Sparsity and Inverse Problems in Statistical Theory and Econometrics

November 29, 2008

Gérad Biau

Consistency of random forests and other averaging classifiers (joint work with L. Devroye and G. Lugosi)

In the last years of his life, Leo Breiman promoted random forests for use in classification. He suggested using averaging as a means of obtaining good discrimination rules. The base classifiers used for averaging are simple and randomized, often based on random samples from the data. He left a few questions unanswered regarding the consistency of such rules. In this talk, we give a number of theorems that establish the universal consistency of averaging rules.

Peter Bickel

Kernel Representations and Kernel Density Estimation

There has been a great deal of attention in recent times particularly in machine learning to representation of multivariate data points x by $K(x, \cdot)$ where K is positive and symmetric and thus induces a reproducing kernel Hilbert space.The idea is then to use the matrix $||K(X_i, X_j)||$ as a substitute for the empirical covariance matrix of a sample $X_1, \ldots, X_n$ for PCA and other inference.(Jordan and Fukumizu(2006) for instance. Nadler et. al(2006) connected this approach to one based on random walks and diffusion limits and indicated a connection to kernel density estimation.By making at least a formal connection to a multiplication operator on a function space we make further connection and show how clustering results of Beylkin ,Shih and Yu (2008) which apparently differ from Nadler et al. can be explained.

Peter Bühlmann

Stability Selection for High-Dimensional Data

Despite remarkable progress over the past 5 years, estimation of high-dimensional structure, such as in graphical modeling, cluster analysis or variable selection in (generalized) regression, remains difficult. Among the main problems are: (i) the choice of an appropriate amount of regularization; (ii) a potential lack of stability of a solution and quantification of evidence or significance of a selected structure or of a set of selected variables.

We introduce the new method of stability selection which addresses these two major problems for high-dimensional structure estimation, both from a practical and theoretical point of view. Stability selection is based on subsampling in combination with (high-dimensional)selection algorithms. As such, the method is extremely general and has a very wide range of applicability. Stability selection provides finite sample control for some error rates of false discoveries and hence a transparent principle to choose a proper amount of regularization for structure estimation or model selection. Maybe even more importantly, results are typically remarkably insensitive to the chosen amount of regularization. Another property of stability selection is the empirical and theoretical improvement over pre-specified selection methods. We prove for randomized Lasso that stability selection will be model selection consistent even if the necessary conditions needed for consistency of the original Lasso method are violated. We demonstrate stability selection for variable selection, Gaussian graphical modeling and clustering, using real and simulated data.

This is joint work with Nicolai Meinshausen.

David Hardoon

Sparse Canonical Correlation Analysis

We present a novel method for solving Canonical Correlation Analysis (CCA) in a sparse convex framework using a least squares approach. The presented method focuses on the scenario when one is interested in (or limited to) a primal representation for the first view while having a dual representation for the second view. Sparse CCA (SCCA) minimises the number of features used in both the primal and dual projections while maximising the correlation between the two views. The method is demonstrated on two paired corpuses of English-French and English-Spanish for mate-retrieval. We are able to observe, in the mate-retreival, that when the number of the original features is large SCCA outperforms Kernel CCA (KCCA), learning the common semantic space from a sparse set of features.

Stefan Haufe:

Groupwise sparsity enforcing estimators for solving the EEG/MEG inverse problem

Cerebral current flows are directly related to information transfer in the brain and thus an excellent means for studying the mechanisms of cognitive processing. Electro- and Magneto-encephalography, EEG and MEG, are noninvasive measures of these electric currents (EEG) or their respective accompanying magnetic fields (MEG). The reconstruction of the cerebral current density from EEG/MEG measurements is an ill-posed inverse problem. As the forward mapping from the current sources to the external sensors is linear, the inverse problem may be formulated as a highly under determined linear system of equations, which has no unique solution. The common strategy to deal with this ambiguity is regularization, i.e. fitting the data with an additional penalization of the sources. Both l2-norm and l1-norm based penalties have been proposed based on the neurophysiologically motivated assumptions of smoothness and sparsity, respectively.

However, in practise it is often observed that the estimated current densities are either too distributed (l2-norm) or too scattered (l1-norm) to be neurophysiologically plausible.

In this talk we present two straightforward approaches that achieve a compromise between smoothness and sparsity and are consequently able to deliver sources with plausible extent. Our penalties are based on the l1-norm where, however, care has been taken to ensure rotational invariance of the estimated current density, which is a R3 —-¿ R3 vector field. These considerations lead us to an l1,2-matrix-norm regularized problem, in which the three variables belonging to a certain vectorial predictor are required to become jointly sparse. The groups of variables may also be extended to contain variables for several measurements which are assumed to be generated by the same set of sources. This strategy is particularly helpful in the presence of multiple trials with common zero-mean activity (e.g. oscillations) which cannot be linearly averaged in sensor space. We show empirically that the inversion of single trials with joint sparsity regularization leads to a noise reduction similar to the one that can be achieved by signal averaging in sensor space.

Technically, our problem formulations lead to convex but nondifferentiable objective functions which are equivalent to the group lasso (Ming & Yuan, 2005) known in statistics. The optimization is carried out either by casting the problem as an instance of second-order-cone programming or by employing an active-set algorithm (Roth & Fischer, 2008) which exploits the sparse structure of the solution. The latter approach enables us to fit models with several thousands of observations and several millions of variables in reasonable time.

References:

[1] Haufe S, Nikulin VV, Ziehe A, Müller KR, Nolte G. Combining sparsity and rotational invariance in EEG/MEG source reconstruction. NeuroImage, 42(2):726738, 2008.

[2] Haufe S, Nikulin VV, Ziehe A, Müller KR, Nolte G. Estimating vector fields using sparse basis field expansions. In Advances in Neural Information Processing Systems 21. MIT Press, Cambridge, MA, 2009. In press.

Zakria Hussain

Matching pursuit algorithms in machine learning

I will describe a generic matching pursuit algorithm that can be used in machine learning for regression, subspace methods (kernel PCA and kernel CCA) and classification (given time). I will also describe some generalisation error bounds upper bounding their loss. Some of these bounds will be formed using standard sample compression bounds whilst others will be amalgamations of traditional learning theory techniques such as VC theory and Rademacher complexities. This is joint work with John Shawe-Taylor.

Joel Horowitz

VARIABLE SELECTION IN NONPARAMETRIC ADDITIVE MODELS

We consider a nonparametric additive model of a conditional mean function in which the number of variables and additive components may be much larger than the sample size but the number of non-zero additive components is small relative to the sample size. The statistical problem is to determine which additive components are non-zero. The additive components are approximated by truncated series expansions with B-spline bases. The adaptive group LASSO is used to select non-zero components. We give conditions under which this procedure selects the non-zero components correctly with probability approaching one as the sample size increases. Following model selection, oracle-efficient, asymptotically normal estimators of the non-zero components can be obtained by using existing methods. The results of Monte Carlo experiments show that the adaptive group LASSO procedure works well with samples of moderate size.

Alois Kneip

The prediction error in functional regression

The talk considers functional linear regression, where scalar responses Y are modeled in dependence of random functions. We propose a smoothing splines estimator for the functional slope parameter based on a slight modification of the usual penalty. Theoretical analysis concentrates on the error in an out-of-sample prediction of the response for a new random function. It is shown that rates of convergence of the prediction error depend on the smoothness of the slope function and on the structure of the predictors. We then prove that these rates are optimal in the sense that they are minimax over large classes of possible slope functions and distributions of the predictive curves. For the case of models with errors-in-variables the smoothing spline estimator is modified by using a denoising correction of the covariance matrix of discretized curves. The methodology is then applied to a real case study where the aim is to predict the maximum of the concentration of ozone by using the curve of this concentration measured the preceding day.

Volker Krätschmer

Inverse problems in empirical risk attitudes

Supported by several recent investigations the empirical pricing kernel paradox might be considered as a stylized fact. In Chabi-Yo, Garcia & Renault (2008) simulation studies have been presented which suggest that this paradox might be caused by regime switching of stock prices in financial markets. Alternatively, we want to emphasize a microeconomic view. Based on an economic model with state dependent utilities for the financial investors we succeed in explaining the paradox by changes of the risk attitudes. Theoretically, the change behaviour is compressed by the pricing kernels. As a starting point for empirical insights we shall develop and investigate an inverse problem in terms of data fits for estimated basic values of the pricing kernel. Also numerical solutions of this problem will be presented.

Sébastien Loustau

Statistical performances of SVM Regularization in Classification

Ill-posed inverse problems $Af = g$ are usually characterized by the non-continuity of the inverse operator $A - 1$. It means that small perturbations of the observations g have strong consequences over the solution f. To overcome this difficulty, regularization scheme has been proposed in the literature, such as Tikhonov regularization. It has been applied in Learning with rather great success. One of the main example is the well-known SVM regularization, given by: min

$$\min_{f \in \mathcal{H}} \Big(\frac{1}{n} \sum_{i=1}^{n} l(Y_i, f(X_i i)) + \alpha_n \|f\|_{\mathcal{H}}^2,$$

where $\mathcal{H}$ is commonly a Reproducing Kernel Hilbert Space and α_n is a smoothing parameter. A major statistical problem is to calibrate α_n. In this talk, we propose to study the statistical performances of SVM regularization for binary classification. We give rates of convergence for an optimal choice of the smoothing parameter. We also consider the problem of adaptation and propose similar performances for an adaptive convex combination of SVM. We illustrate these theoretical results with practical experiments over real-world data sets and discuss some possible theoretical and practical improvements.

Alexander Meister

Nonparametric estimation of the error distribution in software testing

We introduce an estimation procedure for the error distribution based on additive relations in a nonparametric setting with application to software testing. Therein, we face a nonlinear statistical inverse problem, which can be solved by Fourier methods. We derive exact confidence intervals and prove rate-optimality for the derived method.

Hannes Nickisch and Matthias Seeger

Variational Inference and Experimental Design in Sparse Linear Models

Sparsity is a fundamental concept in modern statistics, and often the only general principle available at the moment to address novel learning applications with many more variables than observations. Despite the recent advances of the theoretical understanding and the algorithmics of sparse point estimation, higher-order problems such as covariance estimation or optimal data acquisition are seldomly addressed for sparsity-favouring models, and there are virtually no scalable algorithms.

We provide an approximate Bayesian inference algorithm for sparse linear models, that can be used with hundred thousands of variables. Our method employs a convex relaxation to variational inference and settles an open question in continuous Bayesian inference: The Gaussian lower bound relaxation is convex for a class of super-Gaussian potentials including the Laplace and Bernoulli potentials.

Our algorithm reduces to the same computational primitives used for sparse estimation methods, but requires Gaussian marginal variance estimation as well. We show how the Lanczos algorithm from numerical mathematics can be employed to compute the latter.

We are interested in Bayesian experimental design, a powerful framework for optimizing measurement architectures. We have applied our framework to problems of magnetic resonance imaging design and reconstruction.

Benedikt Poetscher:

Confidence Sets Based on Sparse Estimators and Related Results on the Distribution of Penalized Maximum Likelihood Estimators

We show that confidence sets based on sparse estimators are necessarily large and we discuss the finite-sample as well as the asymptotic distribution of penalized maximum likelihood estimators in a simple framework. (The talk is based partly on joint work with Hannes Leeb and Ulrike Schneider.)

Nora Serdyukova

Approximation of Random Fields in High Dimension

We consider the ε-approximation by n-term partial sums of the Karhunen-Loève expansion to d-parametric random fields of tensor product-type in the average case setting. We investigate the behavior, as $d \to \infty$, of the information complexity of approximation with error not exceeding a given level ε. It was recently shown that for this problem one observes the curse of dimensionality (intractability) phenomenon. We aim to give the exact asymptotic expression for the information complexity.

Alexandre Tsybakov

Some methods of sparse recovery

We suggest some new methods of sparse recovery in deterministic and statistical models. Examples include problems with missing data, constrained minimization and other. We prove sparsity oracle inequalities both for the case of exact sparse model and for approximately sparse solutions.

Sara van der Geer

The incoherence condition in additive models

(Joint work with Lukas Meier and Peter Bu"hlmann)

We extend the idea of regularization using the Lasso, to the case of an additive model with p components, p being larger than the sample size n.

Our method has a group Lasso type structure, and penalizes non-smoothness of the components in the additive model. To arrive at a sparsity oracle inequality, we need an incoherence condition which generalizes the incoherence conditions used for the Lasso. Bickel et al. (2008) impose the "restricted eigenvalue assumption", which is closely related to the "compatibility condition" in van de Geer (2007), which we simply call "Condition C". We will formulate a version of such a "Condition C" for additive models. To verify it, we discuss the case of random design. We prove new results for weighted empirical processes, which make the transition from random to fixed design possible, and which only requires a population version of "Condition C". A consequence is that the sparsity oracle property of our procedure holds when the variables are independent, and that also various dependency structures are allowed.

References

Bickel, P. J., , Ritov, Y. and Tsybakov, A. (2008). Simultaneous analysis of Lasso and Dantzig selector. The Annals of Statistics, to appear .

van de Geer, S. A. (2007). The deterministic Lasso. JSM Proceedings.

Anneleen Verhasselt

Nonnegative garrote in additive models using Psplines

(With Anestis Antoniadis, Irène Gijbels)

The nonnegative garrote method was proposed as a variable selection method by Breiman (1995). In this talk we consider additive modeling and apply the nonnegative garrote method for selecting among the d independent variables. For initial estimation of the d unknown univariate functions, we use P-splines estimation (Eilers & Marx (1996)) and backfitting is applied to deal with the additive modeling. We compare the pro- posed method involving P-splines with some other methods for additive models. The finite-sample performance of the procedure is investigated via a simulation study and an illustration with real data is provided.

References

Breiman, L. (1995). Better subset regression using the nonnegative garrote, Technometrics, 37, 373384.

Eilers, P. & Marx, B. (1996). Flexible smoothing with B-splines and penalties, Statistical Science, 11, 89102.