Mathematisches Forschungsinstitut Oberwolfach

Report No. 52/2004

New Inference Concepts for Analysing Complex Data

Organised by
Jianqing Fan (Chapel Hill)
Klaus-Robert Müller (Berlin)
Vladimir Spokoiny (Berlin)

November 14th – November 20th, 2004

Abstract. The main purpose of this workshop was to assemble international leaders from statistics and machine learning to identify important research problems, to cross-fertilize between the disciplines, and to ultimately start coordinated research efforts toward better solutions. The workshop focused on discussing modern methods for analysis complex high dimensional data with applications to econometrics, finance, biomedicine, genomics etc.

Mathematics Subject Classification (2000): 62G,62H,68T.

Introduction by the Organisers

The workshop *New Inference Concepts for Analysing Complex Data*, organised by Jianqing Fan (Chapel Hill), Klaus-Robert Müller (Berlin) and Vladimir Spokoiny (Berlin) was held November 14st–November 20th, 2004. This meeting was well attended with about 45 participants with broad geographic representation from all continents. This workshop was a nice blend of researchers with various backgrounds from the areas of statistics and machine learning.

The main purpose of this workshop was to assemble international leaders from statistics and machine learning in the Institute: to identify important research problems, to cross-fertilize between the disciplines, and to ultimately start coordinated research efforts towards better solutions.

The program included more than 25 talks organized in sections on various topics: *Support Vector Machines* (B. Schölkopf, A.J. Smola, G.Wahba), *mathematical finance* (H. Dette, W. Härdle, S.X. Chen), *dimension reduction* (Y. Xia, A. Dalalyan, M. Kawanabe, E. Mammen, H. H. Zhang), *non-parametric smoothing* (D. Belomestny, J. Polzehl), *Boosting* (P. Bühlmann, Y. Ritov), *genomics* (A. Nobel, Hepping Zhang), *classification* (S. van de Geer, G. Blanchard), *statistical inverse problem* (M. Reiss, A. Goldenshluger), *Functional regression* (H.-G. Müller,

T.Cai), *Clustering* (J. M. Buhmann, A. Nobel) among others. P. Bühlmann gave an overview on statistical methods of *boosting* while D. Yekutieli made an extended presentation on *False Discovery Rate*.

There was one organized discussion on *Smoothing methods in classification*. The workshop gave an excellent opportunity for exchanging the opinions and expertise as well as for discussing various topics in different areas of modern mathematical statistics and machine learning theory. The discussion revealed a lot of common ideas and principles in these two fields but also differences in the methodology and approaches. An exchange of ideas can clearly contribute to the both fields. Already during workshop some new projects were originated that involve both statisticians and people from machine learning society.

The workshop was attended by a number of young statisticians and gave an excellent opportunity for training: both by attending the high level presentation, by presenting their own results and by participating at the numerous informal discussions.

Workshop: New Inference Concepts for Analysing Complex Data

Table of Contents

Abstracts

Local likelihood modelling via stagewise aggregation

Denis Belomestny
(joint work with V.Spokoiny)

The aim of this talk is to propose a new method of spatially adaptive nonparametric estimation based on aggregating a family of local likelihood estimates. Local likelihood approach was intensively discussed last years, see e.g. Tibshirani and Hastie (1987), Staniswalis (1989), Loader (1996). We refer to Fan, Farmen and Gijbels (1998) for a nice and detailed overview of local maximum likelihood approach and related literature. The approach is very general and applies to many statistical models. An important issues for local likelihood modeling is the choice of localization (smoothing) parameters. Different types of model selection techniques based on the asymptotic expansion of the local likelihood are mentioned in Fan, Farmen and Gijbels (1998) which includes global and variable bandwidth selection. However, the performance of estimators based on bandwidth selection is often rather unstable, see e.g. Breiman (1996). This suggests that in some cases, the attempt to identify the true local model is not necessarily the right thing to do. One approach to reduce variability in model selection is model mixing or aggregation. Yang (2004), Catoni (2001) among other suggested global aggregated procedures that achieves the best estimation risks over the family of given "weak" estimates. Nemirovski (2000), Juditsky and Nemirovski (2000) developed the aggregation procedures that achieves up to some log-factor the minimal risk in the class of all convex combinations of "weak" estimates in regression setup. Tsybakov (2003) discussed the asymptotic minimax rate for aggregation. Aggregation for density estimation has been investigated by Li and Barron (1999). A pointwise aggregation has not been yet considered to the best of our knowledge.

We propose a new approach towards local likelihood modelling which is based on the idea of the spatial aggregation of a "nested" family of local likelihood estimates ("weak" estimates) $\tilde{\theta}^{(k)}$. The main idea is, given the sequence $\{\tilde{\theta}^{(k)}\}$ to construct in a data driven way the "optimal" aggregated estimate $\hat{\theta}(x)$ separately at each point x. "Optimality" means that this estimate satisfies some kind of oracle inequality, that is, its pointwise risk does not exceed the smallest pointwise risk among the all "weak" estimates up to a logarithmic multiple.

Our algorithm can be roughly described as follows. Let $\{\tilde{\theta}^{(k)}(x)\}$, $k = 1, \ldots, k$ be an "ordered" sequence of weak local likelihood estimates at a point x. A new aggregated estimate of $\theta(x)$ is constructed sequentially by mixing the previously constructed aggregated estimate $\hat{\theta}^{(k-1)}$ with the current "weak" estimate $\tilde{\theta}^{(k)}$:

$$\hat{\theta}^{(k)} = \gamma_k \tilde{\theta}^{(k)} + (1 - \gamma_k)\hat{\theta}^{(k-1)},$$

where the mixing parameter γ_k (which may depend on the point x) is defined using a measure of statistical difference between $\hat{\theta}^{(k-1)}$ and $\tilde{\theta}^{(k)}$. In particular, γ_k is equal

to zero if $\hat{\theta}^{(k-1)}$ lies outside the confidence interval of $\tilde{\theta}^{(k)}$. In view of the sequential and poinwise nature of the algorithm, the suggested procedure is called *Spatial Stagewise Aggregation* (SSA). An important feature of the proposed procedure is that it is very simple and transparent and applies in a unified manner for a big family of different models like Gaussian, binary, Poisson regression, density estimation, classification etc. The procedure does not require any splitting of the sample as many other aggregation procedures do, cf. Tsybakov (2003). The SSA procedure can be easily studied theoretically. We establish precise nonasymptotic "oracle" results which apply under very mild conditions in a uniform manner to many different statistical models and problems. We also show that the oracle property automatically implies spatial adaptivity of the proposed estimate.

References

[1] Breiman, L. (1996). Stacked regression. *Machine Learning*, **24** 49–64.

[2] Catoni, O. (1999). "Universal" aggregation rules with exact bias bounds. Preprint.

[3] Fan, J. and Gijbels, I. (1996). *Local polynomial modelling and its applications*. Chapman & Hall, London.

[4] Juditsky, A. and Nemirovski, A. (2000). Functional aggregation for nonparametric estimation. *Ann. Statist.*, **28** 682–712.

[5] Li, J. and Barron, A. (1999). Mixture density estimation. In. S.A. Sola, T.K. Leen, and K.R. Mueller, editors, *Advances in Neural Inforamtion proceedings systems* **12**

[6] Loader, C. R. (1996). *Local likelihood density estimation*. Academic Press.

[7] Staniswalis, J.C. (1989). The kernel estimate of a regression function in likelihood-based models. *Journal of the American Statistical Association*, **84** 276–283.

[8] Tibshirani, J.R., and Hastie, T.J. (1987). Local likelihood estimation. *Journal of the American Statistical Association*, **82** 559–567.

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

[10] Yang, Y. (2001). Adaptive regression for mixing. *Journal of the American Statistical Association*, **96** 574–588.

[11] Yang, Y. (2004). Aggregating regression procedures to improve performance. *Bernoulli* **10** no. 1, 25–47

Oracle bounds and algorithm for optimal dyadic tree classification

Gilles Blanchard

(joint work with C.Schäfer, Y.Rozenholc, K-R. Müller)

Overview. We present a new algorithm to build a single dyadic classification tree for multiclass data. Although not as effective in terms of raw generalization error as recent large margin classifiers or ensemble methods, single classification trees possess important added values: they are easier to interpret for practitioners, they are naturally adapted to multi-class situations and they provide additional and finer information through conditional class probability (hereafter ccp) estimation. We start with the *a priori* that we accept to lose a little on the performance side in order to get these advantages as a counterpart. Furthermore, we show on

experiments that our method outperforms classical single tree methods (Quinlan's C4.5).

The fact that we are considering only *dyadic* tree models allows us to build an algorithm to find such a tree that *globally* minimizes some arbitrary (penalized) loss criterion. This is essential because the greedy way that classical decision tree methods (C4.5, CART) build up the tree can be shown to yield arbitrary bad results in some cases. The present ODCT algorithm finds some of its sources in [4], where a similar method is proposed for regression in 2D problems with equispaced data, and oracle inequalities are derived for the L_2 norm. Our results extend this setting to higher dimension, other loss functions, and arbitrary distribution of the observations. We also present an algorithm that is computationally more efficient for this more general setting.

Dyadic partitions. We assume that observations x lie in $[0,1]^d$ and class y belongs to a finite set $\{1,\dots,S\}$. We oberve an i.i.d. sample $(X_i,Y_i)_{i=1\dots n}$ following an unknown probability distribution P.

The base model we consider is given by *dyadic trees* or equivalently *dyadic partitions*. A dyadic partition $\mathcal{B}$ is defined as a partition of $[0,1]^d$ formed only of hyperrectangles of the form $\Pi_{i=1}^d \left[\frac{k_i}{2^{j_i}} \frac{k_i+1}{2^{j_i}}\right]$. Equivalently, it can be seen as a dyadic decision tree where $[0,1]^d$ is recursively divided in half into smaller pieces, each "cut" being allowed only through the center of the piece and perpendicular to one of the coordinate axes.

For a fixed partition $\mathcal{B}$ we define the histogram estimator $\widehat{f_{\mathcal{B}}}$ of $P(Y|X)$:

$$\forall b \in \mathcal{B}, \qquad \forall x \in b, \qquad \widehat{f_{\mathcal{B}}}(x,y) = \frac{N_{b,y}}{\sum_y N_{b,y}}, \tag{1}$$

where $N_{b,y}$ denotes the number of training points of class y falling in bin b.

Penalized minimum empirical loss. The definition of our final estimator is $\widehat{f_{\widehat{\mathcal{B}}}}$ where a suitable partition $\widehat{\mathcal{B}}$ is selected via the following penalized empirical risk minimization:

$$\widehat{\mathcal{B}} = \operatorname*{Arg\,Min}_{\mathcal{B}} \frac{1}{n} \sum_{i=1}^{n} \ell(\widehat{f}_B, X_i, Y_i) + \gamma|\mathcal{B}|, \tag{2}$$

where ℓ denotes a loss function; we consider classification error (when we predict the majority class according to f), minus log-likelihood ($\ell(f,x,y) = -\log f(x,y)$), or square error ($\ell(f,x,y) = \|f(x,\cdot) - \widetilde{y}\|^2$, where $\widetilde{y}$ denotes the k-dimensional vector which has 1 as the y-th coordinate and 0 elsewhere).

Oracle bounds. Let f^* denote either the true conditional probability distribution of the class $P(Y|X)$ (in the case the log-likelihood or square loss is used) or the Bayes classifier (in the case the misclassification loss function is used). Note that f^* is the minimizer of the true average loss.

The theoretical properties that our method enjoys are summed up in the following "theorem template":

Theorem Template 1. *Let $f^*(x,y) = P(Y=y|X=x)$, and $\mathcal{Z} = (X_i,Y_i)_{i=1\dots n}$ an i.i.d. sample of size n drawn according to P. Then for a suitable choice of γ,*

the estimator $\widehat{f}$ *defined by* (2) *satisfies the following oracle-type inequality:*

$$E_{\mathcal{Z}} E\left[\ell(\widehat{f}) - \ell(f^*)\right] \leq 2 \inf_{\mathcal{B}} \inf_{f \in \mathcal{C}_{\mathcal{B}}} \left(E\left[\ell(f) - \ell(f^*)\right] + 2\gamma|\mathcal{B}| + \frac{C}{n} \right), \tag{3}$$

where $\mathcal{C}_{\mathcal{B}}$ *denotes the set of ccp functions that are piecewise constant on the bins of* $\mathcal{B}$.

This theorem is satisfied by the three mentioned loss functions under additional technical assumptions depending on the loss function, that we do not detail here; in all cases the most important assumption is that γ should be greater that a function of order $\log(n)/n$.

Note that for classification loss, $E\left[\ell(\widehat{f}) - \ell(f^*)\right]$ is the excess loss with respect to the Bayes classifier. For square loss, it is the average sum over classes of square difference between the estimate and the true cpp, while for log-likelihood loss, it is the average conditional Kullback-Leibler divergence between these same quantities.

Anisotropy adaptivity. The most prominent consequence of these theoretical properties is that our algorithm is *adaptive to anisotropy*, which means that if the target function $P(Y|X)$ is more regular in one axis direction than another, this property will be "caught" by the algorithm – because the target is best approximated by dyadic trees that have more cuts in the less regular direction (i.e. "elongated" bins) and the selected tree will be of this type. As an extreme case, if some directions are in fact pure noise and irrelevant to the classification task, they will be ignored.

Formally, we are able to prove that in the case of square error loss, this anisotropy property can be quantified in terms of attaining the minimax rate of convergence for certain classes of weak Hölder functions where the Hölder smoothness is not the same in all directions. Donoho [4] considered very closely related anistropy classes in the case of regression with Gaussian white noise and when P is the Lebesgue measure. In our case the setting is more general as we are able to obtain convergence rates for an arbitrary P - albeit under somewhat stronger assumptions on the function class.

Exact algorithm. In contrast to CART or C4.5 where only an approximate, greedy solution of the optimization problem is computed, we are able to propose an algorithm to compute the exact optimum of eq. (2). This method is inspired by an algorithm initially proposed by Donoho [4] that we further improve to yield better computing efficiency (in the case of higher-dimensional, non-equispaced data).

We proved that if we assume that k_{max}, the maximum number of cuts along a given direction, is of the order of $\log(n)$, the complexity of the dictionary-based algorithm is of order $\mathcal{O}(n \log(n)^{d+1})$, as opposed to n^d for the naive approach taking all bins into account. This is a very important improvement but the complexity is unfortunately still exponential in the dimension, which means that the full algorithm can only be expected to be usable for smaller values of d.

Experiments. Experimental results on real and simulated data showed the following about the ODCT method:

- it generally yields better results than C4.5

- it effectively discards non-informative dimensions
- it adapts to anistropy and irregular data distributions as expected from the theory
- it is robust wrt. to flipping noise.

References

[1] A. Barron, L. Birgé, and P. Massart. Risk bounds for model selection via penalization. *Probability theory and related fields*, 113:301–413, 1999.

[2] G. Blanchard, C. Schäfer, and Y. Rozenholc. Oracle bounds and exact algorithm for dyadic classification trees. In J. Shawe-Taylor and Y. Singer, editors, *Proceedings of the 17th Conference on Learning Theory (COLT 2004)*, number 3210 in Lectures Notes in Artificial Intelligence, pages 378–392. Springer, 2004.

[3] L. Breiman, J. Friedman, J. Olshen, and C. Stone. *Classification and Regression Trees*. Wadsworth, 1984.

[4] D. Donoho. Cart and best-ortho-basis: a connection. *Annals of Statistic*, 25:1870–1911, 1997.

[5] J.R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, 1993.

[6] C. Scott and R. Nowak. Near-minimax optimal classification with dyadic classification trees. In *Proc. Neural Information Processing Systems*, 2003.

[7] C. Scott and R. Nowak. Minimax optimal classification with dyadic decision trees. Technical Report TREE0403, Rice University, 2004.

Boosting: a Statistical Perspective

Peter Bühlmann

Boosting algorithms have been proposed in the machine learning literature by Schapire ([13]) and Freund ([8], [9]), see also [14]. These first algorithms have been developed as ensemble methods. Boosting has been empirically demonstrated to be very accurate in terms of classification, notably the so-called AdaBoost algorithm ([9]).

We will explain that boosting can be viewed as a nonparametric optimization algorithm in function space, as first pointed out by Breiman ([1], [2]). This view turns out to be very fruitful to adapt boosting for other problems than classification, including regression and survival analysis.

We will mainly focus on boosting with the squared error loss (L_2Boosting; cf. [10]) and discuss the following: L_2Boosting for nonparametric additive and second-order interaction modeling ([5]); L_2Boosting for high-dimensional linear models and overcomplete dictionaries ([4]) and its relation to Matching Pursuit ([12]), Lasso and LARS ([7]); conjugate gradient descent L_2Boosting as an alternative to gradient descent type boosting ([11]); L_2Boosting with degrees of freedom penalties ([6]) and its relation to Breiman's nonnegative garrote estimator ([3]).

References

[1] Breiman, L.: Arcing classifiers. Annals of Statistics **26**, 801–824 (1998).

[2] Breiman, L.: Prediction games & arcing algorithms. Neural Computation **11**, 1493–1517 (1999).

[3] Breiman, L.: Better subset regression using the nonnegative garrote. Technometrics **37**, 373–384 (1995).
[4] Bühlmann, P.: Boosting for high-dimensional linear models. Preprint (2004).
[5] Bühlmann, P., Yu, B.: Boosting with the L_2loss: regression and classification. J. American Statistical Association **98**, 324–339 (2003).
[6] Bühlmann, P., Yu, B.: Boosting, Model Selection, Lasso and Nonnegative Garrote. Preprint (2004).
[7] Efron, B., Hastie, T., Johnstone, I., Tibshirani, R.: Least angle regression. To appear in Annals of Statistics (2004).
[8] Freund, Y. (1995): Boosting a weak learning algorithm by majority. Information and Computation **121**, 256–285 (1995).
[9] Freund, Y., Schapire, R.E.: Experiments with a new boosting algorithm. In Machine Learning: Proc. Thirteenth International Conference, pp. 148–156. Morgan Kauffman, San Francisco (1996).
[10] Friedman, J.H.: Greedy function approximation: a gradient boosting machine. Annals of Statistics **29**, 1189–1232 (2001).
[11] Lutz, R.W., Bühlmann, P.: Conjugate direction boosting for regression. Preprint (2004).
[12] Mallat, S., Zhang, Z. (1993). Matching pursuits with time-frequency dictionaries. IEEE Transactions Signal Processing **41**, 3397–3415 (1993).
[13] Schapire, R.E.: The strength of weak learnability. Machine Learning **5**, 197–227 (1990).
[14] Schapire, R.E.: The boosting approach to machine learning: an overview. In: MSRI Workshop on Nonlinear Estimation and Classification (Eds. Denison, D.D., Hansen, M.H., Holmes, C.C., Mallick, B., Yu, B). Springer, New York (2002).

Data clustering in imaging

Joachim M. Buhmann
(joint work with Tilman Lange, Volker Roth and Mikio L. Braun)

Data clustering describes a set of frequently employed techniques in exploratory data analysis to extract "natural" group structure in data. Such groupings need to be validated to separate the signal in the data from spurious structure. In this context, finding an appropriate number of clusters is a particularly important model selection question. We introduce a measure of cluster stability to assess the validity of a cluster model. This stability measure quantifies the reproducibility of clustering solutions on a second sample and it can be interpreted as a classification risk w.r.t. class labels produced by a clustering algorithm. The preferred number of clusters is determined by minimizing this classification risk as a function of the number of clusters. Convincing results are achieved on simulated as well as gene expression data sets. Comparisons to other methods demonstrate the competitive performance of our method and its suitability as a general validation tool for clustering solutions in real world problems.

Prediction in Functional Linear Regression

T. Tony Cai
(joint work with Peter Hall)

In the problem of functional linear regression we observe data

$$\{(X_1, Y_1), \ldots, (X_n, Y_n)\},$$

where the X_i's are independent and identically distributed as a random function X, defined on an interval I, and the Y_i's are generated by the regression model,

$$Y_i = a + \int_I b\, X_i + \epsilon_i.$$

Here, a is a constant, denoting the intercept in the model, and b is a square-integrable function on I, representing the slope function.

There has been substantial recent work on methods for estimating the slope function in linear regression for functional data analysis, typically by methods based on functional principal components. However, as in the case of more conventional, finite-dimensional regression, much of the practical interest in the slope centers on its application for the purpose of prediction, rather than on its significance in its own right. Thus, while there is an extensive literature on properties of $\hat{b}$, for example on convergence rates of $\hat{b}$ to b (see e.g. Ferraty and Vieu, 2000; Cuevas *et al.*, 2002; Cardot and Sarda, 2003; Hall and Horowitz, 2004), there is arguably a still greater need to understand the manner in which $\hat{b}$ should be constructed in order to optimize the prediction of $\int_I bx$, or of $a + \int_I bx$. This is the problem addressed in the present paper.

Estimation of b is intrinsically an infinite-dimensional problem. Therefore, unlike slope estimation in conventional finite-dimensional regression, it involves smoothing or regularization. The smoothing step is used to reduce dimension, and the extent to which this should be done depends on the use to which the estimator of b will be put, as well as on the smoothness of b. It is in this way that the problem of estimating $\int_I bx$ is quite different from that of estimating b. The operation of integration, in computing $\int_I \hat{b}\, x$ from $\hat{b}$, confers additional smoothness, with the result that if we smooth $\hat{b}$ optimally for estimating b then it will usually be oversmoothed for estimating $\int_I bx$.

We show that the problems of slope-function estimation, and of prediction from an estimator of the slope function, have very different characteristics. While the former is intrinsically nonparametric, the latter can be either nonparametric or semiparametric. In particular, the optimal mean-square convergence rate of predictors is n^{-1}, where n denotes sample size, if the predictand is a sufficiently smooth function. In other cases, convergence occurs at a polynomial rate that is strictly slower than n^{-1}. At the boundary between these two regimes, the mean-square convergence rate is less than n^{-1} by only a logarithmic factor. More generally, the rate of convergence of the predicted value of the mean response in

the regression model, given a particular value of the explanatory variable, is determined by a subtle interaction among the smoothness of the predictand, of the slope function in the model, and of the autocovariance function for the distribution of explanatory variables.

References

[1] T. Cai and P. Hall, *Prediction in functional linear regression*, Technical Report, available at *http://stat.wharton.upenn.edu/~tcai*.

Nonparametric Estimation of Expected Shortfalls

Song Xi Chen

The paper evaluates the properties of nonparametric estimators of the expected shortfall, an increasingly popular risk measure in financial risk management. It is found that the existing kernel estimator based on a single bandwidth does not offer variance reduction, which is surprising considering that kernel smoothing reduces the variance of estimators for the value at risk and the distribution function. We reformulate the kernel estimator such that two different bandwidths are employed in the kernel smoothing for the value at risk and the shortfall itself. We demonstrate by both theoretical analysis and simulation studies that the new kernel estimator achieves a variance reduction. The paper also covers the practical issues of bandwidth selection and standard error estimation.

Key Words: Kernel estimator; Risk Measures; Smoothing bandwidth; Value at Risk; Weak dependence.

References

[1] Artzner, P., Delbaen, F., Eber, J-M. and Heath, D. (1999), Coherent measures of risk, *Mathematical Finance*, **9**, 203–228.
[2] Billingsley, P. (1968), *Convergence of probability measures,* New York: Wiley.
[3] Bosq, D. (1998), *Nonparametric Statistics for Stochastic Processes*, Lecture Notes in Statistics, **110**. Heidelberg: Springer-Verlag.
[4] Cai, Z.-W. (2002), Regression quantiles for time series data, *Econometric Theory*, **18** 169–192.
[5] Cai, Z.-W. and Roussas, G. G. (1997), Smoothed estimate of quantile under association, *Statistics and Probability Letters*, **36**, 275–287.
[6] Cai, Z.-W. and Roussas, G. G. (1998), Efficient estimation of a distribution function under quadrant dependence, *Scandinavian Journal of Statistics*, **25**, 211–224.
[7] Chen, S. X. and Tang, C.Y. (2003). Nonparametric inference of Value at Risk for dependent financial returns. Research report, Department of Statistics, Iowa State University.
[8] Doukhan, P. (1994), *Mixing*. Lecture Notes in Statistics, **85**. Heielberg: Springer-Verlag.
[9] Duffie, D. and Pan, J. (1997), An overview of value at risk, *Journal of Derivative*, 7–49.
[10] Embrechts, P., Klueppelberg, C., and Mikosch, T. (1997), Modeling Extremal Events for Insurance and Finance, Berlin: Springer-Verlag.
[11] Falk, M. (1984), Relative deficiency of kernel type estimators of quantiles, *Annals of Statistics*, **12**, 261–268.
[12] Fan, J. and Gijbels, I. (1996), *Local Polynomial Modeling and Its Applications*. London: Chapman and Hall.

[13] Fan, J., Gu, J. and G. Zhou (2003), Semiparametric Estimation of Value-at-Risk, *Econometrics Journal*, to appear.
[14] Fan, J. and Yao, Q. (2003), *Nonlinear Time Series: Nonparametric and Parametric Methods*, New York: Springer-Verlag.
[15] Föllmer, H. and Schied, A. (2002), Convex measures of risk and trading constraints, *Finance and Stochastics* to appear.
[16] Frey, R. and McNeil, A. J. (2002), VaR and expected shortfall in portfolios of dependent credit risks: conceptual and practical insights, *Journal of Banking and Finance*, 26, 1317–1334.
[17] Gouriéroux, C., Scaillet, O. and Laurent, J.P. (2000), Sensitivity analysis of Values at Risk, *Journal of empirical finance*, **7**, 225–245.
[18] Jorion, P. (2001), *Value at Risk*, 2nd Edition, New York: McGraw-Hill.
[19] Masry, E. and Tjøstheim, D. (1995), " Nonparametric estimation and identification of nonlinear arch time series," *Econometric Theory*, **11**, 258–289.
[20] Reiss, R. D. and Thomas, M. (2001), *Statistical Analysis of Extreme Values: with applications to insurance, finance, hydrology, and other fields, 2nd ed.*, Boston: Birkhäuser Verlag.
[21] Rosenblatt, M. (1956), A central limit theorem and a the α-mixing condition, *Proc. Nat. Acad. Sc. U.S.A.*, **42**, 43–47.
[22] Scaillet, O. (2003a), Nonparametric estimation and sensitivity analysis of expected shortfall, *Mathematical Finance* to appear.
[23] Scaillet, O. (2003b), Nonparametric estimation of conditional expected shortfall, HEC Geneve mimeo.
[24] Sheather, S. J. and Marron, J. S. (1990), Kernel quantile estimators, *Journal of American Statistical Association*, **85**,410–416.
[25] Yokoyama, R. (1980), Moment bounds for stationary mixing sequences, *Probability Theory and Related Fields*, **52**, 45–87.
[26] Yoshihara, K. (1995), The Bahadur representation of sample quantiles for sequences of strongly mixing random variables, *Statistics and Probability Letters*, **24**, 299–304.

Dimension Reduction in the Model of Nonparametric Regression

Arnak Dalalyan
(joint work with A. Iouditski and V. Spokoiny)

We consider the model of nonparametric regression

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

with d-dimensional design $\{X_i\}_i \subset \mathbf{R}^d$ and additive i.i.d. Gaussian noise ξ_i. Assume that only the projection of $x \in \mathbf{R}^d$ on a m-dimensional subspace $\mathcal{I} \subset \mathbf{R}^d$ (with $m \ll d$) accounts for the fluctuations of f, that is $f(x) = g(R^T x)$ for a function $g : \mathbf{R}^m \to \mathbf{R}$ and a matrix $R \in \mathbf{R}^{d \times m}$ such that $R^T R = I_m$. Our goal is to recover the index subspace $\mathcal{I} = \mathrm{Im}(R) = \mathrm{Im}(\nabla f)$ when both g and R are unknown.

The method we propose for estimating $\mathcal{I}$ is a modification of the one presented in [1]. The main idea is to proceed in a recursive way: starting from a crude estimation of ∇f, get a general information about $\mathcal{I}$, then use this information to improve the estimator of ∇f, and so on. Thus there are two main issues. Firstly, find the best method to extract the information on $\mathcal{I}$ from the estimator of ∇f.

Secondly, exploit as well as possible the information on $\mathcal{I}$ to obtain an improved estimation of ∇f.

In [1], a PCA on the projections of the estimated gradient on some directions has been used to solve the first issue. In order to guarantee a good behavior of the method, the number of directions should be small relative to the sample size n. In our work, we propose a new method for recovering the index subspace from an estimator of the gradient which allows to make use a large number (polynomial in n) of projections. The resulting estimator is shown to be rate-optimal: it converges with the rate $n^{-1/2}$.

References

[1] M. Hristache, A. Juditsky, J. Polzehl and V. Spokoiny, *Structure adaptive approach for dimension reduction*, Ann. Statist. **29** (2001), 1537–1566.

Estimation of integrated volatility in continuous time financial models with applications to goodness-of-fit testing

Holger Dette

(joint work with Mark Podolskij and Mathias Vetter)

Properties of a specification test for the parametric form of the variance function in diffusion processes $dX_t = b(t, X_t)\,dt + \sigma(t, X_t)\,dW_t$ are discussed. The test is based on the estimation of certain integrals of the volatility function. If the volatility function does not depend on the variable x it is known that the corresponding statistics have an asymptotic normal distribution. However, most models of mathematical finance use a volatility function which depends on the state x. In this paper we prove that in the general case, where σ depends also on x the estimates of integrals of the volatility converge stably in law to random variables with a non-standard limit distribution. The limit distribution depends on the diffusion process X_t itself and we use this result to develop a bootstrap test for the parametric form of the volatility function, which is consistent in the general diffusion model.

References

[1] D.J. Aldous, G.K. Eagleson, *On mixing and stability of limit theorems*, Ann. of Prob. **6** (1978), 325–331.

[2] Y. Ait-Sahalia, *Testing continuous time models of the spot interest rate*, Rev. Financ. Stud. **9** (1996), 385-426.

[3] V. Corradi, H. White, *Specification tests for the variance of a diffusion*, J. Time Series **20** (1999), 253-270.

[4] H. Dette, C. von Lieres und Wilkau, *On a test for a parametric form of volatility in continuous time financial models*, Financ. Stoch. **7** (2003), 363-384.

Adaptive support vector machines

Sara A. van de Geer
(joint work with B. Tarigan)

We examine the problem of labeling a feature $X \in \mathbf{X}$ of an item. For example, the item may be a mushroom, the feature its physical characteristics, and the label Y classifies it as edible or not. We consider the case where Y is binary, say $Y \in \{-1, 1\}$. A *classifier* is a function $f : \mathbf{X} \to \mathbf{R}$. Associated with f is the classification rule: given X, predict Y by the sign of $f(X)$. Thus, a prediction error occurs if $Yf(X) < 0$.

Suppose (X, Y) are random variables with distribution P. The risk of a classifier f is now defined as

$$R(f) = P(Yf(X) < 0).$$

Let $\eta(X) = P(Y = 1|X)$ be the regression of Y on X. The classifier with minimal risk is *Bayes rule*

$$f^* = 2\mathrm{l}\{\eta > 1/2\} - 1.$$

A training set $\{(X_i, Y_i)\}_{i=1}^n$ is a sample of i.i.d. copies from (X, Y). Using this training set, we propose to estimate f^* in the following way. Let $\psi = (\psi_1, \ldots, \psi_m)^T : \mathbf{X} \to \mathbf{R}^m$ be some feature mapping. For $\alpha \in \mathbf{R}^m$, let f_α be the linear combination

$$f_\alpha = \sum_k \alpha_k \psi_k.$$

Let z_+ denote the positive part of $z \in \mathbf{R}$, and define the *hinge* loss function

$$l(z) = (1 - z)_+, \ z \in \mathbf{R}.$$

The support vector machine (SVM) empirical loss in f is now

$$L_n(f) = \frac{1}{n} \sum_{i=1}^n l(Yf(X_i)).$$

The ℓ_1 penalized SVM estimator is defined as

$$\hat{f}_n = \arg \min_{|f_\alpha| \le K/2} L_n(f_\alpha) + \lambda_n \sum_k |\alpha_k|.$$

Here, λ_n is a *smoothing parameter*, and $K \ge 2$ is a given constant.

We prove an oracle inequality for the SVM *excess risk* $L(\hat{f}_n) - L(f^*)$ of this estimator. Here $L(f) = E(1 - Yf(X))_+$ is the SVM theoretical loss. Under conditions A, B and C below, the excess risk is not much larger than the excess risk of an *oracle*, which trades off non-sparseness in the number of non-zero coefficients (the estimation error) against the approximation error of a sparse representation.

Conditions A, B and C are as follows. Let Q be the distribution of X and $\|\cdot\|_p$ be the $L_p(Q)$ norm ($p \ge 1$).

Condition A *For some constants $\kappa \ge 1$ and $\sigma > 0$, we have for all $|f| \le K/2$*

$$L(f) - L(f^*) \ge \|f - f^*\|_1^\kappa / \sigma^\kappa.$$

Condition B *The smallest eigenvalue* $\lambda^2_{\min}$ *of* $\Sigma = \int \psi\psi^T dQ$ *is non-zero.*
Condition C *We have*

- $m \le n^D$, *where* $D \ge 1$,
- Q *has density* q *with respect to some given* σ*-finite measure* μ, *and* $q \le c_q$, *where* $c_q \ge 1$ *is given,*
- $\int \psi_k^2 d\mu \le 1$ *and* $|\psi_k| \le \sqrt{n/\log n}$ *for all* k.

For a vector $\alpha \in \mathbf{R}^m$, let its number of $\alpha_k \neq 0$ be denoted by $N(\alpha)$. The "estimation error" $V_n(N)$ is roughly speaking the error due to the variability in the sample, of an estimator using only a given set of N nonzero coefficients. It is defined as

$$V_n(N) = 2\delta^{\frac{2}{2\kappa-1}} \left[4\sigma\lambda_n^2/\lambda^2_{\min} N\right]^{\frac{\kappa}{2\kappa-1}},$$

where $0 < \delta \le 1/2$ is fixed, but otherwise arbitrary. The best trade off between estimation error and approximation error would yield excess risk

$$\epsilon_n = \inf\{V_n(N(\alpha)) + L(f_\alpha) - L(f_*) : \ |f_\alpha| \le K/2\}.$$

Theorem 1 *Assume conditions A, B and C are met. Take*

$$\lambda_n = cc_q DK^2\sqrt{\log n/n},$$

with c *an appropriate universal constant. Then for a universal constant* c_0,

$$\mathbf{P}(L(\hat{f}_n) - L(f^*) > (1+4\delta)\epsilon_n) \le c_0 \exp[-K^2 \log n/2].$$

Example. Suppose $\mathbf{X} \subset [0,1]^2$ and that f^* is a boundary fragment

$$f^*(u,v) = 2\mathrm{l}\{v < g^*(u)\} - 1, \ (u,v) \in [0,1]^2.$$

Let $2^L \le \sqrt{n/\log n}$ and $\{\psi_{j,l} : \ j = 1,\ldots,2^{l-1}, \ l = 1,\ldots,L\}$ be the Haar system on $[0,1]$ up to resolution level L. We use the expansion

$$f_\alpha(u,v) = \sum_i \sum_j \sum_k \sum_l \alpha_{i,j,k,l}\psi_{i,k}(u)\psi_{j,l}(v).$$

Suppose now that Q is uniform on the grid

$$\mathbf{X} = \{(k2^{-L}, l2^{-L}) : \ k,l \in \{1,2,\ldots,2^L\}\}.$$

Assume moreover that for some $\gamma \ge 0$,

$$1/2 \le \frac{|\eta(u,v) - \eta(u,g^*(u))|}{|v - g^*(u)|^\gamma} \le 1, \ \forall \ (u,v) \in \mathbf{X}.$$

Finally, assume that for some $s > 1/(1+\gamma)$,

$$\sum_{k=1}^{2^L} |g^*(k2^{-L}) - g^*((k-1)2^{-L})|^{1/s} \le 1.$$

Then we have $\epsilon_n = O\left(\log^3 n/n\right)^\rho$, where $\rho = \kappa s/((2\kappa-1)s+1)$, and $\kappa = 1+\gamma$.

The result of Theorem 1 can be compared with the kernel SVM's which are often used in literature (see for example Schölkopf and Smola (2002)). The eigenvalues of kernels generally decrease very fast. We therefore briefly sketch a result under a different set of conditions, more adapted a kernel setup.

Let us define $I(\alpha) = \sum_k |\alpha_k|$. Let

$$\mathbf{C} = \{f_\alpha : I(\alpha) \leq 1,\ |f_\alpha| \leq K/2\}.$$

Let $H(\cdot, \mathbf{C}, \nu)$ be the entropy of $\mathbf{C} \subset L_2(\nu)$, where ν is a probability measure.

Condition D *There exists constants $h > 0$, and $s_0 \geq 1$ such that for all probability measures ν,*

$$H(\epsilon, \mathbf{C}, \nu) \leq \frac{1}{h}\epsilon^{-\frac{1}{2s_0}},\ \forall\ \epsilon > 0.$$

One may think of s_0 as the smoothness of the kernel, and h as the "width".

We slightly extend our definition of the ℓ_1 penalized SVM estimator, incorporating the possibility of penalty on the smoothing parameter λ. Let

$$\hat{f}_n = \arg \min_{|f_\alpha| \leq K/2} \left\{ \min_{\lambda > 0} \left\{ L_n(f_\alpha) + \lambda I(\alpha) + b_0 \lambda^{-\frac{\gamma_0}{s_0 - \gamma_0}} \right\} \right\}.$$

Here $\gamma_0 > s_0$ and $b_0 > 0$ are given constants.. Under conditions A and D, there exist a constant $\tilde{\lambda}$ depending on γ_0, s_0 and b_0, and a constant C depending on γ_0 s_0 and K, such that for all $|f_\alpha| \leq K/2$,

$$\mathbf{E}\left[L(\hat{f}_n) - L(f^*) + \tilde{\lambda}^{\gamma_0} I^{\gamma_0/s_0}(\hat{f}_n) \right]$$

$$\leq C \left[\left(\sqrt{nh}\tilde{\lambda} \right)^{-r} + \tilde{\lambda}^{\gamma_0} I^{\gamma_0/s_0}(f_\alpha) + L(f_\alpha) - L(f^*) \right].$$

Here,

$$r = \frac{2\kappa s_0 \gamma_0}{2(\kappa - 1)s_0\gamma_0 + \gamma_0 - 2\kappa s_0}.$$

The result can be optimized by taking the parameters γ_0, h and $\tilde{\lambda}$ appropriately. However, the optimal choice depends on κ which is generally unknown. On the other hand, the conditions of Theorem 1 allow one to adapt to unknown κ.

References

[1] B. Schölkopf and A. Smola, *Learning with Kernels*, (2002), MIT Press, Cambridge.

[2] B. Tarigan and S.A. van de Geer, *Adaptivity of support vector machines with ℓ_1 penalty*, Techn. Report**MI 2004-14** (2004), University of Leiden.

Recovering edges of an image from noisy tomographic data

Alexander Goldenshluger
(joint work with Vladimir Spokoiny)

In this paper we address the problem of recovering edges of an image from noisy tomographic data. The original image is modeled by function f defined on the unit disc $B^2(o,1) \subset R^2$. Assume that f is smooth apart from a discontinuity jump along a smooth curve. The problem of edge recovery from tomographic data is to estimate the discontinuity curve from noisy measurements of line integrals of f.

The problem of edge detection arises in numerious imaging applications. For example, images with discontinuitites along edges are ubiquitous in medical applications; here edges bring important information about body regions with different levels of metabolic activity. Thus edge recovery is an important step in processing tomographic images.

Although various methods and proposals are widely used in practice, [see, e.g., Faridani et. al. (1992), Katsevich, Ramm (1995), Srinivasa et. al (1995)] theoretical limitations in the problem of edge detection from the Radon data are yet to be understood. What is the best attainable accuracy in recovering edges from noisy observations of projections? Which methods can achieve this optimal performance? The goal of the present paper is to provide a theoretical perspective on these questions and to develop easily implemented nearly–optimal algorithm for edge recovery in tomographic images.

In this paper we consider the white noise Radon tranform model, and our focus is on direct recovery of the edge rather than on estimating the whole image. We assume that the edge can be represented as the boundary of a convex set, and propose a method for estimating support function of this set. Then the boundary is recovered as the envelope of the estimated supporting lines. We analyze theoretical properties of the proposed estimation scheme and show that it is nearly optimal in order in the sense of the rates of convergence.

References

[1] Faridani, A., Ritman, E., and Smith, K. (1992). Local tomography. *SIAM J. Appl. Math.* **52**, 459–484.

[2] Katsevich, A. and Ramm, A. (1995). New methods for finding values of the jumps of a function from its local tomographic data. *Inverse Problems* **11**, 1005–1023.

[3] Srinivasa, N., Ramakrishnan, K. R. and Rajgopal, K. (1992). Edge detection from projections. *IEEE Trans. Med. Imaging* **11**, 76-80.

Primal-Dual Algorithms of Stochastic Approximation

A. Juditsky
(joint work with A. Nemirovski, and Yu. Nesterov)

The problem of convex stochastic optimization consists to find a solution to

$$\min f(x), \quad \text{subject to } x \in G.$$

Here $G \subseteq \mathbb{R}^M$ is a closed convex set and f is a convex function. To find the minimizer the stochastic optimization method is allowed to use the noisy observations of the subgradient of f at the search points: $f'(x_i) + e_i$, $i = 1, 2, ...$

We are specifically interested in the case of large scale, i.e. of high dimension M of the problem, and of "simple set" G (such as a simples or a hyperoctahedre). An original version of the subgradient descent – the *mirror descent* algorithm – has been proposed for that type of problems in [1]. It has been proved in [1] that, for instance, in the case when the set G is such that $\|x\|_1 \leq R$ for any $x \in G$, the approximate solution $\bar{x}_n$ after n steps of mirror descent process satisfies:

$$Ef(\bar{x}_n) - f(x^*) = O\left(RL\sqrt{\frac{(\ln n + \ln M)}{n}}\right). \tag{1}$$

Where L is the "intensity" of the subgradient observation:

$$L^2 = \sup_{x \in G} \|Ef'(x) + e\|_\infty^2,$$

and $E(\cdot)$ stands for the expectation with respect to the noise distribution. It has been also shown in [1] that this performance rate cannot be significantly improved.

The subject of the presented work is to study the properties of a family of primal-dual stochastic approximation algorithms, of which the mirror descent of [1] and [2] is a particular sample. These methods attain the performance bound [1], but due to the improved structure they can be better tuned to satisfy particular requirements of statistical applications.

We present two examples of using the proposed methods in the statistical problems of functional aggregation and excess risk minimization.

References

[1] A. Nemirovski, D. Yudin, *Problem complexity and method efficiency in optimization*, Wiley-Interscience Series in Discrete Mathematics, Chichester etc.: John Wiley & Sons. XV, 1983.

[2] A. Juditsky, A. Nemirovski, *Functional aggregation for nonparametric regression*, Ann. Statist. 28 (2000), no. 3, 681-712.

A Dynamic Semiparametric Factor Model for Implied Volatility String Dynamics

Wolfgang K. Härdle
(joint work with Matthias R. Fengler, Enno Mammen)

Successful trading, hedging and risk managing of option portfolios crucially depends on the accuracy of the underlying pricing models. Departing from the pioneering foundations of option theory laid by Black and Scholes (1973), Merton (1973) and Harrison and Kreps (1979), new valuation approaches are continuously developed and existing models are refined. However, despite these pervasive developments, the model of Black and Scholes (1973) remains the pivot in modern financial theory and the benchmark for sophisticated models, be it from a theoretical or practical point of view.

The crucial parameter in option valuation by BS is the market volatility. Since it is unknown, one studies *implied* volatility, which is derived by inverting the BS formula for a cross section of options with different strikes and maturities traded at the same point in time. Implied volatilities display a remarkable curvature across the strike dimension, and – albeit to a lesser degree – a term structure across time to maturity. For a given time to maturity the phenomenon is called *smile* or *smirk*. This dependence given by the mapping $\hat{\sigma}_t : (\kappa, \tau) \to \hat{\sigma}_t(\kappa, \tau)$, where κ denotes the strike dimension scaled in moneyness and τ time to maturity, is called *implied volatility surface* (IVS). The index t denotes time-dependence. Apparently, it is in contrast with the BS framework in which volatility is assumed to be a constant across strikes, time to maturity and also time.

There is a considerable amount of literature which aims at reconciling this empirical antagonism with financial theory. Generally speaking, this can be achieved by including another degree of freedom into option pricing models: well-known examples are stochastic volatility 0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 Time to maturity models, (Hull and White; 1987; Stein and Stein; 1991; Heston; 1993), models with jump diffusions, Bates (1996a,b), or models building on general Lévy processes, e.g. based on the inverse Gaussian, Barndorff-Nielsen (1997), and generalized hyperbolic distribution, Eberlein and Prause (2002). These approaches capture the smile and term structure phenomena and the complexity of its dynamics to some extent, Das and Sundaram (1999); Bergomi (2004).

Nevertheless, the BS model and the IVS enjoy much popularity. Partly, this may be due to the fact that the IVS is derived from instantaneous option prices, and is thus a widely accepted state variable reflecting current market sentiments, Bakshi et al. (2000). More importantly, however, the IVS plays a decisive role in trading: market makers at plain vanilla desks continuously monitor and update the IVS they trade on; and exotic derivatives trader calibrate their pricing engines with an estimate of the IVS. This is particularly obvious for the pricing systems relying on the local volatility models. Initially developed by Dupire (1994) and Derman and Kani (1994), they are in wide-spread use in form of the highly efficient implementations by Andersen and Brotherton-Ratcliffe (1997) and Dempster and

Richards (2000). Thus, refined statistical model building of the IVS determines vitally the accuracy of applications in trading and risk-management.

In modelling the IVS one faces two main challenges. First, the data design is degenerated: due to institutional conventions, observations of the IVS occur only for a small number of maturities such as one, two, three, six, nine, twelve, 18, and 24 months to expiry on the date of issue. Consequently, implied volatilities appear in a row like pearls strung on a necklace, or, in short: as 'strings'. Options belonging to the same string have a common time to maturity. As time passes, the strings move through the maturity axis towards expiry while changing levels and shape in a random fashion. Second, also in the moneyness dimension, the observation grid does not cover the desired estimation grid at any point in time. Thus, even when the data sets are huge, for a large number of cases implied volatility observations are missing for certain sub-regions of the desired estimation grid. This is particularly virulent when transaction based data are used. However, despite their appearance as strings, implied volatilities are thought as being the observed structure of a smooth surface. This is because in practice one needs to price and hedge OTC options whose expiry dates do not coincide with the expiry dates of the options that are traded at the futures exchange.

For the semi- or nonparametric approximations to the IVS that recently have been promoted by Aït-Sahalia and Lo (1998); Rosenberg (2000); Aït-Sahalia et al. (2001b); Cont and da Fonseca (2002); Fengler et al. (2003); Fengler and Wang (2003), this design may pose difficulties. The fit appears very rough, and there are huge holes in the surface, since the bandwidths are too small to bridge the gaps between the maturity strings. In order to remedy this deficiency one would need to strongly increase the bandwidths which may induce a large bias. Moreover, since the design is time-varying, the bandwidths would also need to be adjusted anew for each trading day, which complicates daily applications. Parametric models, e.g. as in Shimko (1993), Ané and Geman (1999), and Brockhaus et al. (2000, Chap. 2) among others, are less affected by these data limitations, but appear to offer too little functional flexibility to capture the salient features of IVS patterns. Thus, parametric estimates may as well be biased.

We propose a dynamic semiparametric factor model (DSFM), which approximates the IVS in a finite dimensional function space. The key feature is that we only fit in the local neighborhood of the design points. Our approach is a combination of methods from functional principal component analysis and backfitting techniques for additive models.

Let us denote the (log)-implied volatility by $Y_{i,j}$, where the index i is the number of the day ($i = 1, ..., I$), and $j = 1, ..., J_i$ is an intra-day numbering of the option traded on day i. The observations $Y_{i,j}$ are regressed on two-dimensional covariables $X_{i,j}$ that contain moneyness $\kappa_{i,j}$ and maturity $\tau_{i,j}$. Moneyness is defined as $\kappa_{i,j} \stackrel{\text{def}}{=} K_{i,j}/F_{t,i,j}$, i.e. strike $K_{i,j}$ divided by the underlying futures price $F_{t,i,j}$ at time $t_{i,j}$. We also considered the one-dimensional case in which $X_{i,j} = \kappa_{i,j}$. However, since modelling the entire surface is more interesting, we will present results for this case only. The DSFM is given by:

$$m_0(X_{i,j}) + \sum_{l=1}^{L} \beta_{i,l} m_l(X_{i,j}) \ , \tag{1}$$

where m_l are smooth basis functions ($l = 0, \ldots, L$). The IVS is approximated by a weighted sum of smooth functions m_l with weights $\beta_{i,l}$ depending on time i. The factor loading $\beta_i \stackrel{\text{def}}{=} (\beta_{i,1}, \ldots \beta_{i,L})^\top$ forms an unobserved multivariate time series. By fitting model (1), to the implied volatility strings we obtain approximations $\mathbf{w}_i$. We argue that the VAR estimation based on $\mathbf{w}_i$ is asymptotically equivalent to estimation based on the unobserved β_i. A justification for this is given in Fengler et al. (2004) where the relations to Kalman filtering are discussed.

The model is found to have an approximate 10% better performance than the typical naïve trader models. Finally, we devise a generalized vega-hedging strategy for exotic options that are priced in the local volatility framework. The generalized vega-hedging extends the usual approaches employed in the local volatility framework.

References

[1] Alexander, C., *Principles of the Skew*, RISK, **14** (2001), 29–32.

[2] Avellaneda, M. and Zhu, Y., *An E-ARCH Model for the Term-Structure of Implied Volatility of FX Options*, Applied Mathematical Finance, **4**, (1997), 81–100.

[3] Baksi, G., Cao, C. and Chen, Z., *Do call and underlying prices always move in the same direction?*, Review of Financial Studies, **13**, (2000), 549–584.

[4] Black, F. and Scholes, M., *The pricing of options and corporate liabilities*, Journal of Political Economy, **81**, 1973, 637–654.

[5] Cont, R. and da Fonseca, J., *The Dynamics of Implied Volatility Surfaces*, Quantitative Finance, **2**, (2002), 45–602.

[6] Fengler, M., Härdle, W. and Villa, C., *The dynamics of implied volatilities: A common principle components approach*, Review of Derivatives Research, **6**, (2003), 179–202.

[7] Fengler, M., Härdle, W. and Mammen, E., *Semiparametric State Space Factor Models*, CASE Discusion Paper, Humboldt-Universität zu Berlin, (2005).

[8] Fengler, M., Härdle, W. and Schmidt, P., *Common Factors Govering VDAX Movements and the Maximum Loss*, Journal of Financial Markets and Portfolio Management, **16**, (2002), 16–29.

[9] Härdle, W. and Hlavka, Z., *Dynamics of State Price Densities*, SFB Discussion Paper, (2005).

[10] Härdle, W. and Yatchew, A., *Dynamic Nonparametric State price Density Estimation using Constrained least Squares nd the Bootstrap*, Discussion Paper 1, Sonderforschungsbereich 373, Humboldt-Universität zu Berlin, (2001).

[11] Hafner, R. and Wallmeier, M., *The Dynamics of DAX Implied Volatilities*, International Quarterly Journal of Finance, **1**, (2001), 1–27.

[12] Härdle, W., Klinke, S. and Müller, M., *XploRe - Learning Guide*, Heildelberg: Springer Verlag, (2000).

[13] Ramsay, J. O. and Silverman, B. W., *Functional Data Analysis*, Springer, (1997)

[14] Skiadopoulos, G., Hodges, S. and Clewlow, L., *The Dynamics of S&P 500 Implied Volatility Surface*, Review of Derivatives Research, **3**, (1999), 263–282.

Efficient estimation of additive models with link function

Enno Mammen
(joint work with Joel Horowitz)

In this talk efficient estimation of an additive nonparametric component m_1 is discussed for a generalized additive model

$$E[Y^i|X^i] = G[m_0 + m_1(X_1^i) + ... + m_D(X_D^i)]$$

for i.i.d. tuples $X^i = (X_1^i, ..., X_D^i)$, Y^i. We discuss a two-step procedure. In a first step an orthogonal series estimate $\widetilde{m}_0, ..., \widetilde{m}_D$ is fitted for all components of the additive model. In a second step a local linear fit is used for m_1: define $\bar{m}_1(x_1)$ as $\widehat{a}$ where $(\widehat{a}, \widehat{b})$ is the minimizer of

$$\sum_{i=1}^{n} \left\{Y^i - G[\widetilde{m}_0 + a + b(X_1^i - x_1) + \widetilde{m}_1(X_1^i) + ... + \widetilde{m}_D(X_D^i)]\right\}^2 K_h(X_1^i - x_1).$$

We propose to use a one-step Newton approximation to $\bar{m}_1(x_1)$ as estimate of m_1. Let us denote this estimate by $\widehat{m}_1$. This estimate can be compared to an estimate in the above model where the additive functions $m_2, ..., m_D$ are known ("oracle model"). A local linear estimate in this model could be defined as $\widehat{m}_1^{oracle}$ where $\widehat{m}_1^{oracle}(x_1)$ is a one-step Newton approximation to $\widehat{a}$ and where now $(\widehat{a}, \widehat{b})$ is the minimizer of

$$\sum_{i=1}^{n} \left\{Y^i - G[m_0 + a + b(X_1^i - x_1) + m_1(X_1^i) + ... + m_D(X_D^i)]\right\}^2 K_h(X_1^i - x_1).$$

Our main result is that our estimate $\widehat{m}_1$ and the oracle estimate $\widehat{m}_1^{oracle}(x_1)$ are asymptotically equivalent. This can be interpreted as an efficiency result because our estimate is doing as well as if the other components would be known.

In a second part of the talk a general result is shown for the additive model without link function:

$$E[Y^i|X^i = m_0 + m_1(X_1^i) + ... + m_D(X_D^i).$$

For a general class of smoothing estimates $\widehat{m}_1^{oracle}$ that are available in the oracle model we show the following result. There exists an estimate $\widehat{m}_1$ in the additive model that is asymptotically equivalent to $\widehat{m}_1^{oracle}$. The class of smoothing estimates includes local polynomials, regression splines, smoothing splines and orthogonal series estimates.

Functional regression for sparse and noisy data

Hans-Georg Müller
(joint work with Jane-Ling Wang, Fang Yao)

Functional linear regression models can be classified according to whether predictor and responses are functions or vectors. Each of these cases gives rise to different

procedures and analyses. We consider functional regression models where both predictor X and response Y is a random function in L^2, according to

$$E[Y(t)|X] = \alpha(t) + \int \beta(s,t)X(s)ds.$$

While this model has been well studied in the case where entire trajectories are observed without noise (compare [1]), we extend its applicability to the case where the trajectories are sampled on an irregular and sparse time grid and where the measurements are corrupted with additional noise. The number of measurements and their location per trajectory are assumed to be random. Under smoothness assumptions, Gaussian assumptions on processes and errors, and assuming a positive probability to sample each predictor and response trajectory at more than one point, we obtain asymptotic consistency with rates of convergence for estimates of the regression parameter function β and the prediction of individual response trajectories from sparse and noisy measurements of the predictor trajectories. The derivations make use of classical perturbation theory for Hilbert-Schmidt operators and of results in [2]. The method uses algorithms discussed in [3]. The proposed functional regression method is illustrated with data from biomedical longitudinal studies.

References

[1] Ramsay, J. & Dalzell, C.J., *Some tools for functional data analysis*, Journal of the Royal Statistical Society, Series B **53** (1991), 539–572.

[2] Yao, F., Müller, H. G. & Wang, J. L., *Functional data analysis for sparse longitudinal data*, Journal of the American Statistical Association (2005).

[3] Yao, F., Müller, H. G., Clifford, A. J., Dueker, S. R., Follett, J., Lin, Y., Buchholz, B. A., Vogel, J. S., *Shrinkage estimation for functional principal component scores with application to the population kinetics of plasma folate*, Biometrics **59** (2003), 676–685.

Subspace Clustering and Exploratory Analysis of Gene Expression Data

Andrew Nobel

(joint work with Xing Sun)

Exploratory analysis of gene expression data typically begins by clustering the rows and columns of the experimental data matrix, yielding a division of the heat map into non overlapping cells. Cells colored bright red (large values) or bright green (small values) are viewed as representing significant sample-gene interactions, and are subject to further analysis. Subspace clustering, also known as bi-clustering, looks directly for sample variable interactions satisfying a pre-specified criterion. The resulting clusters can overlap, and need not cover the heat map.

The talk will begin with an overview of subspace clustering in the context of gene expression data. The remainder of the talk will be devoted to an overview of ongoing work on the analysis of subspace clustering, beginning with several results on significance tests (p-values) for subspace clusters having a given size and aspect ratio. Implications of these results for the noise sensitivity of algorithms using hereditary clustering criteria will be discussed. Applications of subspace clustering to classification, and some efforts to refine our existing significance analysis, will also be presented.

Spatially Adaptive Smoothing: A Propagation-Separation Approach

Jörg Polzehl
(joint work with Vladimir Spokoiny)

Edges and homogeneous regions are often the most interesting structures in images. Image enhancement should make such structures more appealing. Polzehl and Spokoiny (2000) offered a new adaptive method of nonparametric estimation, *Adaptive Weights Smoothing (AWS)*, in the context of image denoising that exactly focused on this goal. The method employed the structural assumption of a valid local constant approximation of the image in a local vicinity of each pixel and tried to adaptively recover this vicinity from the data. We now revise and extend this method to handle more general situations and to improve on the quality of the original procedure. Generalizations include exponential family models, e.g. binary or Poisson regression, and piecewise smooth models.

Let us assume the image to be given as gray values Y_i in pixel (voxel) X_i arranged on a two or three dimensional grid. The problem of image enhancement can then be formulated in form of a varying coefficient model

$$Y_i \sim \mathbf{P}_{f(X_i)} \tag{1}$$

with parameter $f(X_i)$ depending on location X_i. The function $f(x)$ describes the structure of the image and is the quantity we are interested to estimate.

Traditional approaches to this problem are e.g. nonlinear diffusion, wavelets and Markov random field methods. Popular nonparametric methods for varying coefficient models include kernel estimates and local polynomials.The latter methods focus on a localization of the model using a kernel function K and a bandwidth h to assign weights $w_j(x) = K((X_j - x)/h)$ to each observation (X_j, Y_j) when estimating $f(x)$. These weights determine the local model at point x. Parameter estimates are then obtained by weighted (localized) likelihood or least squares yielding e.g. kernel estimates in the form

$$\hat{f}(x) = \sum_j^n w_j(x) Y_j / \sum_j^n w_j(x) \tag{2}$$

The resulting estimates are smooth, with smoothness carried over from the kernel K. This restricts there use in image enhancement.

We employ a related but more general approach. Instead of prespecifying the weights $w_j(x)$ defining the local model in x we allow them to depend on the unknown image structure. We then attempt to recover both the unknown image structure and the corresponding optimal local models (weights) from the data.

We now describe the basic idea in its simplest case where $Y_i = f(X_i) + \varepsilon_i$ with $\mathbf{E}\varepsilon_i = 0$ and $\mathbf{D}\varepsilon_i = \sigma^2$. Let us assume that in each point x there exists a local neighborhood $U(x)$ such that the function $f(x)$ can be well approximated in $U(x)$ by a constant. If we knew this neighborhood $U(x)$ by an oracle we would define local weights as $w_j(x) = I_{X_j \in U(x)}$ and use these weights to estimate $f(x)$ by (2). On the other hand, if we have good estimates $\hat{f}(x)$ we can use this information to infer on the sets $U(x)$ by testing the hypothesis $H : f(X_j) = f(x)$. A weight $w_j(x)$ can be assigned based on the value of a test statistic $T_j(x)$, assigning zero weights if $\hat{f}(X_j)$ and $\hat{f}(x)$ are significantly different. This provides us with a weight matrix $W(x) = Diag(w_1(x), \ldots, w_n(x))$ that determines a local model in x.

We utilize both steps in an iterative procedure. We start with a very local model in each point X_i given by weights $w_j^{(0)}(X_i) = w_{ij}^{(0)} = K_{\mathrm{loc}}(l_{ij}^{(0)})$ with $\boldsymbol{l}_{ij}^{(0)} = |X_i - X_j|/h^{(0)}$. The initial bandwidth $h^{(0)}$ is chosen very small. K_{loc} is a kernel function supported on $[-1, 1]$, i.e. weights vanish outside a ball $U_i^{(0)}$ of radius $h^{(0)}$ centered in X_i. We then iterate two steps, estimation of $f(x)$ and refining the local models. New weights are generated as $w_{ij}^{(k+1)} = K_{\mathrm{loc}}(\boldsymbol{l}_{ij}^{(k)}) K_{\mathrm{st}}(\boldsymbol{s}_{ij}^{(k)})$ with $\boldsymbol{l}_{ij}^{(k)} = |X_i - X_j|/h^{(k)}$ and $\boldsymbol{s}_{ij}^{(k)} = T_{ij}^{(k)}/\lambda$ increasing the bandwidth h with each iteration k. We use $T_{ij}^{(k)} = N_i^{(k)}(\hat{f}^{(k)}(X_i) - \hat{f}^{(k)}(X_j))^2/(2\sigma^2)$ with $N_i = \sum_j w_{ij}$ as a test statistic. The penalty $\boldsymbol{s}_{ij}^{(k)}$ effectively measures the statistical difference of the current estimates in X_i and X_j. Due to this term we have propagation of weights within homogeneous regions and separation of regions, or edge preservation, if the estimated parameters become significantly different.

For large bandwidths this procedure may introduce an estimation bias in case of $f(x)$ changing smoothly with parameters, i.e. when our local constant assumption is violated. Therefore we introduce a kind of memory in the procedure, that ensures that the quality of estimation will not be lost with iterations. This basically means that we compare the new estimate $\tilde{f}^{(k)}$ with the previous estimate $\tilde{f}^{(k-1)}$ to define a memory parameter $\eta_i = K_{\mathrm{mem}}(\boldsymbol{m}_i^{(k)})$ with $\boldsymbol{m}_i^{(k)} = (2\sigma^2\tau)^{-1} \sum_j K_{\mathrm{loc}}(\boldsymbol{l}_{ij}^{(k)})(\hat{f}^{(k)}(X_i) - \hat{f}^{(k)}(X_j))^2$. This leads to the definition $\hat{f}^{(k)}(X_i) = \eta \tilde{f}^{(k)}(X_i) + (1-\eta)\hat{f}^{(k-1)}(X_i)$. The resulting algorithm reads as follows

- **Initialization:** Set $h^{(1)}$ (unit: distance between adjacent pixel), $k = 1$, $N_i^{(1)} = \sum_j w_{ij}^{(1)}$, $S_i^{(1)} = \sum_j w_{ij}^{(1)} Y_j$ with $w_{ij}^{(1)} = K_{\mathrm{loc}}(\boldsymbol{l}_{ij}^{(1)})$.
- **Adaptation (Computing the weights):** For every pair i, j, compute the penalties

$$\begin{aligned} \boldsymbol{l}_{ij}^{(k)} &= |X_i - X_j|/h^{(k)}, \\ \boldsymbol{s}_{ij}^{(k)} &= \lambda^{-1} T_{ij}^{(k)} = \lambda^{-1} N_i^{(k-1)} (\hat{f}^{(k)}(X_i) - \hat{f}^{(k)}(X_j))^2/\sigma^2. \end{aligned}$$

Now compute the weights $w_{ij}^{(k)}$ as $w_{ij}^{(k)} = K_{\mathrm{loc}}\big(\boldsymbol{l}_{ij}^{(k)}\big) K_{\mathrm{st}}\big(\boldsymbol{s}_{ij}^{(k)}\big)$. Define $W_i^{(k)} = \mathrm{diag}\{w_{i1}^{(k)}, \ldots, w_{in}^{(k)}\}$.

- **Local estimation:** compute new local MLE estimates $\hat{f}_{X_i}^{(k)}$ of f_{X_i}

$$\tilde{f}_{X_i}^{(k)} = S_i^{(k)} / \tilde{N}_i^{(k)} \qquad \text{with} \qquad \tilde{N}_i^{(k)} = \sum_j w_{ij}^{(k)}, \quad S_i^{(k)} = \sum_j w_{ij}^{(k)} Y_l \,.$$

- **Adaptive control (memory):** compute the memory parameter as $\eta_i = K_{\mathrm{mem}}(\boldsymbol{m}_i^{(k)})$). Define

$$\hat{f}_{X_i}^{(k)} = \eta \tilde{f}_{X_i}^{(k)} + (1-\eta) \hat{f}_{X_i}^{(k-1)} \quad \text{and} \quad N_i^{(k)} = \eta \tilde{N}_i^{(k)} + (1-\eta) N_i^{(k-1)}$$

- **Stopping:** Stop if $h^{(k)} \geq h_{max}$, otherwise set $h^k = a_h h^{k-1}$, increase k by 1 and continue with the adaptation step.

This basic procedure can be modified to allow for gray values Y_i following a distribution $\mathbf{P}_{f(X_i)}$ from an exponential family replacing the test statistics T_{ij} by $T_{ij} = N_i Q(\hat{f}(X_i), \hat{f}(X_j))$ with $Q(\theta, \theta')$ denoting the Kullback-Leibler distance between $\mathbf{P}_\theta$ and $\mathbf{P}_{\theta'}$.

Local polynomial models can be handled assuming a local model

$$f(x) = \theta(X_i) \Psi(x - X_i)$$

with functions $\Psi(x - X_i)$ forming a polynomial basis centered in X_i. Here we assume that in each point x there exists a local neighborhood $U(x)$ such that over $U(x)$ the vector of parameters $\theta(x)$ can be well approximated by a constant, see Polzehl and Spokoiny (2004).

The proposed procedure involves several parameters. The most important one is the scale parameter λ in the statistical penalty. The special case $\lambda = \infty$ simply leads to a kernel estimate with bandwidth h_{max}. We propose to chose λ as the smallest value satisfying a propagation condition. We require that if the local assumption is valid globally then with high probability the final estimate for $h_{max} = \infty$ in every point coincides with the global estimate. The value λ provided by this condition does not depend on the unknown function $f(x)$ and can therefore be approximately found by simulations. P or the number of parameters in a polynomial model.

The second parameter of interest is the maximal bandwidth h_{max} which controls both numerical complexity of the algorithm and smoothness within homogeneous regions. The scale parameter τ in $\boldsymbol{m}_i$ can also be chosen to meet the propagation condition.

We propose a new adaptive smoothing procedure with interesting properties for applications in image analysis. For a detailed description and theoretical results including rate optimality we refer to Polzehl and Spokoiny (2002, 2004). For examples illustrating the performance of the procedures we refer to [2], [3] and to the material provided on our webside http://www.wias-berlin.de/project-areas/stat/projects/adaptive-image-processing.html . An implementation of the

algorithm will be available from http://cran.r-project.org/ with the next version of the aws-package for R.

References

[1] J. Polzehl and V. Spokoiny, *Adaptive weights smoothing with applications to image segmentation*, J. of Royal Stat. Soc., Series **B**, **62** (2000), 135–354.

[2] J. Polzehl and V. Spokoiny, *Local likelihood modeling by adaptive weights smoothing.*, WIAS-Preprint 787, 2002.

[3] J. Polzehl and V. Spokoiny, *Spatially adaptive regression estimation: propagation-separation approach*, WIAS-Preprint , 2004.

Nonlinear methods for linear inverse problems with error in the operator

Markus Reiss

(joint work with Marc Hoffmann)

In [3] we consider nonlinear estimation methods for statistical inverse problems in the case where the operator is not exactly known. For a canonical formulation a Gaussian operator white noise framework is developed. For some unknown multivariate function f and compact operator K the observation is given by

$$g_\varepsilon = Kf + \varepsilon \dot{W}, \quad K_\delta = K + \delta \dot{B},$$

where $\dot{W}$ is an L^2-white noise and $\dot{B}$ is an operator white noise, characterized by the fact that in an orthonormal basis the infinite matrix representation consists of i.i.d. standard Gaussian entries.

We study the performance of nonlinear estimation methods for the small noise asymptotics $\delta, \varepsilon \to 0$ independently, under spatially inhomgeneous smoothness assumptions on f and a degree t of ill-posedness of the operator K. Combining the inversion (INV) provided by the Galerkin projection method, as proposed in the linear case by [2], on a large wavelet approximation space and wavelet thresholding as adaptive regularisation technique (REG), we investigate the two conceptually different methods

$$\text{method I:} \qquad \text{observations } g_\varepsilon,\ K_\delta \xrightarrow{INV} \hat{f}^{prelim}_{\delta,\varepsilon} \xrightarrow{REG} \text{estimator } \hat{f}^{I}_{\delta,\varepsilon},$$

$$\text{method II:} \qquad \text{observations } g_\varepsilon,\ K_\delta \xrightarrow{REG} \hat{g}_\varepsilon,\ \hat{K}_\delta \xrightarrow{INV} \text{estimator } \hat{f}^{II}_{\delta,\varepsilon}.$$

Both methods are provably rate optimal over a wide range of smoothness classes measured in the d-dimensional Besov space $B^s_{p,p}$, the optimal rate being $\max\{\delta, \varepsilon\}^{2s/(2s+2t+d)}$. Different limitations for the two methods in the case of $\delta \gg \varepsilon$ and a small degree s of smoothness are discussed. The impact of thresholding the wavelet representation of an operator is also of independent interest and can be combined with iterative solution methods as in [1].

References

[1] Cohen, A., Hoffmann, M. and Reiß, M. (2004): Adaptive wavelet Galerkin methods for linear inverse problems. *SIAM J. Numer. Anal.*, to appear.

[2] Efromovich, S. and Koltchinskii, V. (2001): On inverse problems with unknown operators. *IEEE Trans. Inf. Theory* **47**, 2876–2894.

[3] Hoffmann, M. and Reiß, M. (2004): Nonlinear methods for linear inverse problems with error in the operator, *Preprint Weierstraß Institut Berlin*, to appear.

Some theory for generalized boosting algorithms

Ya'acov Ritov

(joint work with Peter. J. Bickel, Alon Zakai)

The talk was based on Bickel, Ritov and Zakai (2004). We give a review of various aspects of boosting, clarifying the issues through a few simple results, and relate our work and that of others to the minimax paradigm of statistics. We consider the population version of the boosting algorithm and prove its convergence to the Bayes classifier, as a corollary of a general result about Gauss-Southwell optimization in Hilbert space. We then investigate the algorithmic convergence of the sample version, and give bounds to the time until perfect separation of the sample. We conclude by some result on the statistical optimality of the L_2 boosting.

We consider a standard classification problem: Let $(X,Y),(X_1,Y_1),\ldots,(X_n,Y_n)$ be an independent identically distributed sample, where $Y_i \in \{-1,1\}$ and $X_i \in \mathcal{X}$. The goal is to find a good classification rule, $\rightarrow\{-1,1\}$.

The AdaBoost algorithm was originally defined, Schapire (1990), Freund (1995), and Freund and Schapire (1996) as an algorithm to construct a good classifier by a "weighted majority vote" of simple classifiers. To be more exact, let $\mathcal{H}$ be a set of simple classifiers. The AdaBoost classifier is given by $\operatorname{sgn}\bigl(\sum_{m=1}^{M}\lambda_m h_m(x)\bigr)$, where $\lambda_m \in R$, $h_m \in \mathcal{H}$, are found sequentially.

Let $\mathcal{F}_\infty$ be the linear span of $\mathcal{H}$. That is,

$$\mathcal{F}_\infty = \bigcup_{k=1}^{\infty}\mathcal{F}_k,\ \text{where } \mathcal{F}_k = \{\sum_{j=1}^{k}\lambda_j h_j :\ \lambda_j \in R,\ h_j \in \mathcal{H},\ 1 \le j \le k\}\ .$$

A number of workers have noted, Breiman (1998,1999), Friedman, Hastie and Tibshirani (2000), Mason, Bartlett, Baxter and Frean (2000), and Schapire and Singer (1999) that the AdaBoost classifier could be viewed as $\operatorname{sgn}\bigl(F(X)\bigr)$, where F is found by a greedy algorithm minimizing $n^{-1}\sum_{i=1}^{n}\exp\bigl(-Y_iF(x_i)\bigr)$ over $\mathcal{F}_\infty$.

The general boosting algorithm is therefore of the following type. Given a loss function $W : R \rightarrow R^+$, we consider a greedy sequential procedure for finding a function F that minimizes $EW\bigl(YF(X)\bigr)$. That is, given $F_0 \in \mathcal{H}$ fixed, we define

for $m \geq 0$:

$$\lambda_m(h) = \underset{\lambda \in R}{\operatorname{argmin}}\ EW\Big(Y\big(F_m(X) + \lambda h(X)\big)\Big)$$
$$h_m = \underset{h \in \mathcal{H}}{\operatorname{argmin}}\ EW\Big(Y\big(F_m(X) + \lambda_m(h) h(X)\big)\Big)$$
$$F_{m+1} = F_m + \lambda_m(h_m) h_m.$$

From this point of view the algorithm appeared to be justifiable, since as was noted in Breiman (1999) and Friedman, Hastie, and Tibshirani (2000), the corresponding expression $E\exp\big(-YF(X)\big)$, obtained by replacing the sum by expectation, is minimized by $F(x) = \frac{1}{2}\log\Big(P(Y=1|X)/P(Y=-1|X)\Big)$, provided the linear span $\mathcal{F}_\infty$ is dense in the space $\mathcal{F}$of all functions in a suitable way. However, it was also noted that the empirical optimization problem necessarily led to rules which would classify every training set observation correctly and hence not approach the Bayes rule whatever be n. Jiang (2003) established that, for observation centered stumps, the algorithm converged to nearest neighbor classification, a good but rarely optimal rule.

The standard boosting algorithm is an example of a Gauss-Southwell procedure. We started with a general statement of a Gauss-Southwell algorithm in vector space. We then gave conditions that ensures that population version of AdaBoost or other boosting algorithm converges.

Then we study the Bayes consistency properties of the sample versions of the boosting algorithms we considered in Section 2. In particular, we shall

(i) Show that under mild additional conditions, there will exist a random sequence $m_n \to \infty$ such that $\hat{F}_{m_n} \xrightarrow{P} F_\infty$ where $\hat{F}_m$ is defined below as the mth sample iterate and moreover that such a sequence can be determined using the data.

(ii) Comment on the relationship of this result to optimization for penalized versions of W. The difference is that the penalty forces $m < \infty$ to be optimal while with us, cross-validation (or a test bed sample) determines the stopping point. We shall see that the same dichotomy applies later when we "boost" using the method of sieves for nonparametric regression studied by Barron, Birge and Massart (1999) and Baraud (2001).

We study that through a general argument on the convergence of an algorithm applied to the sample, when it is known that its population version converges. We coined this proof as the golden chain argument.

The smoothing of the boosting algorithm is based on an early stopping. We establish that the results by Györfi et al. (2002) are relevant to early stopping of boosting algorithm using test bed sub-sample.

We propose a regularization of L_2 boosting which we view as being in the spirit of the original proposal, but, unlike it, can be shown for, suitable $\mathcal{H}$, to achieve minimax rates for estimation of $E(Y|X)$ under quadratic loss for $\mathcal{P}$ for which $E(Y|X)$ is assumed to belong to a compact set of functions such as a ball in

Besov space if $X \in R$ or to appropriate such subsets of spaces of smooth functions in $X \in R^d$ — see, for example, the classes $\mathcal{F}$of Györfi et al. (2003). In fact, they are adaptive in the sense of Donoho et al (1995) for scales of such spaces. We note that Bühlmann and Yu (2003) have introduced a version of L_2 boosting which achieves minimax rates for Sobolev classes on R adaptively already. However, their construction is in a different spirit than that of most boosting papers. They start out with $\mathcal{H}$ consisting of one extremely smooth and complex function and show that boosting reduces bias (roughness the function) while necessarily increasing variance. Early stopping is still necessary and they show it can achieve minimax rates. In essence, our argument is based on using a finite sieve of weak classifiers at each step of the algorithm, moving to the next sieve only after the improvement in each stage falls below a given threshold.

References

[1] Andersen, P. K. and Gill R. D. (1982). Cox's regression model for counting processes: A large sample study. *Ann. Stat.* **10**, 1100–1120.

[2] Baraud, Y. (2001) Model Selection for regression on a random design. *Tech. Report*, U. Paris Sud.

[3] Barron, A. Birgé, L., Massart, P. (1999). Risk bounds for model selection under Penalization. *Prob. Theory and Related Fields*, **113**, 301–413.

[4] Bartlett, P. L., Jordan, M. I., and McAuliffe, J. D. (2003) Convexity, Classification, and Risk Bounds. *Tech. Report* **638**, Department of Statistics, University of California at Berkeley.

[5] Bickel, P. J. and Millar, P. W. (1992). Uniform convergence of probability measures on classes of functions. Statistica Sinica 2(1992), 1-15.

[6] Bickel, P. J. and Ritov, Y. (2003). The golden chain. A comment. *Ann. Statist.*, to appear.

[7] Bickel, P. J., Ritov, Y., and Zakai, A. (2004). Some theory for generalized boosting algorithms. Manuscript.

[8] Breiman, L. (1999) Prediction games and arcing algorithms. *Neural Computation* **11**, 1493-1517.

[9] Breiman, L. (1998) Arcing classifiers (with discussion). *Ann. Statist.* **26**, 801–849.

[10] Breiman L. (2000) Some infinity theory for predictor ensembles *Technical Report* U.C. Berkeley

[11] Bühlmann P. (2002) Consistency for L_2 boosting and matching pursuit with trees and tree type base functions. *Technical Report* ETH Zürich

[12] Bühlmann and Yu (2003). Boosting the L_2 loss: Regression and classification. *JASA*, to appear

[13] Donoho, D., Johnstone, I.M. Kerkyacharian, G. and Picard, D. (1995). Wavelet shrinkage: asymptopia (with discussion). *J. Roy. Statist. Soc. Ser. B***57**, 371–394.

[14] Friedman, J. H., Hastie, T., and Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting (with discussion). *Ann. Statist.* **28**, 337–407.

[15] Freund, Y. (1995). Boosting a weak learning algorithm by majority. *Information and Computation* **121**, 256–285.

[16] Freund, Y. and Schapire, R. E. (1996). Experiments with a new boosting algorithm. *Machine Learning: Proc. 13th International Conference*, 148–156. Morgan Kauffman, San Francisco.

[17] Györfi, G., Kohler, M., Krzyżak, A., and Walk, H. (2002). *A Distribution Free Theory of Nonparametric Regression*. Springer, New York.

[18] Jiang W. (2002) Process consistency for ADABOOST. Technical Report 00-05, Dept. of Statistics, Northwestern University.

[19] Lee, Y., Lin, Y., and Wahba, G. (2002). Multicategory Support Vector Machines, Theory, and Application to the Classification of Microarray Data and Satellite Radiance Data . Technical Report 1064. Department of Statistics, University of Wisconsin-Madison. Submitted.
[20] Luenberger, D. G. (1984) *Linear and Nonlinear Programming*. Addison-Wesley Publishing Company, Reading.
[21] Lugosi, G. and Vayatis, N. (2004) On the Bayes-risk consistency of boosting methods. *Ann. Statist.* **32**, 30–55.
[22] Mallat S. and Zhang Z. (1993) Matching pursuit with time frequency dictionaries. *IEEE Transactions on Signal Processing* **41**, 3397–3415.
[23] Mammen, E. and Tsybakov, A. (1999). Smooth discrimination analysis. *Ann. Statist.* **27**, 1808–1829.
[24] Mason, L., Bartlett, P., Baxter, J., and Frean, M. (2000). Functional gradient techniques for combining hypotheses. In Schölkopg, Smola, A., Bartlett, P., and Schurmans, D. (edts.) *Advances in Large Margin Classifiers*, MIT Press, Boston.
[25] Schapire, R. E. (1990). The strength of weak learnability. *Machine Learning* **5**, 197–227.
[26] Schapire, R. E., and Singer, Y. (1999). Improved boosting algorithms using confidence related predictions. *Machine Learning*, **37**, 297–336.
[27] Tsybakov, A. (2001). Optimal aggregation of classifiers in statistical learning. *Technical Report*, U. of Paris IV.
[28] Yang, Y. (1999) Minimax nonparametric classification – Part I Rates of convergence, Part II Model selection, *IEEE Trans. Inf. Theory* **45**, 2271–2292.
[29] Zhang, T. and Yu. B. (2003). Boosting with early stopping: convergence and consistency. Tech Report 635, Stat Dept, UCB.
[30] Zhang T. (2004) Statistical Behaviour and Consistency of Classification Methods based on Convex Risk Minimization. *Ann. Statist.*, **32**, 56–134.

Kernel Methods for Implicit Surface Modeling

Bernhard Schölkopf

(joint work with Olivier Chapelle, Joachim Giesen, Simon Spalinger, Florian Steinke, Christian Walder)

We describe methods for computing an implicit model of a hypersurface that is given only by a finite sampling. The methods work by mapping the sample points into a reproducing kernel Hilbert space (RKHS) and estimating hyperplanes in that space.

Suppose we are given a sampling $x_1, \ldots, x_m \in \mathcal{X}$, where the domain $\mathcal{X}$ is some hypersurface in Euclidean space $\mathbf{R}^d$. Today the most popular approach to approximately recover $\mathcal{X}$ from the sampling is to add connectivity information to the data by transforming them into a triangle mesh. But recently also implicit models, where the surface is modeled as the zero set of some sufficiently smooth function, gained some popularity. We use expansions in terms of a positive definite kernel k to model this function. Note that for every such kernel, there exists a map Φ into an associated RKHS $\mathcal{H}$ such that we have $k(x, x') = \langle \Phi(x), \Phi(x') \rangle$ for all $x, x' \in \mathcal{X}$; i.e., k computes the inner product in $\mathcal{H}$ (see e.g. [2]).

We discuss two methods. The first one computes a hyperplane (or more generally a *slab*) in the RKHS, with the property that it contains most of the sampling

points. To this end, we consider the following quadratic program:[1]

$$\begin{aligned}
&\underset{\mathbf{w}\in\mathcal{H},\boldsymbol{\xi}^{(*)}\in\mathbb{R}^m,\rho\in\mathbb{R}}{\text{minimize}} && \frac{1}{2}\|\mathbf{w}\|^2+\frac{1}{\nu m}\sum_i(\xi_i+\xi_i^*)-\rho && (1)\\
&\text{subject to} && \delta-\xi_i\le\langle\mathbf{w},\Phi(x_i)\rangle-\rho\le\delta^*+\xi_i^* && (2)\\
&\text{and} && \xi_i^{(*)}\ge 0. && (3)
\end{aligned}$$

Here, ν is a parameter that controls certain geometric properties of the solution, involving the points which come to lie on the hyperplane(s), and the ones that are outside of the slab; the parameters δ and δ^* control the width of the slab (see [1] for details).

This problem can be expressed in dual form as

$$\begin{aligned}
&\underset{\boldsymbol{\alpha}\in\mathbb{R}^m}{\text{minimize}} && \frac{1}{2}\sum_{ij}(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)k(x_i,x_j)-\delta\sum_i\alpha_i+\delta^*\sum_i\alpha_i^* && (4)\\
&\text{subject to} && 0\le\alpha_i^{(*)}\le\frac{1}{\nu m}\text{ and }\sum_i(\alpha_i-\alpha_i^*)=1. && (5)
\end{aligned}$$

Once this quadratic program is solved, we can evaluate for each test point x whether it satisfies $\delta\le\langle\mathbf{w},\Phi(x)\rangle-\rho\le\delta^*$. In other words, we have an implicit description of the region in input space that corresponds to the region in between the two hyperplanes in the RKHS. For $\delta=\delta^*$, this is a single hyperplane, corresponding to a hypersurface in input space.[2] To evaluate this criterion, we use the kernel expansion

$$\langle\mathbf{w},\Phi(x)\rangle=\sum_i(\alpha_i-\alpha_i^*)k(x_i,x). \tag{6}$$

A close relationship to one-class Support Vector (SV) methods as well as to SV regression algorithms exists. Moreover, the method can be made more robust by including points which lie equidistantly on the two sides of the surface, and forcing the hyperplane to assume corresponding distances to them. In this case, the value of (6) can be used to give an estimate of the signed distance function to the surface.

In the second method, we use not only points sampled from the surface, but also "background" points sampled uniformly from a grid covering a hyper-rectangle enclosing $\mathcal{X}$. As an estimation algorithm, we use *oriented PCA* performed in an RKHS. The algorithm attempts to estimate a direction $\mathbf{w}$ in the RKHS to maximize the ratio of two variances, one being the variance of the background points projected onto $\mathbf{w}$ and the other one being the variance of the surface points projected onto $\mathbf{w}$. A hyperplane orthogonal to $\mathbf{w}$ is then used as a model of the surface.

[1] Here and below, the superscript $(*)$ simultaneously denotes the variables with and without asterisk, e.g., $\boldsymbol{\xi}^{(*)}$ is a shorthand for $\boldsymbol{\xi}$ and $\boldsymbol{\xi}^*$.

[2] subject to suitable conditions on k

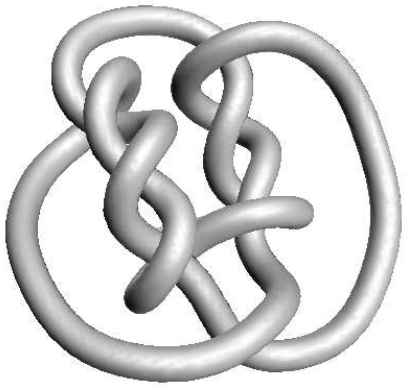

FIGURE 1. An example of applying a variant of the first algorithm, taken from [1].

References

[1] B. Schölkopf, J. Giesen, & S. Spalinger, *Kernel Methods for Implicit Surface Modeling*, to appear in the proceedings of NIPS'2004.
[2] B. Schölkopf & A. Smola, *Learning with Kernels*, MIT Press, Cambridge, MA (2002)

Examples of Statistical Learning (in Life Sciences)

YOUNG K. TRUONG
(joint work with X. Lin, C. Beecher, A. Cutler, S. Young, S. Simmons)

Three examples of learning are presented based on human data from genomics, metabolomics and functional Magnetic Resonance Imaging (fMRI). The first example is about building a gene network that shows regulatory relationships between genes and a brief survey of Bayesian network involving functional modeling is given. The second applies support vector machines (SVM), and random forest (RF), for outlier detection, variable selection and classification based on a metabolomic data set. Some new features of RF such as importance of variables and multidimensional scaling as well as a new window interface (POWERMV) will be illustrated. The selected predictors may be discussed in the context of the biochemistry of the disease. The third is about the exploration and analysis of large fMRI datasets. We briefly describe the practical issues in preprocessing the data, followed with statistical methods involving time series and independent component analysis to detect the region of interest.

References

[1] S. Simmons, X. Lin, C. Beecher, Y. Truong and S. Young, *Active and passive learning to explore a complex metabolism data set*, International Federation Classification Association (2004), to appear.
[2] Y. Truong, X. Lin, C. Beecher, A. Cutler and S. Young, *Learning a complex metabolomic dataset using random forests and support vector machines*, Knowledge Discovery and Data Mining, Seattle, Washington, (2004), to appear.

The multicategory support vector machine and the multicategory penalized likelihood estimate

Grace Wahba

(the talk is based on the work of and work with Yoonkyung Lee, Yi Lin, Stephen A. Ackerman, Hao Helen Zhang,Xiwu Lin)

This talk was about building a model for the classification of objects into one of two or one of several categories; firstly, via the estimation of the posterior probability of each category using penalized likelihood methods, and, secondly, using the recently popular support vector machine (SVM) method.

We briefly reviewed the Neyman Pearson lemma and the Bayes rule for optimal classification, and went on to review the representer theorem [2], to describe the penalized likelihood estimate for two categories [12], and to describe the support vector machine [1] proposed by V. Vapnik and collaborators in the early 90's. We noted that [9] showed in the two category case that the SVM is estimating the *sign* of the log odds ratio, just exactly what you need to implement the Bayes rule.

Our main contribution to the theory and practice of SVMs is the development of the multicategory SVM (MSVM) as discussed in [6] and [7]. We also briefly described a multicategory penalized likelihood estimate from [8], which is constrasted with the MSVM in [13]. Some remarks were made concerning when one would like to use the penalized likelihood estimate and when, the SVM or MSVM.

We gave theoretical details from [6] and showed pictures from [7], where the MSVM was used to classify vectors of upwelling radiances at twelve wavelenghts as observed by the MODIS satellite, into clear sky, water cloud or ice cloud pixels. Impressive results were obtained. Preprints or reprints of papers with Wahba as co-author are available via `http://www.stat.wisc.edu/~wahba` click on TRLIST, and for earlier papers click on `golden oldies`.

Other contributions to the SVM literature by the author, coauthors/former students include [3] [4] [5] [10] [11] [14] [15] [16].

References

[1] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines*. Cambridge University Press, 2000.

[2] G. Kimeldorf and G. Wahba. Some results on Tchebycheffian spline functions. *J. Math. Anal. Applic.*, 33:82–95, 1971.

[3] Y. Lee. *Multicategory Support Vector Machines, Theory, and Application to the Classification of Microarray Data and Satellite Radiance Data*. PhD thesis, Technical Report 1062, Department of Statistics, University of Wisconsin, Madison WI, 2002.

[4] Y. Lee and C.-K. Lee. Classification of multiple cancer types by multicategory support vector machines using gene expression data. *Bioinformatics*, 19:1132–1139, 2003.

[5] Y. Lee, Y. Lin, and G. Wahba. Multicategory support vector machines. *Computing Science and Statistics*, 33:498–512, 2001.

[6] Y. Lee, Y. Lin, and G. Wahba. Multicategory support vector machines, theory, and application to the classification of microarray data and satellite radiance data. *J. Amer. Statist. Assoc.*, 99:67–81, 2004.

[7] Y. Lee, G. Wahba, and S. Ackerman. Classification of satellite radiance data by multicategory support vector machines. *J. Atmos. Ocean Tech.*, 21:159–169, 2004.
[8] X. Lin. *Smoothing Spline Analysis of Variance for Polychotomous Response Data*. PhD thesis, Department of Statistics, Uiversity of Wisconsin, Madison WI, 1998. Also Statistics Dept TR 1003 available via Grace Wahba's website.
[9] Y. Lin. Support vector machines and the bayes rule in classification. *Data Mining and Knowledge Discovery*, 6:259–275, 2002.
[10] Y. Lin, Y. Lee, and G. Wahba. Support vector machines for classification in nonstandard situations. *Machine Learning*, 46:191–202, 2002.
[11] Y. Lin, G. Wahba, H. Zhang, and Y. Lee. Statistical properties and adaptive tuning of support vector machines. *Machine Learning*, 48:115–136, 2002.
[12] F. O'Sullivan, B. Yandell, and W. Raynor. Automatic smoothing of regression functions in generalized linear models. *J. Amer. Statist. Assoc.*, 81:96–103, 1986.
[13] G. Wahba. Soft and hard classification by reproducing kernel Hilbert space methods. *Proc.National Academy of Sciences*, 99:16524–16530, 2002.
[14] G. Wahba, Y. Lin, Y. Lee, and H. Zhang. On the relation between the GACV and Joachims' $\xi\alpha$ method for tuning support vector machines, with extensions to the non-standard case. Technical Report 1039, Statistics Department University of Wisconsin, Madison WI, 2001.
[15] G. Wahba, Y. Lin, Y. Lee, and H. Zhang. Optimal properties and adaptive tuning of standard and nonstandard support vector machines. In D. Denison, M. Hansen, C. Holmes, B. Mallick, and B. Yu, editors, *Nonlinear Estimation and Classification*, pages 129–148. Springer, 2002.
[16] G. Wahba, Y. Lin, and H. Zhang. Generalized approximate cross validation for support vector machines. In A. Smola, P. Bartlett, B. Scholkopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 297–311. MIT Press, 2000.

Estimating Dimension Reduction Directions via Conditional Density Functions

Xia Yingcun

In this paper, we propose two new approaches to estimate the efficient dimension reduction (EDR) directions based on the definition directly. Compared with the inverse regression methods [1, 2, 3], the new methods require no strong assumptions on the design of covariates and the link function, and have better performance than the inverse regression methods for finite samples. Compared with the direct regression methods [4, 5, 6], which can only estimate the EDR directions in the regression mean, the new methods can detect the EDR directions exhaustively. Consistency of the estimators are proved. Especially, the root-n rate can be achieved when the efficient dimension is less than 4 regardless of the number of covariates. The corresponding algorithms are also proved to be convergent.

References

[1] Li, K. C. (1991) Sliced inverse regression for dimension reduction (with discussion). *Amer. Statist. Ass.*, **86**, 316-342.
[2] Li, K. C. (1992) On Principal Hessian Directions for Data Visualization and Dimension Reduction: Another Application of Stein's Lemma. *Journal of the American Statistical Association*, **87** 1025-1039.
[3] Cook, R. D., and Weisberg, S. (1991) Sliced Inverse Regression for Dimension Reduction: Comment. *Amer. Statist. Ass.*, **86**, 328-332

[4] Härdle, W. and Stoker, T. M. (1989) Investigating smooth multiple regression by method of average derivatives. *J. Amer. Stat. Ass.* **84** 986-995.
[5] Hristache, M., Juditski, A, Polzehl, J., Spokoiny, V. (2001). Structur adaptive approach for dimension reduction. *Annals of Statistics* **29** 1537–1566.
[6] Xia, Y., Tong, H., Li, W. K. and Zhu, L. (2002) An adaptive estimation of dimension reduction space (with discussions). *J. Roy. Statist. Soc. B.*, **64**, 363-410.

Controlling the False Discovery Rate in large complex Studies

Daniel Yekutieli
(joint work with Yoav Benjamini)

The False Discovery Rate (FDR) criterion, its variations, and procedures that control it will be reviewed. Its relevance to very large problems, such as the analysis of locations on the chromosome associated with complex traits (QTL analysis), the analysis of gene expression data, and behavioral genetics will be discussed. We shall explain the complex nature of the large problems encountered in these areas, and the need for new methodologies. We shall present hierarchical testing, where the research question is organized in tree structure, and the testing is organized in a way that will assure control of the FDR. While keeping in sight the actual genetic problems. Practical issues of implementation will be addressed as well.

References

[1] Benjamini, Y., Yekutieli, D., (2001) *The control of the False Discovery Rate in Multiple Testing under Dependency*, Annals of Statistics, **29** (4), 1165–1188.
[2] Benjamini, Y., Yekutieli, D., (2004) *False Discovery Rate Adjusted Confidence Intervals for Selected Parameters*, to appear in the Journal of the American Statistics Association.
[3] Benjamini, Y., Yekutieli, D., (2002) *Hierarchical FDR testing of trees of hypotheses*, Research Paper 02-02 the Department of Statistics and OR, Tel Aviv University

Variable Selection via COSSO in Nonparametric Regression Models

Hao Helen Zhang
(joint work with Yi Lin)

Consider the regression model $y_i = f(\mathbf{x}_i) + \epsilon_i,\ i = 1, ..., n$, where f is the unknown regression function to be estimated, $\mathbf{x}_i = (x_i^{(1)}, ..., x_i^{(d)}) \in \Pi_{j=1}^{d} \mathcal{X}^{(j)}$ is a d-dimensional vector of covariates, and the ϵ's are independent noises with mean 0 and variance σ^2. For parametric regression models, traditional variable selection approaches include best subset selection and stepwise selection. Recently shrinkage methods, such as the nonnegative garrotte (Breiman 1995), the LASSO (Tibshirani 1996), and the SCAD (Fan & Li 2001), have been proposed to shrink regression coefficients towards zero and hence select the optimal subset.

In [1], we developed a new regularization method for model selection and model fitting in nonparametric regression models. Under the framework of the smoothing

spline ANOVA (SS-ANOVA), we can decompose any multivariate function f as $f(\mathbf{x}) = b + \sum_{j=1}^{d} f_j(x^{(j)}) + \sum_{j<k} f_{jk}(x^{(j)}, x^{(k)}) + \cdots$, in terms of a constant, main effect components, two-way interaction components, and so on. Assume each main component lies in some reproducing kernel Hilbert space (RKHS) $\mathcal{H}^{(j)}$ over $\mathcal{X}^{(j)}$. Then $f \in \mathcal{F}$ which is a subspace of the tensor product space $\otimes_{j=1}^{d} \mathcal{H}^{(j)}$, since only low order terms are typically retained in the decomposition for interpretability. We call our method COmponent Selection and Smoothing Operator (COSSO), in which the penalty functional is the sum of RKHS norms of function components in the SS-ANOVA decomposition. When the noise is Gaussian and f is the additive model, the COSSO solves

$$\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \{y_i - f(\mathbf{x}_i)\}^2 + \tau J(f), \quad \text{with} \quad J(f) = \sum_{j=1}^{d} \|P^j f\|,$$

where P^j is the projection operator onto $\mathcal{H}^{(j)}$, $\|\cdot\|$ is the norm defined in the RKHS $\mathcal{F}$, and $\tau > 0$ is the smoothing parameter. The penalty $J(f)$ is different from the squared norm employed in the traditional smoothing spline method.

Theoretical properties, such as the existence and the rate of convergence of the COSSO estimator, are studied in [1]. The COSSO minimizer exists in $\mathcal{F}$ and lies in a finite dimensional space. The solution is further shown to be unique when evaluated at the data points. Asymptotically, the COSSO estimator in the additive model has a rate of convergence $n^{-\ell/(2\ell+1)}$, where ℓ is the order of smoothness of the components. In the special case of a tensor product design with periodic functions, the COSSO conducts model selection by applying a novel soft thresholding type operation to the function components, and selects the correct model structure with probability tending to one. For tuning τ, both GCV and five-fold cross validation are tried. A generic algorithm is then proposed to find the COSSO minimizer by iteratively solving the smoothing spline and the nonnegative garrotte. Compared with the MARS, the COSSO gives very competitive performances in both simulations and real examples.

In [2], we further generalized the COSSO as a penalized likelihood method for nonparametric regression in exponential families. The general framework developed there allows the treatment of many types of responses such as non-normal responses, binary and polychotomous responses, and the event counts data. With some special way of kernel construction, the COSSO fits and selects continuous and categorical covariates in a unified manner. The existence of the COSSO penalized likelihood estimator is established under some mild assumptions. An iterative algorithm is developed using Newton-Raphson method, and to further improve the computation speed in large sample size problems, we also give a subset basis algorithm using the parsimonious basis approach.

A related work, likelihood basis pursuit (LBP), was developed in our earlier paper [3]. In the context of SS-ANOVA models, each functional component is represented by a large number of basis functions which were chosen to be compatible with variable selection and model building. Basis pursuit is then applied to obtain the optimal ANOVA decomposition in terms of having the smallest l_1 norm on the

coefficients. The LBP produces sparse solutions which greatly facilitate the variable selection process. Though both give possible generalizations of the LASSO to nonparametric regression models, the LBP conducts coefficient shrinkage while the COSSO applies component shrinkage in some function space.

References

[1] Lin, Y. and Zhang, H. H. *Component selection and smoothing in smoothing spline analysis of variance models*, revised for journal, (2003).

[2] Zhang, H. H. and Lin, Y. *Component selection and smoothing for nonparametric regression in exponential families*, submitted, (2004).

[3] Zhang, H. H., Wahba, G., Lin, Y., Voelker, M., Ferris, M., Klein, R. and Klein, B. *Variable selection and model building via likelihood basis pursuit*, Journal of American Statistical Association, **99**, (2004), 659-672.

Genetic Analysis of Ordinal Traits and Statistical Challenges

Heping Zhang

(joint work with Rui Feng, Hongtu Zhu)

Genetic mechanisms underlying many human diseases and conditions. The existing genetic analysis methods require, however, that the diseases or conditions must either be dichotomized or measured by a quantitative trait such as blood pressure for hypertension. In the latter case, normality is generally assumed for the trait. On the other hand, many diseases and conditions are rated on ordinal scales such as cancer and mental and behavioral conditions. Our objective is to establish a framework to conduct genetic analysis for ordinal traits. We proposed and exploited a latent variable, proportional odds logistic model that relates inheritance patterns to the distribution of the ordinal trait. I will present simulation studies and real examples to demonstrate that the power of our proposed model to detect genetic effects was substantially higher than other methods based on binary traits. I will also present statistical challenges for understanding the asymptotic distributions of some test statistics and for computing the statistics.

References

[1] H.P. Zhang, R. Feng, and H.T. Zhu, *A latent variable model of segregation analysis for ordinal traits.* JASA **98** (2003), 1023–1034.

[2] R. Feng, J.F. Leckman, and H.P. Zhang, *Linkage analysis of ordinal traits for pedigree data* PNAS **101** (2004), 16739-16744.

Reporter: Denis Belomestny

Participants

Jean-Yves Audibert
CERTIS - Ecole Nationale des
Ponts et Chaussees
19, rue Alred Nobel
Cite Descartes Champs-s-Marne
F-77455 Marne-la-Vallee Cedex 2

Dr. Denis Belomestny

Weierstraß-Institut für
Angewandte Analysis und Stochastik
im Forschungsverbund Berlin e.V.
Mohrenstr. 39
10117 Berlin

Dr. Gilles Blanchard

Fraunhofer Institut FIRST
Intelligent Data Analysis Group
(IDA)
Kekulestr. 7
12489 Berlin

Dr. Peter Bühlmann

Seminar für Statistik
ETH-Zürich
LEO C 17
CH-8092 Zürich

Prof. Dr. Joachim M. Buhmann

Institut für Computational Science
ETH Zentrum
HRS F 31
Hirschengraben 84
CH-8092 Zürich

Prof. Dr. T. Tony Cai

Department of Statistics
The Wharton School
University of Pennsylvania
Philadelphia, PA 19104-6302
USA

Prof. Dr. Jean-Francois Cardoso

Dept. TSI
E. N. S. Telecommunications
46, rue Barrault
F-75634 Paris Cedex

Prof. Dr. Song Xi Chen

Department of Statistics
Iowa State University
315F Snedecor Hall
Ames IA 50011
USA

Prof. Dr. Ying Chen

Wirtschaftswissenschaftl. Fakultät
Lehrstuhl für Statistik
Humboldt-Universität Berlin
Spandauer Str. 1
10178 Berlin

Prof. Dr. Rainer Dahlhaus

Institut für Angewandte Mathematik
Universität Heidelberg
Im Neuenheimer Feld 294
69120 Heidelberg

Prof. Dr. Arnak Dalalyan

Laboratoire de Probabilites
Universite Paris 6
tour 56
4 place Jussieu
F-75252 Paris Cedex 05

Prof. Dr. Holger Dette

Fakultät für Mathematik
Ruhr-Universität Bochum
44780 Bochum

Prof. Dr. Lutz Dümbgen

Mathematische Statistik
und Versicherungslehre
Universität Bern
Sidlerstraße 5
CH-3012 Bern

Prof. Dr. Jianqing Fan

Department of Operations Research
and Financial Engineering
Princeton University
Princeton NJ 08544
USA

Prof. Dr. Jürgen Franke

Fachbereich Mathematik
Universität Kaiserslautern
67653 Kaiserslautern

Prof. Dr. Sara van de Geer

Mathematical Institute
University of Leiden
Niels Bohrweg 1
NL-2300 RA Leiden

Prof. Dr. Friedrich Götze

Fakultät für Mathematik
Universität Bielefeld
Postfach 100131
33501 Bielefeld

Prof. Dr. Alexander Goldenshluger

Dept. of Statistics
University of Haifa
Haifa 31905
ISRAEL

Prof. Dr. Wolfgang Härdle

Wirtschaftswissenschaftl. Fakultät
Lehrstuhl für Statistik
Humboldt-Universität Berlin
Spandauer Str. 1
10178 Berlin

Prof. Dr. Anatoli Iouditski

Lab. de Modelisation et Calcul
Institut IMAG
Univ. Joseph Fourier Grenoble I
B.P. 53
F-38041 Grenoble Cedex 9

Dr. Motoaki Kawanabe

Fraunhofer Institut FIRST
Intelligent Data Analysis Group
(IDA)
Kekulestr. 7
12489 Berlin

Prof. Dr. Enno Mammen

Lehrstuhl für Volkswirtschaftslehre
und Ökonometrie
Universität Mannheim
L 7, 3-5
68161 Mannheim

Prof. Dr. James Stephen Marron

Department of Statistics and
Operations Research
University of North Carolina
Chapel Hill, NC 27599-3260
USA

Prof. Dr. Hans-Georg Mueller

Division of Statistics
University of California
469 Kerr Hall
Davis, CA 95616-8705
USA

Prof. Dr. Klaus-Robert Müller

Fraunhofer Institut FIRST
Intelligent Data Analysis Group
(IDA)
Kekulestr. 7
12489 Berlin

Prof. Dr. Andrew Nobel

Department of Statistics and
Operations Research
University of North Carolina
Chapel Hill, NC 27599-3260
USA

Prof. Dr. Wolfgang Polonik

Department of Statistics
University of California
Davis
One Shields Avenue
Davis CA 95616
USA

Dr. Jörg Polzehl

Weierstraß-Institut für
Angewandte Analysis und Stochastik
im Forschungsverbund Berlin e.V.
Mohrenstr. 39
10117 Berlin

Dr. Gunnar Rätsch

Friedrich-Miescher-Laboratory of
the Max-Planck-Society
Spemannstr. 39
72076 Tübingen

Markus Reiss

WIAS
Mohrenstr. 39
10117 Berlin

Prof. Dr. Yaacov Ritov

Department of Statistics
The Hebrew University of Jerusalem
Mount Scopus
Jerusalem 91905
ISRAEL

Prof. Dr. Bernhard Schölkopf

Max-Planck-Institut für
Biologische Kybernetik
Spemannstraße 38
72076 Tübingen

Prof. Dr. David Siegmund

Department of Statistics
Stanford University
Sequoia Hall
Stanford, CA 94305-4065
USA

Prof. Dr. Alex Smola

Machine Learning Program
National ICT Australia
Australian National University
Canberra ACT 0200
Australia

Prof. Dr. Vladimir Spokoiny

WIAS
Mohrenstr. 39
10117 Berlin

Prof. Dr. Young K.N. Truong

Department of Biostatistics
School of Public Health,201 H
University of North Carolina
Chapel Hill, NC 27599-7400
USA

Prof. Dr. Grace Wahba

Department of Statistics
University of Wisconsin
MSC, 1300 University Ave.
Madison WI 53706-1685
USA

Prof. Dr. Yazhen Wang

Department of Statistics
University of Connecticut
Box U-4120
Storrs, CT 06269-3120
USA

Prof. Dr. Yingcun Xia

Department of Statistics and
Applied Probability, Block S16
National University of Singapore
6 Science Drive 2
Singapore 117546
Singapore

Prof. Dr. Qiwei Yao

Department of Statistics
London School of Economics
Houghton Street
GB-London,WC2A 2AE

Prof. Dr. Daniel Yekutieli

Department of Stat. and Operat. Res
Tel Aviv University
Tel Aviv 69978
ISRAEL

Prof. Dr. Hao Helen Zhang

Dept. of Statistics
North Carolina State University
Raleigh, NC 27695-8203
USA

Prof. Dr. Heping Zhang

Department of Epidemiology and
Public Health
Yale University
60 College Street
New Haven CT 06520-8034
USA

Andreas Ziehe

Fraunhofer Institut FIRST
Intelligent Data Analysis Group
(IDA)
Kekulestr. 7
12489 Berlin