.
[16] G. Winkler and V. Liebscher, *A Potts model for segmentation and jump-detection*, In V. Benes, J. Janacek, and I. Saxl, editors, Proceedings S4G International Conference on Stereology, Spatial Statistics and Stochastic Geometry, Prague June 21 to 24, 1999, Union of Czech Mathematicians and Physicists, Prague (1999), 185–190.
[17] G. Winkler and V. Liebscher, *Smoothers for discontinuous signals*, J. Nonpar. Statist., **14**(1-2) (2002), 203–222.
[18] G. Winkler, V. Liebscher, and V. Aurich, *Probabilistic image smoothing: Recent advances*, In V. Benes, J. Janacek, and I. Saxl, editors, Proceedings S4G International Conference on Stereology, Spatial Statistics and Stochastic Geometry, Prague June 21 to 24, 1999, Union of Czech Mathematicians and Physicists, Prague (1999), 273–278.
[19] G. Winkler, O. Wittich, V. Liebscher, and A. Kempe, *Don't shed tears over breaks*, Jahresbericht der Deutschen Mathematiker-Vereinigung **107**(2) (2005), 57–87.
[20] O. Wittich, A. Kempe, G. Winkler, and V. Liebscher, *Complexity penalized sums of squares for time series': Rigorous analytical results*, Math. Nachr. (2006), In print.
[21] O. Wittich, V. Liebscher, and G. Winkler, *The Family of Mumford-Shah Functionals in Dimension One*(2008), In progress.
[22] O. Wittich, V. Liebscher, and G. Winkler, *The Potts model is sparse* (2008) In progress.

Smooth interpolation

Henry Wynn
(joint work with Hugo Maruri-Aguilar)

1. Introduction

Techniques from computational commutative algebra were first applied to design of experiments in [5]. The fundamental principle is to study the design through a related algebraic object: the *design ideal*. By this means a linear saturated model for the response on the design can be constructed and confounding relations induced by the design can be generalised to essentially, any design, see [6] and [7].

We are concerned about extending algebraic techniques to produce interpolators which also satisfy desired smoothness properties. We give a general proposal and illustrate with an example.

2. Algebraic interpolation

For a set of nonnegative integers $\alpha = (\alpha_1, \ldots, \alpha_k)$, a monomial is the power product $x^\alpha := x_1^{\alpha_1} \cdots x_k^{\alpha_k}$, and a polynomial is a linear combination of monomials. Let $\mathbb{R}[x_1, \ldots, x_k] = \mathbb{R}[x]$ be the set of all polynomials in indeterminates $x_1, \ldots, x_k$ and coefficients in $\mathbb{R}$. $\mathbb{R}[x]$ is known as the polynomial ring. A term order $\prec$ on $\mathbb{R}[x]$ is a total order on the set of all monomials in the indeterminates $x_1, \ldots, x_k$. The leading term of a polynomial is the highest term with respect to $\prec$ with non zero coefficient.

A design $\mathcal{D}$ is a set of n distinct points in $\mathbb{R}^k$, where k is the number of factors. The design ideal is $I(\mathcal{D}) = \{f \in \mathbb{R}[x] : f(x) = 0, x \in \mathcal{D}\}$, that is, the set of all polynomials that, as polynomial functions, vanish over the design points.

Given a term ordering $\prec$, a Gröbner basis for $I(\mathcal{D})$ is a finite subset $G_\prec \subset I(\mathcal{D})$ such that the ideal generated by the set of leading terms of polynomials in $G_\prec$ coincides with the ideal generated by the leading terms of all polynomials in $I(\mathcal{D})$, see [2]. Computation of Gröbner bases is implemented in computer programs such as CoCoA or Singular, see also [2].

Linear independence of monomials over a design can then be studied using the design ideal. Given a term order $\prec$ and a Gröbner basis $G_\prec$ for $I(\mathcal{D})$, $G_\prec = \{g_1, \ldots, g_m\}$, then every polynomial $f \in \mathbb{R}[x]$ can be expressed modulo $I(\mathcal{D})$ as

$$f = \sum_{i=1}^{m} g_i s_i + r$$

where $s_i \in \mathbb{R}[x]$ and the remainder r is unique and is composed only with monomials that cannot be divided by the leading terms of $G_\prec$. We call this set of monomials as the set of standard monomials.

The quotient ring $\mathbb{R}[x]/I(\mathcal{D})$ is the set of equivalence classes created by the above decomposition. As a $\mathbb{R}$-vector space, $\mathbb{R}[x]/I(\mathcal{D})$ is isomorphic to the set of polynomial functions $\varphi : \mathcal{D} \mapsto \mathbb{R}$. Moreover, the set of standard monomials forms a monomial basis for $\mathbb{R}[x]/I(\mathcal{D})$ and thus they form the support for any interpolator model over the design. In other words, for a set of observations at the design points $\{y_x, x \in \mathcal{D}\}$, algebraic techniques always identify the support for a saturated model. We refer to this model as the algebraic interpolator $\widehat{y}(x)$, as it satisfies $\widehat{y}(x) = y_x$ for $x \in \mathcal{D}$. Algebraic interpolators satisfy desirable properties such as marginality, see [4] and are of minimal average degree, see [1].

3. Smoothing the interpolator

One potential drawback of algebraic interpolators is that they are not necessarily smooth. A smooth version of the univariate algebraic interpolator is obtained by considering

$$\widetilde{y}(x) = \widehat{y}(x) - s(x)g(x), \tag{1}$$

where $g(x)$ is the single polynomial which forms the Gröbner basis in the univariate case and $s(x) \in \mathbb{R}[x]$. As $g(x) = 0$ for $x \in \mathcal{D}$ then $\widetilde{y}(x)$ is still an interpolator.

We select $s(x)$ to be a high degree polynomial, that is $s_t(x) = \sum_{i=0}^{t} \theta_i x^i$. The coefficients $\theta_0, \ldots, \theta_t$ of $s_t(x)$ are chosen to minimise a measure of smoothness for $\widetilde{y}_t(x) = \widehat{y}(x) - s_t(x)g(x)$. The following measure of smoothness is proposed

$$\phi_2 = \int_{\mathcal{X}} |\widetilde{y}_t''(x)|^2 dx, \tag{2}$$

where the integration is carried out over a desired interval $\mathcal{X} \subset \mathbb{R}$.

The measure ϕ_2 is quadratic in the unknown parameters $\theta_0, \ldots, \theta_t$, and the minimum solution can be found in closed form. Let $\widetilde{y}_t^*(x)$ be the interpolator that minimises Equation 2. The smooth interpolator $\widetilde{y}_t^*(x)$ is a linear in terms of the observations y_x. The minimal value of ϕ_2 decreases to a limit, as function of t.

For example, consider the response values $1, -5, 2, 7$ at design points $0, 1, 2, 3$ and $\mathcal{X} = [0, 3]$. The algebraic interpolator $\widehat{y}(x)$ gives a value of $\sqrt{\phi_2}$ of 63.023. Using the method above described gives values $63.019, 44.168, 44.168, 42.329$ for $\sqrt{\phi_2}$ when using $t = 0, 1, 2, 3$, respectively.

For the present smoothing problem, if instead we minimise ϕ_2 by searching over all functions with absolutely continuous second derivatives, the minimum value of ϕ_2 is achieved with the cubic spline. This property of cubic splines was first shown by [3]. In the example above, the minimal value of $\sqrt{\phi_2}$ using a cubic spline is 42.028.

The above procedure to smooth algebraic interpolators extends to problems for which the design $\mathcal{D}$ lies in higher dimensions. The procedure starts with setting a term order $\prec$ and then computing the algebraic interpolator $\widehat{y}(x)$ for a given data set. The smooth interpolator is

$$\widetilde{y}(x) = \widehat{y}(x) - \sum_{i=1}^{m} s_i(x) g_i(x), \tag{3}$$

where $g_i(x)$ are polynomials in the Gröbner basis for $I(\mathcal{D})$ and $s_i(x)$ are polynomials of $\mathbb{R}[x]$ whose coefficients are selected to minimise a measure of smoothness. For example, the simplest instance of $s_i(x)$ is to set $s_i(x) = \theta_i$, and then $\widetilde{y}(x)$ has m parameters. The measure of smoothness to minimise is

$$\phi_2 = \sum \left(\frac{\partial^2 \widetilde{y}(x)}{\partial x_i \partial x_j} \right)^2 = ||H||^2 = \text{trace}(H^T H), \tag{4}$$

where H is the Hessian matrix for $\widetilde{y}(x)$. This minimisation problem is quadratic in the parameters of the $s_i(x)$ and the smooth interpolator $\widetilde{y}^*(x)$ is linear on the observations y_x.

4. Further work

The algebraic methodology proposed in Section 3 produces a smooth polynomial interpolator, which in a limiting condition, turns to a cubic spline. However, the methodology still depends on the term ordering used and on the specified form for $s_i(x)$.

A proposal that avoids both inconveniences is to interpolate using dummy observations on a regular grid that contains the original design $\mathcal{D}$. These dummy

observations turn to be parameters which are selected to minimise Equation (4). That is, for a set of observations y_x taken on design points $x \in \mathcal{D}$, to interpolate an extended set $\{y_x, z_b\}$ where z_b are dummy observations taken at points $\{b\}$ such that the extended design $\{x, b\}$ is a regular grid. The interpolator does not depend on term orderings and the smoothing problem is quadratic in the parameters and it has closed solution.

For example, consider observations $5, -4, 0, -1, 8$ taken at the Latin hypercube design points $(-2,-2), (-1,0), (0,2), (1,-1), (2,1)$. The extended design is a 5×5 grid with levels $-2, -1, 0, 1, 2$ for each variable. The minimisation problem returns a smooth interpolator $\widetilde{y}^*(x)$ whose value for $\sqrt{\phi}_2$ is 10.5. This compares favorably with that for the algebraic interpolator of 28.2.

References

[1] Y. Bernstein, H. Maruri-Aguilar, S. Onn, E. Riccomagno and H. Wynn, *Minimal average degree aberration and the state polytope for experimental designs*, (2007), submitted.

[2] D. Cox, J. Little and D. O'Shea, *Ideals, varieties and algorithms*, Springer (1996).

[3] J.H. Holladay, *A smoothest curve approximation*. Mathematical Tables and Other Aids to Computation **11** (1957), 233-243.

[4] P. McCullagh and J.A. Nelder, *Generalized Linear Models*, Chapman & Hall (1983).

[5] G. Pistone and H. Wynn, *Generalised confounding with Gröbner bases*, Biometrika **83**(3) (1996), 653-666.

[6] G. Pistone, E. Riccomagno and M.P. Rogantin, *Algebraic Statistics methods for DOE*, Research report 18, Politecnico di Torino (2006).

[7] G. Pistone, E. Riccomagno and H. Wynn, *Algebraic Statistics*, Chapman & Hall (2000).

Reporter: Ursula Gather

Participants

Sylvain Arlot
Laboratoire de Mathematiques
Universite Paris Sud (Paris XI)
Batiment 425
F-91405 Orsay Cedex

Prof. Dr. Rudolf J. Beran
Department of Statistics
University of California
Davis
One Shields Avenue
Davis CA 95616
USA

Prof. Dr. Lucien Birge
Laboratoire de Probabilites-Tour 56
Universite P. et M. Curie
4, Place Jussieu
F-75252 Paris Cedex 05

Prof. Dr. Lawrence D. Brown
Department of Statistics
The Wharton School
University of Pennsylvania
3730 Walnut Street
Philadelphia , PA 19104-6340
USA

Dr. Peter Bühlmann
Seminar für Statistik
ETH-Zürich
LEO C 17
CH-8092 Zürich

Prof. Dr. Raymond J. Carroll
Department of Statistics
Texas A & M University
College Station , TX 77843-3143
USA

Prof. Dr. Gerda Claeskens
OR & Business Statistics
K.U. Leuven
Naamsestraat 69
B-3000 Leuven

Prof. Dr. P. Laurie Davies
Fachbereich Mathematik
Universität Duisburg-Essen
45117 Essen

Prof. A. Philip Dawid
Centre for Mathematical Sciences
University of Cambridge
Wilberforce Road
GB-Cambridge CB3 OWB

Prof. Dr. Holger Dette
Fakultät für Mathematik
Ruhr-Universität Bochum
Universitätsstr. 150
44801 Bochum

Stefanie Donauer
Faculteit Wiskunde en Informatica
Vrije Universiteit Amsterdam
De Boelelaan 1081 a
NL-1081 HV Amsterdam

Prof. Dr. Lutz Dümbgen
Mathematische Statistik
und Versicherungslehre
Universität Bern
Sidlerstraße 5
CH-3012 Bern

Prof. Dr. Stephen E. Fienberg
Dept. of Statistics
Carnegie Mellon University
Pittsburgh , PA 15213
USA

Prof. Dr. Ursula Gather
Fachbereich Statistik
Universität Dortmund
44221 Dortmund

Prof. Dr. Edward I. George
Department of Statistics
The Wharton School
University of Pennsylvania
3730 Walnut Street
Philadelphia , PA 19104-6340
USA

Dr. Peter Grünwald
CWI
Postbus 94079
NL-1090 GB Amsterdam

Prof. Dr. Wolfgang Härdle
Wirtschaftswissenschaftl. Fakultät
Lehrstuhl für Statistik
Humboldt-Universität Berlin
Spandauer Str. 1
10178 Berlin

Prof. Dr. Peter G. Hall
Dept. of Mathematics and Statistics
University of Melbourne
Melbourne VIC 3010
AUSTRALIA

Prof. Dr. Nils Lid Hjort
Department of Mathematics
University of Oslo
P. O. Box 1053 - Blindern
N-0316 Oslo

Dr. Geurt Jongbloed
Delft Institute of Applied
Mathematics
Delft University of Technology
Mekelweg 4
NL-2628 CD Delft

Prof. Dr. Arne Kovac
School of Mathematics
University of Bristol
University Walk
GB-Bristol BS8 1TW

Prof. Dr. Hans Rudolf Künsch
Seminar für Statistik
ETH-Zentrum Zürich
LEO D2
Leonhardstr. 27
CH-8092 Zürich

Claire Lacour
Universite Rene Descartes
UFR Mathematiques et Informatique
45, rue des Saints-Peres
F-75270 Paris Cedex 6

Prof. Dr. Hannes Leeb
Department of Statistics
Yale University
P.O.Box 208290
New Haven , CT 06520-8290
USA

Marloes Maathuis
Seminar für Statistik
ETH-Zentrum Zürich
LEO D2
Leonhardstr. 27
CH-8092 Zürich

Lukas Meier
Departement Mathematik
ETH-Zentrum
Rämistr. 101
CH-8092 Zürich

Dr. Monika Meise
Fachbereich Mathematik
Universität Duisburg-Essen
45117 Essen

Thoralf Mildenberger
Fachbereich Statistik
Universität Dortmund
44221 Dortmund

Dr. Samuel Müller
School of Mathematics and Statistic
University of Western Australia
35 Stirling Highway
Crawley WA 6009
AUSTRALIA

Prof. Dr. Axel Munk
Institut f. Mathemat. Stochastik
Georg-August-Universität Göttingen
Maschmühlenweg 8-10
37073 Göttingen

Dr. Natalie Neumeyer
Department of Mathematics
University of Hamburg
Bundesstr. 55
20146 Hamburg

Prof. Dr. Benedikt M. Pötscher
Institut für Statistik
Universität Wien
Universitätsstr. 5/3
A-1010 Wien

Prof. Dr. Jim Ramsay
2748 Howe St.
Ottawa ON K2B 6W9
CANADA

Prof. Dr. Jorma R. Rissanen
140 Teresita Wy
Los Gatos , CA 95032
USA

Prof. Dr. Elvezio Ronchetti
Department of Econometrics
University of Geneva
Blv.Pont d'Arve 40
CH-1211 Geneve

Prof. Dr. Ritei Shibata
Dept. of Mathematics
Keio University
Hiyoshi 3-14-1, Kohokuku
Yokohama 223-8522
JAPAN

Prof. Dr. Vladimir Spokoiny
Weierstrass-Institute for Applied
Analysis and Stochastics
Mohrenstr. 39
10117 Berlin

Dipl-Math. Rahel Stichtenoth
Fachbereich Mathematik
Universität Duisburg-Essen
45117 Essen

Prof. Dr. Alexandre B. Tsybakov
Laboratoire de Probabilites
Universite Paris 6
4 place Jussieu
F-75252 Paris Cedex 05

Prof. Dr. Ingrid Van Keilegom
Institut de Statistique
Universite Catholique de Louvain
Voie du Roman Pays 20
B-1348 Louvain-la-Neuve

Prof. Dr. Alan Welsh
Centre for Mathematics and its
Applications
Australian National University
Canberra ACT 0200
AUSTRALIA

Prof. Dr. Gerhard Winkler
GSF-Biomathematik und Biometrie
Postfach 1129
85758 Oberschleissheim

Birgit I. Witte
Delft Institute of Applied Mathematics
Delft University of Technology
Mekelweg 4
NL-2628 Delft CD

Prof. Dr. Henry P. Wynn
Department of Statistics
London School of Economics
Houghton Street
GB-London WC2A 2AE