Mathematisches Forschungsinstitut Oberwolfach

Report No. 14/2012

DOI: 10.4171/OWR/2012/14

Frontiers in Nonparametric Statistics

Organised by
Peter Bühlmann, Zürich
Tony Cai, Philadelphia
Axel Munk, Göttingen
Bin Yu, Berkeley

March 11th – March 17th, 2012

Abstract. The goal of this workshop was to discuss recent developments of nonparametric statistical inference. A particular focus was on high dimensional statistics, semiparametrics, adaptation, nonparametric bayesian statistics, shape constraint estimation and statistical inverse problems. The close interaction of these issues with optimization, machine learning and inverse problems has been addressed as well.

Mathematics Subject Classification (2000): 62G05, 62G08, 62G15, 62G20.

Introduction by the Organisers

This workshop was well attended with over 50 participants from around the world. The workshop brought together a nice blend of researchers with various backgrounds, from leading experts to Ph.D. students, from mathematical statisticians to theoretical computer scientists and applied mathematicians.

The workshop featured 27 talks covering a wide range of research problems in nonparametric statistics. A major focus was on and high-dimensional inference, including estimation of large matrices, high-dimensional signal recovery, inference in non- and semiparametric models, shape constrained and adaptive estimation and inference, statistical inverse problems and nonparametric Bayesian theory. The talks were well received and stimulated a lot of discussions among the participants. In addition to the regular talks, there were two informal evening sessions, one tutorial on casual inference and one session of Ph.D. students' talks. These experimental informal sessions turned out to be a great success. As always, the

traditional Wednesday afternoon hike to St. Roman was an enjoyable experience to all participants.

Estimation of Large Matrices Based on Noisy Data

The analysis of high-dimensional data sets, nowadays commonly arising in scientific investigations, poses many statistical challenges not present in smaller scale studies. Among the many interesting applications are genomics, fMRI analysis, risk management, and web search problems. A major part of this workshop was on the very active research area of high-dimensional statistical inference and in particular estimation of large matrices based on noisy observations.

After a brief introduction, the workshop opened on the Monday morning with an hour-long talk by Martin Wainwright (UC Berkeley) entitled "Some Recent Results in High Dimensional Statistics: from Parametrics to Nonparametrics". A major focus of his talk was on recovery of large matrices, including low-rank matrices, sparse matrices, and sparse precision matrices which are closely related to Gaussian graphical model selection. Using regularization as a common theme, Martin presented a set of new results on several interconnected problems, from high-dimensional linear regression, to matrix completion, to sparse graphical models, to nonparametric additive models. Ming Yuan (Georgia Institute of Technology) discussed adaptive estimation of covariance matrices in the high-dimensional setting. He presented a fully data-driven block thresholding estimator which attains the minimax optimal rate simultaneously over each parameter space in a large collection. The results and technical analysis reveal new features that are quite different from the conventional low-dimensional problems.

The discussions on matrix estimation continued on the Monday afternoon. Hui Zou of University of Minnesota considered the choices of the penalty function for estimation of sparse precision matrices and the related computational issues. This problem is directly connected to the sparse Gaussian graphical model selection. Cun-Hui Zhang of Rutgers followed with a talk on matrix completion using calibrated elastic regularization.

Jianqing Fan (Princeton University) considered a multi-factor model for estimating a high-dimensional covariance matrix that is a sum of a low-rank matrix and a sparse matrix. A procedure, called principal orthogonal complement thresholding, was introduced to explore the approximate factor structure. Both theoretical and numerical results were discussed. Huibin Zhou (Yale University) presented a talk focusing on the minimax lower bounds for matrix estimation, introducing a new lower bound technique particularly well suited for treating matrix estimation under the spectral norm. The result can be viewed as a generalization of Le Cam's method in one direction and Assouad's Lemma in another. Applications to optimal estimation of sparse covariance matrix, sparse precision matrix and sparse volatility matrix were discussed.

Noureddine El Karoui (UC Berkeley) reported on the properties of solutions of quadratic programs with linear constraints in the setting where parameters are estimated from noisy data. He studied the impact of distributional assumptions on

the empirical solution of the problem. His results also shed light on the behavior of the high-dimensional ridge regression estimators, as well as techniques such as regularized discriminant analysis. Angelika Rohde's (Universität Hamburg) talk focused on the accuracy of random projections of large Gaussian matrices and Zongming Ma (University of Pennsylvania) presented his latest results on sparse singular value decomposition of high-dimensional low-rank matrices based on noisy observations. He introduced an iterative thresholding estimation procedure for estimating the sparse principle components shown to be theoretically optimal and computationally efficient.

High-dimensional Statistical Inference

As already indicated above, high-dimensional statistics is a very active and ongoing field in many areas of statistics.

After the opening lecture on the Monday morning, Sara van de Geer (ETH Zürich) presented new results on regularization for structured sparsity. She introduced an elegant framework with separable regularizers and made an interesting connection to Wainwright's opening talk which covered decomposable regularizers. The framework allows for a wide variety of important sparsity structures, including e.g. overlapping groups.

Pradeep Ravikumar (University of Texas, Austin) highlighted the importance of Bregman divergence leading to a computationally tractable loss function for single- and multi-index models. His contribution combined "classical nonparametrics" (projection pursuit regression and backfitting) with a new and modern viewpoint where convex optimization replaces notoriously difficult optimization in projection pursuit regression (or neural networks).

Nicolai Meinshausen (Oxford University) impressively demonstrated that qualitative constraints are sufficient for effective regularization in high-dimensional problems. Examples include non-negative matrix factorization and he mainly focused on non-negative least squares in regression. Advantages were high-lighted by practical problems and by demonstrating the issue that no regularization (tuning) parameter is required.

Much discussion was generated by a session on causal inference and dynamic treatment/intervention analysis. Marloes Maathuis (ETH Zurich) focused on the problem of high-dimensional inference for bounds of causal effects, based on observational data: the methods were demonstrated on real applications from biology. Susan Murphy (University of Michigan, Ann Arbor) highlighted major challenges and new problems in the area of dynamic treatment schemes which are of eminent importance in medical research. The field is closely connected to causal inference since a treatment is an active intervention and the effect of the latter is called a causal effect. James Robins (Harvard University) presented novel and sharp results on identifiability of causal effects.

Nonparametric Bayesian Inference

Nonparametric Bayesian analysis has been recently proved to be a very powerful and necessary tool for understanding widely but frequently used ad hoc Bayesian computational methods. Larry Brown (University of Pennsylvania) presented a new class of empirical Bayes estimators for the Poisson decision problem improving remarkably on Robbin's estimator by employing a "convolution" trick. Judith Rousseau (Paris Dauphine) discussed adaptation for different loss functions in a Bayesian setting. She gave counterexamples in semiparametric models to the common adaptation results using a hierarchical prior. Volodia Spokoiny (WIAS Berlin) presented a Bernstein von Mises theorem on the asymtotic posterior under model misspecification. Aad van der Vaart (VU Amsterdam) gave an entirely new insight into the practically important question of asymptotic coverage by Bayesian credible sets. He presented a number of remarkable results showing that the rate of contraction of the posterior depends on the fine properties of the Gaussian prior.

Non and Semiparametric Inference and Applications

In high dimensional models, structural constraints such as additive modeling allow for improved estimation and inference. This issue has been highlighted by Enno Mammen (Universität Mannheim), who considered a generalized varying nonparametric coefficient model where division of the covariates into two groups is not required and interaction terms between all covariates can be included. Rate optimal estimators are constructed and kernel type-estimators are given as the solution of a system of nonlinear integral equations, which provides a suprising link to inverse problems. Hans Georg Müller (UC Davies) introduced new functional volatility processes to model financial volatility and extended principal component analysis for functional data to these processes. Ji Zhu (University of Michigan) suggested a semiparametric framework to estimate parameters of a dynamical system given by a system of ordinary differential equations. Holger Dette (Universität Bochum) introduced a new concept for spectral density estimation for time series based on copulae, generalizing classical methods restricted to covariances, hence relying on normality to some extent. Wolfgang Polonik (UC Davies) proposed novel methods to estimate filamentary structures, using bump hunting techniques linking this to multiscale methods. Finally, he discussed several interesting applications.

Statistical Inverse Problems, Adaptive Confidence Sets and Shape Constrained Inference

A major lesson in the last years understood by the community is that shape constraint estimation and inference, in particluar for difficult statistical inverse problems, allows to overcome the burden of slow minimax convergence rates and non adaptation of confidence sets. Consequently, this has recently initiated much research recently, in particular on computational and inferential aspects. Otmar

Scherzer (Universität Wien) gave a one-hour overview talk on variational methods for solving linear inverse problems and he highlighted interesting links to a variety of statistical inverse problems. In particular, he focused on mathematical imaging and related it to graph cut algorithms on discrete grids. Lutz Dümbgen (Universität Bern) discussed asymptotic results and computational strategies for log concave shape contraint density estimation, in particular for censored data. He introduced bi-log concavity as a new concept to overcome the burden of unimodality for log concave estimators. Linking shape constraints with statistical inverse problems Johannes Schmidt-Hieber (VU Amsterdam) introduced a new multiscale technique to extract qualitaive information, e.g. on the number of modes or points of inflection from deconvolution problems. Klaus Frick (Universität Göttingen) extended the Dantzig estimation method, well known in high dimensional statistics, to multiscale constraints, showing its usefulness for signal detection and imaging recovery problems. He demonstrated that the multiscale Dantzig estimator leads to a unifying concept covering many known and new estimators. Günther Walther's (Stanford University) talk was related to this and he surveyed and compared several multiscale methods and highlighted advantages of the average likelihood ratio statistics. Richard Nickl (Cambridge University) discussed necessary and sufficient conditions for the existence of honest confidence sets in adaptative function estimation problems, this being based on a sharp analysis of certain minimax testing problems.

On the Thursday evening the "New Researchers Session" provided the unique opportunity for Ph.D. students to discuss their current Ph.D. projects with worldwide recognized experts. Within this new format short talks had been given by Yin Xia (University of Pennsylvania) on "Some High Dimensional Hypothesis Testing Problems", Till Sabel (Universität Göttingen) on "Lower Bounds in Estimation of Scale Parameters in Stationary Gaussian Time Series", Diego Colombo (ETH Zürich) on "Learning Causal Information with Hidden Variables", Hannes Sieling (Universität Göttingen) on "Jump Penalized Regression in Exponential Families" and Alain Hauser (ETH Zürich) on "Causal Inference from Interventional Data". This session has been perceived as particularly fruitful by the young researchers as well as the senior scientists and we very much hope that this will serve as a role model for the future Oberwolfach events.

Workshop: Frontiers in Nonparametric Statistics

Table of Contents

Abstracts

Poisson Compound and Empirical Bayes Estimation, Revisited

Lawrence Brown
(joint work with Eitan Greenshtein,Yaacov Ritov)

We investigate a classical non-parametric Poisson empirical Bayes estimation problem and propose an estimator that performs better than the original proposal of Robbins (1955).

Begin with independent Poisson obervations, $Y_i \sim Po(\lambda_i), i = 1, \ldots, p$. Consider the standard decision theoretic estimation problem. Estimate the vector $\lambda = (\lambda_1, \ldots, \lambda_p)$ by $\delta = \delta(Y)$. Consider the average quadratic risk $R(\delta, \lambda) = E_\lambda(p^{-1}\|\delta - \lambda\|^2)$. For a prior distribution, G, the expected risk is denoted by $R(G, \delta) = E_G(R(\Lambda, \delta))$. The Bayes procedure $\delta_G(y) = E(\Lambda|Y = y)$ (with the conditional expectation taken coordinate-wise). The Bayes risk is $B(G) = R(G, \delta_G) = \min_\delta R(G, \delta)$.

Here is the classical empirical Bayes estimator proposed in [1]. Here, G is unknown. The goal is to find an estimator $\tilde{\delta}$ that approximates δ_G sufficiently well so that $R(G, \tilde{\delta}) - B(G)$ is "small" uniformly in G as $p \to \infty$. In this setting, "small" can mean $o(1)$ or sometimes something even smaller, if possible.

The approach we take is consistent with Robbins' original empirical Bayes proposal. Write δ_G as a functional of the marginal distribution $P_G(y) = \int Po_\lambda(y)G(d\lambda)$, i.e., $\delta_G = \Delta(P_G)$. Then use the sample $Y = Y_1, \ldots, Y_p$ to estimate P_G by, say, $\tilde{P}$, and δ_G by $\tilde{\tilde{\delta}} = \Delta(\tilde{P})$. In his paper, Robbins took such an approach. He observed that if G is known then the Bayes estimator can be written as

$$\delta_G(y) = \frac{\int \lambda Po(y|\lambda)G(d\lambda)}{\int Po(y|\lambda)G(d\lambda)} = \frac{\int (y+1)Po(y+1|\lambda)G(d\lambda)}{\int Po(y|\lambda)G(d\lambda)} = \frac{(y+1)P_G(y+1)}{P_G(y)}.$$

Note that this is a function of the marginal distribution P_G. Summarize the observed sample by $Z_Y = \{\mathbb{N}_Y(k)\}$ where $\mathbb{N}_Y(k) = \#\{Y_i : Y_i = k\}$. Then a natural estimator of P_G is $\hat{P}(y) = Z_y/p$. This suggests the following empirical Bayes estimator, which is known as Robbins' estimator for this problem:

$$\hat{\tilde{\delta}}(k) = \frac{(k+1)\hat{P}(k+1)}{\hat{P}(k)} = \frac{(k+1)\mathbb{N}_{Y+1}(k+1)}{\mathbb{N}_Y(k)}.$$

It is clear that for any fixed G and each y, $\hat{P}(y) \to P_G(y)$ as $p \to \infty$ and $\hat{\tilde{\delta}}(y) \to y$. However, there are some serious problems with $\hat{\tilde{\delta}}$:

Problem 1: If $\mathbb{N}(k+1) > 0$ but $\mathbb{N}(k) = 0$ (or is small) then $\hat{\tilde{\delta}}(Y_i = k) = \infty$ (or is probably not desirably accurate).

Problem 2: Any Bayes estimator is monotone in y, but $\hat{\tilde{\delta}}$ is not.

Problem 3: (a subcase of Problem 2) At $y_{(p)} = \max\{y_i\}$ we have $\hat{\tilde{\delta}}(y_{(p)}) = 0$.

The remainder of the construction is devoted to modifying the estimator so as to remedy these problems.

To address Problem 1, pick a small $h > 0$ (called the "corruption" parameter). Let $Q \sim Po(h)$. Choices for h in the range $0.5 \leq h \leq 3$ seem to work well. We will later propose a cross-validation step to choose h. Let $Z = Y + Q$. Use $\mathbb{N}_Y(k)$ as a basis for estimating the marginal distribution of Z. The estimate is $\tilde{P}_Z(z) = \sum_{j=0}^{z} \frac{\mathbb{N}(j)}{p} \frac{h^{z-j}e^h}{(z-j)!}$. Now generate $Q_i \sim_{iid} Po(h)$ and build a corrupted sample $\{Z_i\}$ with $Z_i = Y_i + Q_i$. (Each $Z_i \sim Po(\lambda_i + \boldsymbol{h})$.) Apply Robbins' method to estimate λ_i from this sample via $\tilde{\tilde{\delta}}_{h;1}(z) = \frac{(z+1)\tilde{P}_Z(z+1)}{\tilde{P}_Z(z)} - \boldsymbol{h}$. It is easily checked that $\tilde{\tilde{\delta}}_{h;1}(z) > 0$ for all $z \geq y_{(1)}$. So define $\tilde{\tilde{\delta}}_{h;1}(z) = 0$ for all $z < y_{(1)}$. This guarantees Problem 1 does not happen. However, $\{\tilde{\tilde{\delta}}_{h;1}(z_i)\}$ is a randomized estimator, since $Z_i = Y_i + Q_i$. Such estimators can be improved. To do so, Rao-Blackwellize. Let $\tilde{\tilde{\delta}}_{h;2}(y) = E^Q(\tilde{\tilde{\delta}}_{h;1}(y+Q)) = \sum \frac{h^j e^{-h}}{j!} \tilde{\tilde{\delta}}_{h;1}(y+j)$. The random Q_i have now disappeared. The estimator $\{\tilde{\tilde{\delta}}_{h;2}(y_i)\}$ is a closed-form function of $\{Y_j\}$ through the sufficient statistics $\{\mathbb{N}_Y(k); k = 0, \ldots\}$.

Problem 2 usually persists – $\tilde{\tilde{\delta}}_{h;2}(y)$ need not be monotone in y. So we monotone-ize $\tilde{\tilde{\delta}}_{h;2}$. As a convenient, but rather ad-hoc method, we use the Pool-Adjacent-Violators algorithm developed for least-squares isotonic regression. Koenker and Mizra (unpublished) have proposed a more principled and likely better method that appears to still be computationally feasible. It can be verified that so long as h is not too small, this should also fix any remnant of Problem 3. Call the resulting monotone-ized estimator Δ_h. It remains only to choose the corruption parameter, h. One plausible possibility that generally works well on examples is to directly choose a moderate value of h – say $1 \leq h \leq 3$. A more interesting and flexible choice involves what we call "*inbred* cross-validation":

Let $p < 1$ but not too far from 1. Let $B_i \sim_{ind} Bin(Y_i, p)$. Let $U_i = B_i$ and $V_i = Y_i - U_i$. This yields $U_i \sim Po(p\lambda_i)$, $V_i \sim Po((1-p)\lambda_i)$ and $U_i \perp V_i$. Then use Δ_h on the sample $\{U_i\}$ to estimate $\{p\lambda_i\}$, and use cross-validation on the smaller sample $\{V_i\}$ to choose h. The estimates of $\{p\lambda_i\}$ can be adjusted to estimate $\{\lambda_i\}$. It is possible to also use an additional Rao-Blackwell step here to further improve the estimator, but we did not do so for simulations that we have reported.

Asymptotics of Robbins' method are quite appealing. But simulations we have performed show that actual performance in examples can be quite suboptimal. Here is a slightly informal statement of a theorem we have proved.

If $\{G_k\}$ is a sequence of priors on a bounded set that does not concentrate at a single point then a rate-sharp bound is

$$R(G_k, \hat{\delta}) - B(G_k) = O\Big(\frac{(\log p)^2}{\log\log p}\Big).$$

This is not much different from $(\log p)^2$, and so seems a pretty desirable convergence rate. But behavior in finite (not too large) samples can be much worse than this suggests, as revealed by simulations we have performed for a variety of

examples. For $p = 200$ and G supported within $[0, 20]$, Robbins' estimator can be worse than Δ_h by 5-35% in terms of squared error risk, depending on the form of G.

References

[1] H. Robbins, *An Empirical Bayes Approach to Statistics*, Proc. Third Berkeley Symp. on Prob. Statist. (1955), 157–164.

Of Copulas, Quantiles, Ranks and Spectra: An L_1-approach to spectral analysis

Holger Dette

(joint work with Marc Hallin, Tobias Kley, Stanislav Volgushev)

In this talk we presented a new method to overcome the limitations of conditional location-scale modeling, and to provide statistical tools for a new approach to time series modeling. The traditional nonparametric techniques, such as spectral analysis (in its usual L_2-form), which only account for second-order serial features, cannot handle such objects, and we therefore propose and develop an original, flexible and fully nonparametric L_1-spectral analysis method.

While classical spectral densities are obtained as Fourier transforms of classical covariance functions, we rather define spectral density *kernels*, associated with covariance *kernels* of the form

$$\gamma_k(x_1, x_2) := \mathrm{Cov}(I\{Y_t \le x_1\}, I\{Y_{t-k} \le x_2\}) \tag{1}$$

(Laplace cross-covariance kernels) or

$$\gamma_k(\tau_1, \tau_2) := \mathrm{Cov}(I\{U_t \le \tau_1\}, I\{U_{t-k} \le \tau_2\}) \tag{2}$$

(copula cross-covariance kernels), where $U_t := F_Y(Y_t)$ and F_Y denotes the marginal distribution of the strictly stationary process $\{Y_t\}_{t\in\mathbb{Z}}$. Contrary to covariance functions, the *kernels* $\{\gamma_k(x_1, x_2)|x_1, x_2 \in \mathbb{R}\}$ and $\{\gamma_k(\tau_1, \tau_2)|\tau_1, \tau_2 \in [0, 1]\}$ allow for a complete description of arbitrary bivariate distributions for the couples (Y_t, Y_{t-k}) and arbitrary bivariate copulas of the pairs (U_t, U_{t-k}), respectively, and thus escape the conditional location-scale paradigm. They are able to account for sophisticated dependence features that covariance-based methods are unable to detect, such as time-irreversibility, tail dependence, varying conditional skewness or kurtosis, etc. Special virtues, such as invariance/equivariance (with respect to continuous order-preserving marginal transformations), can be expected from the copula covariance kernels defined in (2).

Classical nonparametric spectral-based inference methods have proven quite effective [Granger (1964), Bloomfield (1976)], essentially in a Gaussian context, where dependencies are fully characterized by autocovariance functions. Therefore, it can be anticipated that similar methods, based on estimated versions of Laplace or copula spectral kernels (associated with Laplace and copula covariance kernels,

respectively) would be quite useful in the study of series exhibiting those features that classical covariance-related spectra cannot account for.

Estimation of Laplace and copula spectral kernels, however, requires a substitute for the *ordinary periodogram* concept considered in the classical approach. We therefore introduce Laplace and copula *periodogram kernels*. While ordinary periodograms are defined via least squares regression of the observations on the sines and cosines of the harmonic basis, our periodogram kernels are obtained via quantile regression in the [Koenker and Bassett, (1978)] sense. A study of their asymptotic properties shows that, just as ordinary periodograms, they produce asymptotically unbiased estimates (more precisely, the mean of their asymptotic distribution is 2π times the corresponding spectrum), and we therefore also consider smoothed versions that yield consistency. Asymptotic results show that copula periodograms, as anticipated, are preferable to the Laplace ones, as their asymptotic behavior only depends on the bivariate copulas of the pairs (U_t, U_{t-k}), not on the (in general unknown) marginal distribution F_Y of the Y_t's.

Unfortunately, copula periodogram kernels are not statistics, since their definition involves the transformation of Y_t into U_t, hence the knowledge of the marginal distribution function F_Y. We therefore introduce a third periodogram kernel, based on the empirical version of F_Y, that is, on the *ranks* of the random variables $Y_1, \ldots, Y_n$, and establish, under mild assumptions, the asymptotic equivalence of that rank-based Laplace periodogram with the copula one. Smoothed rank-based Laplace periodogram kernels, accordingly, seem to be the adequate tools in this context. We conclude with a brief numerical illustration – simulations and an empirical application – of their potential use in practical problems.

References

[1] P. Bloomfield, *Fourier Analysis of Time Series: An Introduction*, Wiley, New York (1976).
[2] H. Dette, M. Hallin, T. Kley, S. Volgushev *Of copulas, quantiles, ranks and spectra: An L_2-approach to spectral analysis*, arXiv: 1111.7205v1 (2011).
[3] C. W. Granger, *Spectral Analysis of Economic Time Series*, Princeton University Press, (1964).
[4] R. Koenker and G. Bassett, *Regression Quantiles*, Econometrica **46** (1978), 33–50.

Shape-constraints for i.i.d. and censored data

Lutz Dümbgen

(joint work with Petro Kolesnyk, Kaspar Rufibach, Richard Samworth, Dominic Schuhmacher)

In the first part we discuss approximation of distributions Q on $\mathbb{R}^d$ by log-concave densities $f = f(\cdot \,|\, Q)$. This means that $f(\cdot \,|\, Q)$ maximizes $\int \log f \, dQ$ over all probability densities f such $\log f$ is concave. As shown by Dümbgen et al. (2011), $f(\cdot \,|\, Q)$ is well-defined if, and only if, Q has finite first absolute moment and is not supported by a hyperplane in $\mathbb{R}^d$. Moreover, the mapping $Q \mapsto f(\cdot \,|\, Q) \in L^1(\mathbb{R}^d)$ is continuous with respect to Wasserstein distance (i.e. weak convergence plus

convergence of first absolute moments). Explicit algorithms to compute $f(\cdot \mid Q)$ in case of discrete distributions Q with finite support are provided by Dümbgen and Rufibach (2011) and Cule et al. (2010).

In the second part we show how to adapt this approach to arbitrarily censored event times. It turns out that a specific version of the EM algorithm yields satisfactory estimators of an unknown event time distribution on $(0, \infty]$, often superior to traditional nonparametric maximum-likelihood estimators. (This part is based on joint work in progress of Dümbgen, Rufibach and Schuhmacher.)

In the last part we introduce a weaker shape constraint: A distribution function F is called bi-log-concave if both $\log F$ and $\log(1 - F)$ are concave. This restriction allows distributions with arbitrarily high modality. We present equivalent characterizations of this shape-constraint. While maximum-likelihood estimation in this class seems to be difficult and may even be impossible, one can combine this shape-constraint with traditional nonparametric confidence regions, e.g. Kolmogorov-Smirnov confidence bands. It turns our that this leads to rather informative and honest confidence regions for F and functionals of F such as moments of arbitrary order. (This part is pased on joint work in progress of Dümbgen and Kolesnyk.)

References

[1] M. L. Cule, R. J. Samworth and M. I. Stewart, *Maximum likelihood estimation of a multi-dimensional log-concave density (with discussion)*, Journal of the Royal Statistical Society, Ser. B **72** (2010), 702–730.

[2] L. Dümbgen, R. Samworth and D. Schuhmacher, *Approximation by log-concave distributions, with applications to regression*, Annals of Statistics **39**(2) (2011), 702–730.

[3] L. Dümbgen and K. Rufibach, Journal of Statistical Software **39**(6) (2011).

Some questions in high-dimensional statistics

Noureddine El Karoui

The talk was concerned with problems in high-dimensional statistics, i.e the setting where we observe n vectors, X_i, in dimension p and p/n has a finite non-zero limit as $n \to \infty$. Specifically, I considered the problem of understanding the properties of solutions of quadratic programs with linear constraints, when parameters are estimated from data.

Rather than making structural assumptions (e.g sparsity in one sense or another) on those parameters, I studied the impact of distributional assumptions on the X_i's on the empirical solution of the problem. I considered the case of elliptical data - which includes the normal distribution as a subcase. The analysis reveals that one can for instance very well estimate the optimal value of the problem, even though naive estimates are quite biased. In a risk management context, it is also possible to predict accurately the future risk of such as an estimate. The solution depends non-trivially on the ellipticity of the data, which is a proxy for its geometry.

I presented extensions for situations involving penalized estimates of covariance, under very weak distributional assumptions. In joint work with Holger Koesters (Bielefeld), we found deterministic equivalents for all the random quantities appearing in the problem. The work also naturally sheds light on the behavior of ridge regression estimators in high-dimension, as well as techniques such as regularized discriminant analysis.

Finally, I discussed briefly regression M-estimates. With techniques and ideas similar to the ones employed above, it is possible to understand their risk (at the time of the talk, this understanding was not yet fully mathematically rigorous). In the case where $p < n$, it is also possible to optimize over the objective function to find the best performing estimator, in the case of Gaussian design. Interestingly, this optimal objective function depends in general of the dimension and of course the distribution of the errors. It should be noted that it does not coincide with natural objective functions coming from maximum-likelihood ideas, which yield suboptimal estimators in this context, despite the well-known fact that they are optimal when $p/n \to 0$. This part was based on joint works with Peter Bickel, Bin Yu and our students Derek Bean and Chingwhay Lim.

Principal Orthogonal Complement Transformation

Jianqing Fan

This paper deals with estimation of high-dimensional covariance with a conditional sparsity structure, which is the composition of a low-rank matrix plus a sparse matrix. By assuming sparse error covariance matrix in a multi-factor model, we allow the presence of the cross-sectional correlation even after taking out common but unobservable factors. We introduce the Principal Orthogonal complEment Thresholding (POET) method to explore such an approximate factor structure. The POET estimator includes the sample covariance matrix, the factor-based covariance matrix (Fan, Fan, and Lv, 2008), the thresholding estimator (Bickel and Levina, 2008) and the adaptive thresholding estimator (Cai and Liu, 2011) as specific examples. We provide mathematical insights when the factor analysis is approximately the same as the principal component analysis for high dimensional data. The rates of convergence of the sparse residual covariance matrix and the conditional sparse covariance matrix are studied under various norms, including the spectral norm. It is shown that the impact of estimating the unknown factors vanishes as the dimensionality increases. The uniform rates of convergence for the unobserved factors and their factor loadings are derived. The asymptotic results are also verified by extensive simulation studies.

Multiresolution Dantzig Estimation for Imaging and Signal Detection

Klaus Frick
(joint work with Philipp Marnitz, Axel Munk, Hannes Sieling)

In many applications the relation of observable data Y (that is assumed to be given on a grid $G = \{1, \dots, n\}^d$) and an underlying, unknown signal u^0 can be modelled as an inverse regression problem: We assume that $u^0 \in U$ for some suitable model space U and that independently

$$Y_\nu \sim P_{(Ku^0)_\nu} \quad \nu \in G. \tag{1}$$

Here $K : U \to \mathbb{R}^G$ is a linear operator that is assumed to model data acquisition and sampling at the same time. Further, we assume that $\mathcal{P} = \{P_\theta\}_{\theta \in \Theta}$ is some one-dimensional standard exponential family of distributions.

In practical situations, the signal u^0 exhibits features of different scales and modality, as for example images contain constant and smooth portions as well as oscillating patterns both of different sizes. Thus, a minimum requirement for modern reconstruction methods is to allow for such spatially varying characteristics. In this talk, we will introduce an estimation paradigm that meets this requirement.

To be more precise, we study *multiresolution Dantzig estimators (MDE)* that are solutions of the following minimization problem

$$\inf_{u \in U} J(u) \quad \text{subject to} \quad \max_{S \in \mathcal{S}} T_S(Y, Ku) \le q. \tag{2}$$

Here $\mathcal{S}$ is some family of subsets of the grid G and $T_S(Y, \theta)$ is a likelihood-ratio type statistic restricted to the set $S \in \mathcal{S}$. The functional $J : U \to \mathbb{R}$ is supposed to encode some notion of cost and q is a threshold that balances costs against data-fit.

Put differently, the constraint in (2) selects those estimators $\hat{u}$ for which the restricted data $\{Y_\nu : \nu \in S\}$ is well described by $\hat{u}$ *on all sets* $S \in \mathcal{S}$ according to a local likelihood-ratio criterion. Among these estimators we then pic the most parsimonious by minimizing J. Hence, the *multiresolution (MR) statistic*

$$T(Y, \cdot) = \max_{S \in \mathcal{S}} T_S(Y, \cdot)$$

is sensitive to local violations of the hypotheses that the data Y is generated by Ku^0, where the spatial resolution for detecting violations is governed by the system $\mathcal{S}$. A popular choice for the system $\mathcal{S}$, for example, is the set of all d-dimensional cubes contained in the grid G (cf. [9])

It is important to note, that numerous well established estimation methods are covered by our SMRE framework (see for example [4, 2]). A further prominent example is the *Dantzig selector* as introduced in [3]. In a Gaussian model with $d = 1$, $U = \mathbb{R}^p$ and $p \gg n$ the Dantzig selector is defined as the solution of

$$\inf_{u \in U} \|u\|_1 \quad \text{subject to} \quad \max_{1 \le i \le p} |K^T(Ku - Y)_i| \le q$$

Setting $J(u) = \|u\|_1$ and $T_i(Y, \theta) = |K^T(\theta - Y)_i|$ gives (2). The system $\mathcal{S}$ consists of all singletons in $\{1, \dots, p\}$ which reveals the uni-scale nature of the Dantzig

selector. This justifies the name multiresolution Dantzig estimators for solutions of (2).

In this talk we study the performance of the above estimation paradigm for two particular problems:

Changepoint estimation. We describe how the above paradigm can be employed for model selection and estimation in one-dimensional changepoint problems, generalizing the work in [1]. Put differently, we assume that U is the set of all piecewise constant functions $u : [0,1] \to \Theta$ and $(Ku)_i = u(i/n)$. Assume that $u^0 \in U$ has $N \in \mathbb{N}$ jumps and that the data $(Y_1, \dots, Y_n)$ is given by (1).

Firstly, we estimate N by the minimal number $\hat{N}$ such that $T(Y, Ku)$ is finite for some u with $\hat{N}$ jumps. Then, we compute a solution of (2) with J being the negative log-likelihood function restricted to the functions with $\hat{N}$ jumps. We discuss how to choose the threshold $q = q_n$ and prove nearly optimal convergence rates for $\hat{u}$ as $n \to \infty$. Additionally, we propose a modified version of the dynamic programming algorithm [6] for the efficient solution of (2).

Image reconstruction. Here, we aim for the reconstruction of an unknown (gray-valued) image $u^0 \in U = L^2([0,1]^d)$ from the data in (1) by computing a solution of (2). We restrict our considerations on *convex* cost functionals J, as for example the total variation semi-norm that is well known to foster smooth solutions while preserving edges. We propose a general algorithmic framework for the solution of (2) with convex costs J based on the combination of the *alternating direction method of multipliers (ADMM)* with Dykstra's projection method (cf. [7]). We illustrate the applicability of our approach for various imaging examples, including deconvolution problems in Poisson models arising in fluorescence microscopy [8].

References

[1] L. Boysen, A. Kempe, V. Liebscher, A. Munk, and O. Wittich. Consistencies and rates of convergence of jump-penalized least squares estimators. *Ann. Statist.*, 37(1):157–183, 2009.

[2] E. Candès and F. Guo. New multiscale transforms, minimum total variation synthesis: Applications to edge-preserving image reconstruction. *Signal Processing*, 82:1519–1543, 2002.

[3] E. Candès and T. Tao. The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.*, 35(6):2313–2351, 2007.

[4] P. L. Davies, A. Kovac, and M. Meise. Nonparametric regression, confidence regions and regularization. *Ann. Statist.*, 37(5B):2597–2625, 2009.

[5] L. Dümbgen and V. G. Spokoiny. Multiscale testing of qualitative hypotheses. *Ann. Statist.*, 29(1):124–152, 2001.

[6] F. Friedrich, A. Kempe, V. Liebscher, and G. Winkler. Complexity penalized M-estimation: fast computation. *J. Comput. Graph. Statist.*, 17(1):201–224, 2008.

[7] K. Frick, P. Marnitz, and A. Munk. Statistical multiresolution Dantzig estimation in imaging: Fundamental concepts and algorithmic framework. *Electron. J. Stat.*, 6:231–268, 2012.

[8] S. W. Hell. Far-Field Optical Nanoscopy. *Science*, 316(5828):1153–1158, 2007.

[9] D. Siegmund and B. Yakir. Tail probabilities for the null distribution of scanning statistics. *Bernoulli*, 6(2):191–213, 2000.

The geometry of adaptive confidence sets

Richard Nickl

We give general sets of necessary and sufficient conditions for the existence of honest confidence sets in adaptative function estimation problems. We give results for L^2-confidence balls as well as L^∞ confidence bands, using a sharp analysis of certain minimax testing problems, generalising work of Ingster and others to the situation relevant here. We highlight the subtle dependence of our existence results on the geometry of the given adaptation problem, and discuss various consequences for the theory of statistical inference in such models.

The main results are of the following flavour:

Let $\Sigma(s)$ be a Hölder ball. For $s > r$ define

$$\tilde{\Sigma}(r, \rho_n) := \left\{ f \in \Sigma(r) : \inf_{g \in \Sigma(s)} \|g - f\|_\infty \geq \rho_n \right\}$$

where $\rho_n \geq 0$. We are thus removing those functions from $\Sigma(r)$ that are not separated away from $\Sigma(s)$ in sup-norm by at least ρ_n.

Can we find a honest confidence band over

$$\mathcal{P}(\rho_n) := \Sigma(s) \cup \tilde{\Sigma}(r, \rho_n)$$

that is also adaptive in the sense that

$$f \in \Sigma(s) \Rightarrow E_f |C_n| \leq L r_n(s) \text{ and}$$
$$f \in \tilde{\Sigma}(r, \rho_n) \Rightarrow E_f |C_n| \leq L r_n(r)$$

where $r_n(s)$ is the optimal rate of estimation in the given Hölder ball?

$\rightarrow$ Ideally $\rho_n = 0$, but the following result shows that this is impossible, and moreover characterises the optimal admissible choice for ρ_n.

Theorem 1. *(Hoffmann, Nickl, 2011) Let $s > r > 0$ and B be given. An adaptive and honest confidence band over*

$$\Sigma(s) \cup \tilde{\Sigma}(r, \rho_n)$$

exists if and only if ρ_n exceeds the minimax rate of testing between the hypotheses

$$H_0 : f_0 \in \Sigma(s) \text{ and } H_1 : f_0 \in \tilde{\Sigma}(r, \rho_n);$$

and this rate is

$$\rho_n \simeq r_n(r) = \left(\frac{\log n}{n} \right)^{\frac{r}{2r+1}}.$$

More precisely:

($\Leftarrow$) *If C_n is a confidence band that is adaptive and honest with level $\alpha < 0.5$ over $\Sigma(s) \cup \tilde{\Sigma}(r, \rho_n)$, then*

$$\liminf_n \frac{\rho_n}{r_n(r)} > 0.$$

($\Rightarrow$) *Suppose B is known and $0 < \alpha < 1$ is given. Then there exists a sequence ρ_n satisfying*

$$\limsup_n \frac{\rho_n}{r_n(r)} < \infty$$

and a confidence band C_n that is honest with level α and adaptive over $\Sigma(s) \cup \tilde{\Sigma}(r, \rho_n)$.

When one separates in L^2, the results change, and show the dependence on the geometry.

Consider again adaptation to a fixed submodel $\Sigma(s)$ with $s > r$, and define

$$\tilde{\Sigma}(r, \rho) \equiv \{f \in \Sigma(r) : \|f - \Sigma(s)\|_2 \geq \rho\},$$

the only difference being that we now 'remove in L^2'.

Instead of asking for the maximal sup-norm diameter of the confidence set to shrink at the optimal rate, we now weaken this requirement and only require the L^2-diameter to shrink at the optimal rate. (FDR-idea: coverage only for sufficiently many points). For a random subset C of L^2 define

$$|C|_2 = \inf \left\{\rho : C \subset \{h : \|h - g\|_2 \leq \rho\}, g \in L^2\right\}.$$

Theorem 2 (Bull and Nickl, 2011)**.** *An L^2 - adaptive and honest confidence set over*

$$\Sigma(s) \cup \tilde{\Sigma}(r, \rho_n)$$

exists if and only if ONE of the following conditions holds true.
a) $s \leq 2r$ and $\rho_n = 0$ $\forall n$,
b) $s > 2r$ and ρ_n exceeds the minimax rate of testing between

$$H_0 : f_0 \in \Sigma(s) \text{ and } H_1 : f_0 \in \tilde{\Sigma}(r, \rho_n);$$

and this rate is

$$\rho_n \simeq n^{-\frac{r}{2r+1/2}}.$$

One can also study the continuous adaptation problem as in Giné and Nickl (2010), and new results on such settings are obtained as well.

References

[1] Hoffmann, M. and Nickl, R. *On adaptive inference and confidence bands.*, Annals of Statistics **39** (2011), 2383-2409
[2] Bull, A. and Nickl, R., *Adaptive confidence sets in L^2*, preprint (arxiv).
[3] Giné, E., Nickl, R. *Confidence bands in density estimation*, Annals of Statistics **38** 1122-1170

Singular Value Decomposition for High-Dimensional Data

Zongming Ma
(joint work with Dan Yang, Andreas Buja)

Singular value decomposition (SVD) is widely used in multivariate analysis for dimension reduction, data visualization, data compression and information extraction in such fields as genomics, imaging, financial markets, etc. However, when used for statistical estimation in high-dimensional low rank matrix models, singular vectors of the noise-corrupted matrix are inconsistent for their counterparts of the true mean matrix (Shabalin and Nobel, 2010). To achieve consistency in estimation and better interpretability, in addition to low-rankness, we further assume that the true singular vectors have sparse representations in a certain basis.

Sparse SVD in high dimensions has been studied by several recent papers. Witten et al. (2009) introduced penalized matrix decomposition which constrains the ℓ_1 norm of the left and right singular vectors to impose sparsity on the solutions. Lee et al. (2010) used penalized least squares for rank-one matrix approximations with ℓ_1 norms of the singular vectors as additive penalties. Both papers focus on obtaining the first pair of singular vectors. The subsequent pairs are then obtained by repeating the same procedure on the residual matrices. This may cause non-identifiability and non-orthogonality issues, and theoretical properties of resulting estimators are not well understood.

The goal of this work is to provide a theoretically optimal and computationally efficient solution to the high dimensional SVD problem. In particular, we propose an iterative thresholding estimation procedure, which has the following distinctive features. First, it does not involve any optimization criterion and is based on a simple matrix computation method. Second, it estimates the subspaces spanned by the leading singular vectors simultaneously as well as the true mean matrix, as opposed to the previous one-pair-at-a-time methods. Hence, it yields orthogonal sparse singular vectors. iterative procedure, our method even outperforms the classical SVD in terms of speed. Fourth, simulation results also show that it has competitive finite sample performance. Last but not least, under normality assumption, the resulting estimators achieve near optimal minimax rates of convergence and adaptivity. We further turn the algorithm into a practical methodology that is fast, data-driven and robust to heavy-tailed noises (Yang et al., 2011).

We now lay out the model assumptions. To start with, we assume the data matrix is the sum of signal and fully exchangeable noise:

$$X = M + Z\,. \tag{1}$$

In (1), the signal matrix $M = (m_{ij})$ is of dimension p_u by p_v and has a multiplicative low-rank structure: $M = UDV' = \sum_{l=1}^{r} d_l \mathbf{u}_l \mathbf{v}_l'$, where $d_1 \geq \cdots \geq d_r$ and the singular vector matrices $U = [\mathbf{u}_1, \ldots, \mathbf{u}_r], V = [\mathbf{v}_1, \ldots, \mathbf{v}_r]$ are both deterministic. The rank $r \ll \min(p_u, p_v)$ is assumed to be fixed in later asymptotic analysis and known throughout. Moreover, the noise matrix $Z = (z_{ij})$ consists of i.i.d. errors as its components. For theoretical analysis, we assume that the z_{ij}'s are i.i.d. $N(0, 1)$.

Furthermore, we assume that the singular vectors are sparse and use the notion of weak ℓ_q ball to quantify the sparsity level. For any p-vector $\mathbf{u}$, we say that $\mathbf{u}$ belongs to the weak ℓ_q ball of radius s, denoted by $\mathbf{u} \in wl_q(s)$, if $|\mathbf{u}|_{(i)} \leq si^{-1/q}$. For $0 < q < 2$, the above condition implies rapid decay of the ordered coefficients of $\mathbf{u}$, and hence sparsity. Altogether, for $q_u, q_v \in (0, 2)$, we focus on parameter spaces characterized by parameters (s_u, q_u, s_v, q_v) as the following:

$$\begin{aligned}\Theta(s_u, q_u; s_v, q_v) = \{M = UDV' : U'U = I_r, V'V = I_r,\\ D = \mathrm{diag}(d_1, \ldots, d_r) > 0,\\ \mathbf{u}_l \in wl_{q_u}(s_u), \mathbf{v}_l \in wl_{q_v}(s_v)\}\,,.\end{aligned} \tag{2}$$

Given the model assumptions, our goal is to estimate the subspaces spanned by the leading left and right singular vectors $span(U), span(V)$ and/or to recover the low rank mean matrix M. For estimating singular subspaces, we use the loss function $L_U(\hat{U}, U) = \|P_{\hat{U}} - P_U\|_2^2$, where $P_U, P_{\hat{U}}$ are the projection matrix onto $span(U)$ and $span(\hat{U})$ and $\|\cdot\|_2$ is the spectral norm. For estimating the mean matrix, we use the loss function $L_M(\hat{M}, M) = \|\hat{M} - M\|_F^2/\|M\|_F^2$, where $\|\cdot\|_F$ is the Frobenius norm.

Iterative thresholding for sparse SVD

repeat

Right-to-Left Multiplication: $U^{(k),mul} = XV^{(k-1)}$.

Left Thresholding: $U^{(k),thr} = (u_{il}^{(k),thr})$, with $u_{il}^{(k),thr} = \eta\left(u_{il}^{(k),mul}, \gamma_{ul}\right)$.

Left Orthonormalization with QR Decomposition: $U^{(k)}R_u^{(k)} = U^{(k),thr}$.

Left-to-Right Multiplication: $V^{(k),mul} = X'U^{(k)}$.

Right Thresholding: $V^{(k),thr} = (v_{jl}^{(k),thr})$, with $v_{jl}^{(k),thr} = \eta\left(v_{jl}^{(k),mul}, \gamma_{vl}\right)$.

Right Orthonormalization with QR Decomposition: $V^{(k)}R_v^{(k)} = V^{(k),thr}$.

until Convergence

We next give a detailed description of the proposed sparse SVD method. Our algorithm originates from "orthogonal iteration" algorithm in the matrix computation literature which seeks the leading eigenvectors for symmetric matrices. We first generalize the orthogonal iteration method to handle asymmetric or even rectangular matrix. Moreover, we modify the algorithm by inserting thresholding steps within each iteration, which wipes out small entries, achieves sparsity, and reduces the variance in the estimator; see also Ma (2011). One more advantage of adding thresholding steps is the computational benefit since the subsequent multiplication and orthonormalization steps are much reduced because of the resulting zeros after thresholding. The proposal is schematically laid out in Algorithm 2. At each thresholding step, we perform entry-wise thresholding. We allow any thresholding function $\eta(x, \gamma)$ that satisfies $|\eta(x, \gamma) - x| \leq \gamma$ and $\eta(x, \gamma)1_{|x|\leq\gamma} = 0$, which includes soft-thresholding with $\eta_s(x, \gamma) = sign(x)(|x| - \gamma)_+$, hard-thresholding with $\eta_h(x, \gamma) = x1_{|x|>\gamma}$, etc. The threshold level γ is set to be $\sqrt{2\log(p_u \vee p_v)}$ [with

$a \vee b = \max\{a, b\}$] under normality assumption and can be chosen by data-driven method for other noise distributions; see Yang et al. (2011) for the details.

Under model (1), with the parameter space defined in (2), we have the following minimax lower bounds.

Theorem 3. *Under model* (1) *and parameter space* (2)*, there exists a constant* c*, s.t., for any estimator* $\tilde{U}$ *and* $\tilde{M}$*,*

$$\begin{aligned} \inf_{\tilde{U}} \sup_{\Theta(s_u,q_u;s_v,q_v)} \mathbb{E}_M L(\tilde{U}, U) &\geq c\, m_u \epsilon^2 , \\ \inf_{\tilde{M}} \sup_{\Theta(s_u,q_u;s_v,q_v)} \mathbb{E}_M L(\tilde{M}, M) &\geq c\, (m_u \vee m_v)\epsilon^2 , \end{aligned}$$

where m_u, ϵ^2 *are given by the following formulas:*

$$\begin{cases} m_u = \frac{s_u^{q_u} d_1^{q_u}}{(\log(p_u \vee p_v))^{q_u/2}}, \quad \epsilon^2 = \frac{\log(p_u \vee p_v)}{d_1^2}, & \text{if } m_u = O(p_u^\alpha),\ 0 < \alpha < 1; \\ m_u = \min\{d_1^2, p_u, s_u^{q_u} d_1^{q_u}\}, \quad \epsilon^2 = d_1^{-2}, & \text{otherwise.} \end{cases}$$

Our estimators achieve near optimal minimax rates of convergence, as is shown in the following upper bound result. Further note that our estimators do not require knowledge of the parameters $(s_u, q_u; s_v, q_v)$ and hence are adaptive.

Theorem 4. *Let* $\hat{U}, \hat{V}$ *be the output of Algorithm 2. Define* $\hat{M} = \hat{U}\hat{D}\hat{V}'$*, where* $\hat{D} = diag(\hat{d}_1, \ldots, \hat{d}_r)$ *with* $\hat{d}_l = \hat{\mathbf{u}}_l' X \hat{\mathbf{v}}_l$*. Under mild conditions, there exists a constant* C*, s.t.,*

$$\begin{aligned} \sup_{\Theta(s_u,q_u;s_v,q_v)} \mathbb{E}_M L_U(\hat{U}, U) &\leq C m_u \epsilon^2 , \\ \sup_{\Theta(s_u,q_u;s_v,q_v)} \mathbb{E}_M L_M(\hat{M}, M) &\leq C(m_u \vee m_v)\epsilon^2 , \end{aligned}$$

where $\epsilon^2 = \frac{\log(p_u \vee p_v)}{d_1^2}$*,* $m_u = \frac{s_u^{q_u} d_1^{q_u}}{(\log(p_u \vee p_v))^{q_u/2}}$ *and* m_v *is defined accordingly.*

The proof of the above theorems can be found in Yang et al. (2012).

References

[1] M. Lee, H. Shen, J.Z. Huang, and J.S. Marron. *Biclustering via sparse singular value decomposition.* Biometrics, **66** (2010), 1087-1095.

[2] Z. Ma. Sparse principal component analysis and iterative thresholding. Available at http://arxiv.org/abs/1112.2432. (2011).

[3] A. Shabalin and A. Nobel. *Reconstruction of a low-rank matrix in the presence of gaussian noise.* Preprint, availabel at http://arxiv.org/abs/1007.4148, (2010).

[4] D.M. Witten, R. Tibshirani, and T. Hastie. *A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis.* Biostatistics, **10** (2009) 515-534

[5] D. Yang, Z. Ma and A. Buja. *A Sparse SVD Method for High-dimensional Data.* Available at http://arxiv.org/abs/1112.2433. (2011).

[6] D. Yang, Z. Ma and A. Buja. *Near optimal sparse SVD in high dimensions.* Technical report, Department of Statistics, University of Pennsylvania. (2012).

High-dimensional estimation of causal effects

Marloes H. Maathuis
(joint work with Diego Colombo, Markus Kalisch, Peter Bühlmann)

We recently introduced a method to estimate bounds on causal effects from observational data, assuming the data are generated from an *unknown* directed acyclic graph (DAG) without hidden confounders. This method, called IDA (Intervention-calculus when the DAG is Absent; see [8, 7]) conceptually consists of two steps: (i) estimating the Markov equivalence class of DAGs from the conditional independence relationships in the observational data, and (ii) listing all DAGs in the estimated equivalence class and estimating the desired causal effect for each of these DAGs. The resulting possible causal effects are collected in a multi-set, which can be summarized by a summary measure of choice. Step (i) is known as causal structure learning. We used the PC-algorithm [13] for this purpose, which requires the so-called faithfulness assumption. This algorithm has been shown to be consistent for high-dimensional sparse graphs under some conditions [6]. Step (ii) concerns the estimation of causal effects when the DAG is given. We used Pearl's do-calculus [9, 10] for this step, which reduces to linear regression with covariate adjustment in the multivariate normal model. For large graphs, we developed a local version of step (ii) that does not require listing all DAGs in the equivalence class, as this quickly becomes computationally infeasible.

I presented a validation of IDA on a high-dimensional yeast gene expression data set [5], as described in [7]. This data set contains both observational and interventional data, obtained under similar conditions. The interventional data contains expression measurements of 5361 genes for 234 single-gene deletion mutant strains, and the observational data contains expression measurements of the same 5361 genes for 63 wild-type cultures. We used the interventional data to estimate the sizes of the causal effects of the 234 deletion genes on the remaining genes, and defined the top 10% of these effects as the "target set". We then applied IDA, as well as Lasso [14] and elastic-net [17] to estimate these causal effects, and obtained rankings of the estimated causal effects for each of the methods. The ROC curve in Figure 1 shows that IDA clearly outperforms the regression methods in identifying causal effects in the target set. This validation indicates that IDA can be a useful new tool for the design of experiments, since it can predict which interventions are likely to show a large effect.

I closed my talk with a discussion of some selected recent work on causal structure learning that might be incorporated into step (i) of IDA:
(a) Estimating an equivalence class based on a combination of observational and interventional data, see [3]. This approach has connections to active learning (see, e.g., [2]), i.e., given the estimated equivalence class, which interventions should be done subsequently to learn as much as possible about the underlying causal structure?

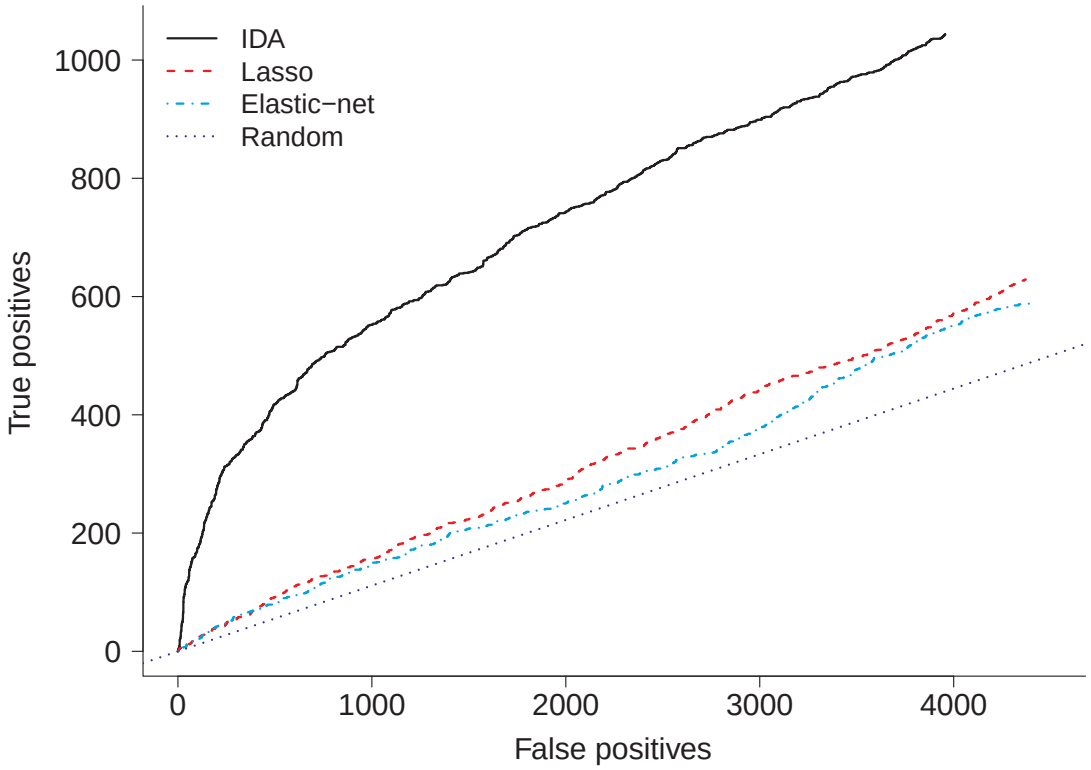


FIGURE 1. *(Taken from* [7]*)* The number of true positives (effects in the target set) versus the number of false positives (effects not in the target set) are plotted for the indicated methods, for the top $q = 1, \ldots, 5000$ predicted effects from the observational data. The target set is defined as the top 10% of the effects as computed from the interventional data.

(b) By imposing additional assumptions one can avoid the issue with the equivalence class. This is for example used in LiNGAM (Linear Non-Gaussian Acyclic Models, see, e.g., [12]) and in general additive noise models (see, e.g., [4, 11]).
(c) Estimating an equivalence class when allowing for arbitrarily many hidden and selection variables, using the FCI algorithm [13] or the much faster (but slightly less informative) RFCI algorithm [1]. See also Jamie Robins' talk for an approach that can sometimes narrow down the equivalence class substantially.
(d) Estimating an equivalence class based on different data sets with overlapping sets of variables, possibly obtained under different conditions, see, e.g., [15, 16].

REFERENCES

[1] D. Colombo, M.H. Maathuis, M. Kalisch, and T.S. Richardson, *Learning high-dimensional DAGs with latent and selection variables*, Ann. Statist. (2012), to appear.
[2] F. Eberhardt, *Almost optimal intervention sets for causal discovery*, Proceedings of the Twenty-Fourth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-08) (Corvallis, Oregon), AUAI Press, 2008, 161–168.
[3] A. Hauser and P. Bühlmann, *Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs*, Submitted, arXiv:1104.2808v1, 2012.

[4] P. Hoyer, D. Janzing, J.M. Mooij, J. Peters, and B. Schölkopf, *Nonlinear causal discovery with additive noise models*, 22nd Annual Conference on Neural Information Processing Systems (NIPS), 2009.
[5] T.R. Hughes, M.J. Marton, A.R. Jones, C.J. Roberts, R. Stoughton, C.D. Armour, H.A. Bennett, E. Coffey, H. Dai, Y.D. He, M.J. Kidd, A.M. King, M.R. Meyer, D. Slade, P.Y. Lum, S.B. Stepaniants, D.D. Shoemaker, D. Gachotte, K. Chakraburtty, J. Simon, M. Bard, and S.H. Friend, *Functional discovery via a compendium of expression profiles*, Cell **102** (2000), 109–126.
[6] M. Kalisch and P. Bühlmann, *Estimating high-dimensional directed acyclic graphs with the PC-algorithm*, J. Mach. Learn. Res. **8** (2007), 613–636.
[7] M. H. Maathuis, Diego Colombo, Markus Kalisch, and Peter Bühlmann, *Predicting causal effects in large-scale systems from observational data*, Nature Methods **7** (2010), 247–248.
[8] M. H. Maathuis, M. Kalisch, and P. Bühlmann, *Estimating high-dimensional intervention effects from observational data*, Ann. Statist. **37** (2009), 3133–3164.
[9] J. Pearl, *Causality. models, reasoning, and inference*, Cambridge University Press, Cambridge, 2000.
[10] J. Pearl, *Causal inference in statstics: an overview*, Statistics Surveys **3** (2009), 96–146.
[11] J. Peters, J.M. Mooij, D. Janzing, and B. Schölkopf, *Identifiability of causal graphs using functional models*, 27th Conference on Uncertainty in Artificial Intelligence (UAI), 2011.
[12] S. Shimizu, P.O. Hoyer, A. Hyvärinen, and A. Kerminen, *A linear non-Gaussian acyclic model for causal discovery*, J. Mach. Learn. Res. **7** (2006), 2003–2030.
[13] P. Spirtes, C. Glymour, and R. Scheines, *Causation, prediction, and search*, second ed., Adaptive Computation and Machine Learning, MIT Press, Cambridge, 2000.
[14] R. Tibshirani, *Regression shrinkage and selection via the lasso*, J. Roy. Statist. Soc. B. **58** (1996), 267–288.
[15] R.E. Tillman and P. Spirtes, *Learning equivalence classes of acyclic models with latent and selection variables from multiple datasets with overlapping variables*, Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS 2011), 2011.
[16] I. Tsamardinos, S. Triantafillou, and V. Lagani, *Towards integrative causal analysis of heterogeneous datasets and studies*, J. Mach. Learn. Res. (2012), To appear.
[17] H. Zou and T. Hastie, *Regularization and variable selection via the elastic net*, J. Roy. Statist. Soc. B. **67** (2005), 301–320.

Some Generalizations of Varying Coefficient Regression Models

Enno Mammen
(joint work with Young K. Lee, Byeong U. Park)

In this talk we consider a generalization of the varying coefficient regression model, proposed by Hastie and Tibshirani (1993). In the classical varying coefficient regression model the covariates are divided into two groups and the model contains only interaction terms between these two groups. In our model we abstain from the division of the covariates into two groups and we allow for interaction terms between all covariates. This broadens the field of applications of varying coefficient models. We discuss optimal rates for the estimation of nonparametric components of the model and we show that these rates can be attained by sieve and penalized least squares estimators. Furthermore, we give a detailed asymptotic distribution theory for kernel type-estimators that are given as the solution of a system of

nonlinear integral equations. This talk reports in particular on the results in Lee, Mammen and Park (2012).

In the Varying Coefficient Regression Model (Hastie and Tibshirani, 1993): one assumes for a response Y and covariates X and Z that $m(x,z) \equiv E(Y|X=x, Z=z)$ takes the form

$$m(x,z) = x_1 f_1(z_1) + \cdots + x_d f_d(z_d)$$

for some unknown univariate functions f_j. This model is simple in structure, gives easy interpretation, and yet is flexible since the dependence of the response variable on the covariates is modeled in a nonparametric way. But it is restrictive: each X_j and Z_j enter in only one "nonparametric interaction term" $X_j f_j(Z_j)$ and the covariates are divided into Z-covariates that are transformed nonparametrically and into X-covariates that are not transformed.

We consider the following generalization of the varying coefficient regression model:

$$g(m(x)) = x_1 \left(\sum_{k \in I_1} f_{1k}(x_k) \right) + \cdots + x_d \left(\sum_{k \in I_d} f_{dk}(x_k) \right),$$

where the index sets $I_j \subset \{1, 2, \ldots, D\}$ are known and each I_j does not include j. The index sets I_j may not be disjoint and g is a known link function.

Our model frees us from the restrictive assumption that covariates have to be divided into X-covariates and Z-covariates. It does not assume that all covariates only appear in one "nonlinear interaction term" $X_j f_{jk}(Z_k)$. In particular, it allows to let all covariates interact with all covariates, that is, take $I_1 \cup \{x_1\} = \cdots = I_d \cup \{x_d\}$. Furthermore, the introduction of a link function allows us to have a binary response. Our model generalizes the Varying Coefficient Regression Model, extended versions of it, the generalized additive model and the functional coefficient model.

We now discuss rate-optimal estimation in our model if n i.i.d. copies (X^i, Y^i) of (X, Y) can be observed. By application of empirical process theory rate-optimality can be easily checked for the estimation of m. Suppose e.g. that f_{jk} have bounded second order derivatives. Then we get by standard empirical process theory that (under some standard regularity conditions) m can be estimated with rate $n^{-2/5}$. This follows because entropy conditions for classes of f_{jk} carry over to the resulting class of functions m. This shows, that under appropriate conditions, m can be estimated with the same rate of convergence as the function f^* in the model $Y = f^*(U) + \varepsilon$, where f^* is the function among f_{jk} ($k \in I_{d_j}$, $1 \le j \le d$) belonging to the largest entropy class. We now will give conditions under which the rates for estimation of m carry over to the estimation of f_{jk}. For this purpose, we need two rewritings of our model. First we write our model as

$$g(m(x)) = \sum_{j=1}^{d} \alpha_j x_j + \sum_{\substack{j<k \\ j,k \in \mathcal{C}_0}} \alpha_{jk} x_j x_k + \sum_{j=1}^{d} x_j \left(\sum_{k \in I_j} f_{jk}(x_k) \right).$$

where each function f_{jk} satisfies the constraint $\int f_{jk}(x_k)w_k(x_k)\,dx_k = 0, k \in \mathcal{C}, 1 \leq j \leq d$; $\int x_k f_{jk}(x_k)w_k(x_k)\,dx_k = 0, j, k \in \mathcal{C}_0$ for nonnegative weight functions w_k with $\mathcal{C}_0 = \{1, 2, \ldots, d\} \cap \mathcal{C}$, $\mathcal{C} = \bigcup_{j=1}^{d} I_j$. Secondly, by collecting those X-elements that pertain to the function argument we can also write by rearrangement of terms:

$$g(m(x,z)) \quad = \quad \sum_{j=1}^{d} \alpha_j x_j + \sum_{\substack{j<k \\ j,k \in \mathcal{C}_0}} \alpha_{jk} x_j x_k + \tilde{x}_1^\top \mathbf{f}_1(x_{r+1}) + \cdots + \tilde{x}_p^\top \mathbf{f}_p(x_{r+p}),$$

where $\tilde{X}_k = \{X_j : r+k \in I_j,\ 1 \leq j \leq d\}$ for $1 \leq k \leq p$ and where we have assumed that $\mathcal{C} = \bigcup_{j=1}^{d} I_j = \{r+1, ..., r+p\}$. We now assume that the smallest eigenvalue of the matrix $E(\tilde{X}_k \tilde{X}_k^\top | X_{r+k} = x_{r+k})$, as a function of x_{r+k}, is bounded away from zero on the support of X_{r+k}. Under this assumption and under some other regularity conditions one can show the following inequality:

$$\|m - m^*\| \geq c \left(\|\alpha - \alpha^*\|_* + \sum_{k \in I_1} \|f_{1k} - f_{1k}^*\| + \cdots + \sum_{k \in I_{d_0}} \|f_{d_0 k} - f_{d_0 k}^*\| \right).$$

for some constant $c > 0$ and for all functions m and m^* with components α, f_{jk} or α^*, f_{jk}^*, respectively ($k \in I_{d_j}$, $1 \leq j \leq d_0$). Here, $\|\alpha\|_*^2 = \int (\sum_{j=1}^{d} \alpha_j x_j + \sum_{\substack{j<k \\ j,k \in \mathcal{C}_0}} \alpha_{jk} x_j x_k)^2 P_X(dx)$. From the inequality we get immediately:

PROPOSITION: *Suppose that an estimator $\hat{m}$ of m satisfies for a null sequence κ_n, that*

$$\int [g(\hat{m}(x)) - g(m(x))]^2 P_X(dx) = O_p(\kappa_n^2).$$

Then, it holds that

$$\int [\hat{f}_{jk}(x_k) - f_{jk}(x_k)]^2 p_{X_k}(x_k)\,dx_k = O_p(\kappa_n^2)$$

for all $k \in I_j$, $1 \leq j \leq d$.

In particular, this result can be used to show that the penalized least squares estimator and the sieve estimator achieve optimal rates.

We now discuss estimation procedures based on kernel smoothing. These estimators allow a complete asymptotic theory. Our estimation procedure is based on maximizing a quasi-likelihood. For this purpose we choose Q as a quasi-likelihood function with $\frac{\partial}{\partial \mu} Q(\mu, y) = \frac{y-\mu}{V(\mu)}$. Here, V is a function for modeling the conditional variance $v(x,z) \equiv \mathrm{var}(Y|X = x, Z = z)$ by $v(x,z) = V(m(x,z))$. The quasi-likelihood for the mean regression function m is then given by $\sum_{i=1}^{n} Q(m(X^i, Z^i), Y^i)$. We maximize the integrated kernel-weighted quasi-likelihood

$$L_{NW}(\alpha, \eta) \equiv \int \sum_{i=1}^{n} Q\Big(g^{-1}(\alpha * X^i + \eta_1(z_1)^\top \tilde{X}_1^i + \cdots + \eta_p(z_p)^\top \tilde{X}_p^i), Y^i\Big) K_h(X^{c,i}, z)\,dz,$$

where

$$\begin{aligned} X^{c,i} &= (X^i_{r+1}, \ldots, X^i_{r+p})^\top, \\ z &= (z_1, \ldots, z_p)^\top, \\ \eta(z) &= (\eta_1(z_1)^\top, \ldots, \eta_p(z_p)^\top)^\top, \\ \eta_k &= \{\eta_{j,r+k} : r+k \in I_j, 1 \le j \le d\}, \\ \alpha * X^i &= \sum_{j=1}^{d} \alpha_j X^i_j + \sum_{\substack{j<k \\ j,k \in \mathcal{C}_0}} \alpha_{jk} X^i_j X^i_k, \end{aligned}$$

and where K_h is a product kernel with bandwidth vector h.

Our nonparametric quasi-likelihood estimator is defined as:

$$(\hat{\alpha}^{NW}, \hat{f}^{NW}) = \arg\max_{(\alpha,\eta)} L_{NW}(\alpha, \eta)$$

where the maximization runs over the tuple of functions $\eta = (\eta_1, \ldots, \eta_p)$, each η_k being a vector of univariate functions that satisfy the constraints $\int \eta_{jl}(u) w_l(u) du = 0$, $r+1 \le l \le r+p, 1 \le j \le d$; $\int u\, \eta_{jl}(u) w_l(u)\, du = 0$, $r+1 \le j, l \le d$. We also considered a local linear quasi-likelihood estimator. For both estimators results on rate optimality and asymptotic normality can be found in Lee, Mammen and Park (2012). There also an algorithm is proposed based on the iterative solution of nonlinear integral equations.

References

[1] Y.K. Lee, E. Mammen, and B.U. Park, *Flexible Generalized Varying Coefficient Regression Models.* (2012) Preprint.

[2] T.J. Hastie and R.J. Tibshirani, *Varying-coefficient models.* Journal of Royal Statistical Society **B55** (1993), 757–796.

Non-negative least squares for high-dimensional data

Nicolai Meinshausen

Many regularization schemes for high-dimensional regression have been put forward. Most require the choice of a tuning parameter, using model selection criteria or cross-validation schemes. We show that a simple non-negative or sign-constrained least squares is a very simple and effective regularization technique for a certain class of high-dimensional regression problems. The sign constraint has to be derived via prior knowledge or an initial estimator. The success depends on conditions that are easy to check in practice. A sufficient condition for our results is that most variables with the same sign constraint are positively correlated. For a sparse optimal predictor, a non-asymptotic bound on the L1-error of the regression coefficients is then proven. Without using any further regularization, the regression vector can be estimated consistently as long as $\log(p)s/n \to 0$ for $n \to \infty$, where s is the sparsity of the optimal regression vector, p the number of variables and n sample size. The bounds are almost as tight as similar bounds

for the Lasso despite the fact that the method does not have a tuning parameter and does not require cross-validation. Network tomography is shown to be an application where the necessary conditions for success of non-negative least squares are naturally fulfilled and empirical results confirm the effectiveness of the sign constraint for sparse recovery.

The data are assumed to be given by a $n \times 1$-vector of real-valued observations $\mathbf{Y}$ and a $n \times p$-dimensional matrix $\mathbf{X}$, where column k of $\mathbf{X}$ contains all n samples of the k-th predictor variable for $k = 1, \ldots, p$. The non-negative least squares (NNLS) regression estimator is defined as

$$\hat{\beta} \;:=\; \text{argmin}_\beta \, \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 \quad \text{such that} \quad \min_k \beta_k \geq 0. \tag{1}$$

The following *Positive Correlation Condition* is the main assumption necessary to show success of non-negative least squares. The positively constrained minimal ℓ_1- eigenvalue of matrix $\mathbf{A}$ is defined as

$$\phi_{pos}^2(\mathbf{A}) := \min \Big\{ \frac{\beta^T \mathbf{A} \beta}{\|\beta\|_1^2} : \min_k \beta_k \geq 0 \Big\},$$

A lower bound on this restricted eigenvalue will be a sufficient condition for sparse recovery success of NNLS. A lower bound on this eigenvalue seems to be a much stricter condition than the *Compatibility Condition* [1]. However, the latter allows for positive and negative regression coefficients, while the *Positive Eigenvalue Condition* is restricted to positive coefficients. There are thus some immediate examples where it is fulfilled, the simplest of which being the case where all entries in the matrix $\mathbf{S} = \mathbf{X}^T\mathbf{X}$ are positive, $\min_{i,j} \hat{\mathbf{S}} \geq \nu > 0$.

Assume that the *Positive Eigenvalue Condition* holds with $\nu > 0$. Choose any $0 < \eta < 1/3$. Assume that the compatibility condition holds with $\phi > 0$ for $L = 4\nu^{-1}$. Setting

$$K_{p,\eta}^2 := 2 \log \big(\frac{\sqrt{2}p}{\sqrt{\pi}\eta} \big)$$

and assuming $\min_{k \in S} \beta_k > K_{p,\eta}\sigma/\sqrt{n\phi}$, it then holds with probability at least $1 - \eta$ that

$$\|\hat{\beta} - \beta^*\|_1 \quad \leq \quad K_{p,\eta} \, (5/\nu + 4/\sqrt{\phi}) \, \frac{s\sigma}{\sqrt{n}} \tag{2}$$

If the conditions hold, NNLS can thus recover the correct sparsity pattern in the absence of any further shrinkage, as long as $\log(p)s/n \to 0$ for $n \to \infty$. The rate of convergence is thus identical as for Lasso-type estimators [2]. A bound on prediction error can also be derived, which has the same rate again as Lasso-type estimators. [3] derived a bound for the prediction error of NNLS without *compatibility condition*, with a consequently $\sqrt{n}$-factor slower convergence. Overall the most compelling aspect of NNLS is that it does not require any tuning parameter beyond the choice of the signs of the individual regression coefficients. This makes the method very simple to understand and efficient to implement.

References

[1] S.A. Van De Geer and P. Bühlmann. On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.

[2] P. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 37:1705–1732, 2009.

[3] M. Slawski, M. Hein, and E. Campus. Sparse recovery by thresholded non-negative least squares. Technical report, Department of Computer Science, University of Saarbruecken, 2011.

Dynamics and Volatility

Hans-Georg Müller

(joint work with Rituparna Sen, Ulrich Stadtmüller, Wenwen Tao, Nicolas Verzelen, Fang Yao)

Common diffusion models for financial trading data include the Black-Scholes continuous time diffusion model for equity prices $X(t)$,

$$\frac{dX(t)}{X(t)} = \mu\, dt + \sigma\, dW(t),$$

with Wiener process W and constants μ, σ. For intra-day trading data, underlying processes are not not observed in the continuum, but rather on a discrete grid of densely spaced time points, $t_j = j\Delta, \quad j = 1, \ldots, [T/\Delta]$. Typical recordings are available on grids with $\Delta = 5$min, so-called "high-frequency" data.

For asymptotic analysis, we consider limits $\Delta \to 0$. Various diffusion models specifically for volatility include the Heston (1993) model:

$$d \log X(t) = \mu X(t)dt + \sqrt{v(t)}X(t)dW_1(t),$$

for a Wiener process W_1, where the stochastic volatility function v follows the Cox-Ingersoll-Ross square root diffusion model

$$dv(t) \quad = a(b - v(t))dt + c\sqrt{v(t)}dW_2(t),$$

with constants a, b, c, where Wiener process W_1 and W_2 have correlation ρ. Such diffusion models feature non-differentiable volatility trajectories.

As an alternative to these diffusion approaches, we consider a class of smooth volatility models, where the structure of the underlying processes is learned from a sample of n realizations of the underlying processes,

$$d \log X_i(t) = \widetilde{\mu}_i(t)\, dt + \widetilde{\sigma}_i(t)\, dW_i(t), \quad 0 \le t \le T, \quad i = 1, \ldots, n,$$

and $\widetilde{\mu}_i(t)\,,\ \widetilde{\sigma}_i(t)$ are i.i.d. copies of the stochastic processes $\widetilde{\mu}$ and $\widetilde{\sigma}$, which are assumed to be smooth but not stationary, and $W_i,\ i = 1, \ldots, n$, are n independent standard Wiener processes. Assume that the log-volatility process is smooth and can be represented as

$$\log(\widetilde{\sigma}^2(t)) = \widetilde{W}(t) + H(t), \quad \widetilde{W}(t) = f(\int_{t-\delta}^{t} \kappa(t - v)dW(v)) \text{ for a } \delta > 0,$$

where the stochastic process H and the function f are continuously differentiable, $H \perp W$; κ is a smooth kernel function with supp(K) $\subseteq [0, \delta]$, and $\int_0^{\tilde{\delta}} \kappa^2(u)du > 0$ for all $\tilde{\delta} > 0$,

$$\kappa(0) = \kappa'(0) = \kappa''(0) = 0, \quad \kappa(\delta) = \kappa'(\delta) = \kappa''(\delta) = 0.$$

The availability of multiple copies is crucial for the application of functional data analysis methodology in this setting. Defining

$$\begin{aligned} Z_\Delta(t) &= \frac{1}{\sqrt{\Delta}} \log\left(\frac{X(t+\Delta)}{X(t)}\right), \\ W_\Delta(t) &= \frac{1}{\sqrt{\Delta}} (W(t+\Delta) - W(t)), \end{aligned}$$

write

$$\begin{aligned} Z_\Delta(t) &= \frac{1}{\sqrt{\Delta}} \int_t^{t+\Delta} \widetilde{\mu}(v)\, dv + \frac{1}{\sqrt{\Delta}} \int_t^{t+\Delta} \widetilde{\sigma}(v)\, dW(v) \\ &= \widetilde{\mu}(t)\sqrt{\Delta} + \widetilde{\sigma}(t) W_\Delta(t) \; + \; R_1(t,\Delta) + R_2(t,\Delta), \end{aligned}$$

where under regularity conditions

$$E[\sup_{t\in[0,T]} |R_1(t,\Delta)|] = O(\Delta^{3/2}), \quad E[\sup_{t\in[0,T]} |R_2(t,\Delta)|] = O(\Delta^{1/2}).$$

The target is the smooth volatility process

$$V(t) = \log(\{\widetilde{\sigma}(t)\}^2),$$

which is related to the observations $Z_\Delta(t_j) = \frac{1}{\sqrt{\Delta}} \log\left(\frac{X(t_j+\Delta)}{X(t_j)}\right)$ by

$$\log(\{Z_\Delta(t_j)\}^2) - q_0 \approx Y_\Delta(t_j) = V(t_j) + U_\Delta(t_j).$$

Here q_0 is a numerical constant and $Y_\Delta(t), U_\Delta(t)$ are stochastic processes,

$$\begin{aligned} q_0 &= E(\log W_\Delta^2(t)) \approx -1.27, \\ Y_\Delta(t) &= \log(\{\widetilde{\sigma}(t) W_\Delta(t)\}^2) - q_0, \quad \text{observed raw volatilities} \\ U_\Delta(t) &= \log(\{W_\Delta(t)\}^2) - q_0, \quad \text{residuals for volatility.} \end{aligned}$$

Setting $G_V(s,t) = \mathrm{cov(V(s), V(t))}$, $\mathrm{E(V(t))} = \mu_\mathrm{V}(\mathrm{t})$, consider the auto-covariance operator of V, $G_V(f)(s) = \int G_V(s,t) f(t)\, dt$, with orthonormal eigenfunctions ϕ_k and eigenvalues λ_k, $k = 1, 2, \ldots$, such that $\lambda_1 \geq \lambda_2 \geq \cdots$ and $\sum_k \lambda_k < \infty$. This leads to the Karhunen-Loève representation

$$V(t) = \mu_V(t) \; + \; \sum_{k=1}^{\infty} \xi_k \phi_k(t),$$

where the ξ_k are uncorrelated random variables that satisfy

$$\xi_k = \int (V(t) - \mu_V(t)) \phi_k(t)\, dt, \quad E\xi_k = 0, \quad \mathrm{var}(\xi_\mathrm{k}) = \lambda_\mathrm{k}.$$

The components of the functional volatility process V can then be estimated from observed trades

$$Z_{ij\Delta} = \frac{1}{\sqrt{\Delta}} \log\left(\frac{X_i(t_j+\Delta)}{X_i(t_j)}\right),\ i=1,\ldots,n,\ j=1,\ldots,[\frac{T}{\Delta}],$$

and raw volatilities $Y_{ij\Delta} = \log(Z^2_{ij\Delta}) - q_0$, as

$$\begin{aligned} E(Y_\Delta(t)) &= \mu_V(t), \\ \mathrm{cov}(\mathrm{Y}_\Delta(\mathrm{s}), \mathrm{Y}_\Delta(\mathrm{t})) &= O(\sqrt{\Delta}) + G_V(s,t), \quad s \neq t. \end{aligned}$$

An estimation approach based on these relations can be implemented with common functional principal component analysis methodology, see PACE2.14 at `http://anson.ucdavis.edu/~mueller/data/pace.html`. The tuning parameters (smoothing parameters, number of included components) may be obtained with various methods, including variants of cross-validation, pseudo-BIC or fraction of variance explained. Under regularity conditions, these estimates are consistent, as the number of observed processes n satisfies $n \to \infty$ and $\Delta(n) \to 0$. This functional approach can be used to predict or classify volatility by applying various functional regression and classification methods ([1]).

A standard dynamic model for the univariate case is

$$X'(t) = f(t, X(t), \theta),$$

with a known "force function" f. Repeatedly observed realizations as considered here make it possible to learn the dynamics of the processes. In this endeavor, one needs to distinguish between Gaussian and non-Gaussian cases. In Gaussian situations, the dynamics are linear ([2]), while in non-Gaussian situations, as likely present in functional volatility processes, a general approach is

$$E\{X^{(1)}(t) - \mu^{(1)}(t) \mid X(t)\} = f(t, X(t)),$$

where f is unknown. One can then use smoothing methods to regress $\hat{X}'(t)$ versus $(t, \hat{X}(t))$ to obtain the unknown forcing function f and residual processes can be used to infer the properties of the drift process $X^{(1)}(t) - \mu^{(1)}(t) - f(t, X(t))$ ([3]).

References

[1] Müller, H.G., Sen, R., Stadtmüller, U. (2011). Functional data analysis for volatility. *J. Econometrics* **165**, 233–245.

[2] Müller, H.G., Yao, F. (2010). Empirical dynamics for longitudinal data. *Annals of Statistics* **38**, 3458–3486.

[3] Verzelen, N., Tao, W., Müller, H.G. (2012). Inferring stochastic dynamics from functional data. *Preprint*

Estimating filamentary structures

Wolfgang Polonik
(joint work with Wanli Qiao)

A filamentary structure in the plane is a collection of one-dimensional curves. This collection is envisioned to have a web-like geometric appearance, and observations taken from a distribution with ridges corresponding to this filamentary structure will tend to cluster along these curves. In other words, the corresponding distribution shows a higher concentration along the individual curves (filaments). This type of geometric structure is observed in several practical situations. Most prominently, the two-dimensional locations of the (observable) galaxies of our universe are known to form such a pattern (cosmic web). Other applications in which filamentary structures play a central role include diffusion tensor imaging, remote sensing, and medical imaging.

While a large number of approaches for the estimation of filaments can be found in the literature, in particular in the area of cosmology, not many of them come with a more theoretical foundation. Some exceptions from the more statistical literature are [1, 2, 3, 4, 5]. In this paper we discuss two novel statistically motivated approaches for filament estimation. As in [3] and [4] our methods also are based on the estimation of integral curves.

For simplicity we only discuss the case of 2-dimensional filamentary structures. Given a vector field $v : \mathbb{R}^2 \to \mathbb{R}^2$ an integral curve $x : [0, T] \to \mathbb{R}^2$ is the solution of the differential equation $\frac{d}{dt}x(t) = V(x(t))$ with a given starting value $x(0) = w$.

1. Method I: A geometric algorithm

The first method for estimating filaments is based on integral curves of the gradient field of an underlying probability density f, i.e. here we have $V(x) = \nabla f(x)$. Integral curves of gradient fields are paths of steepest ascent. We estimate these integral curves by the corresponding integral curves of a kernel density estimator.

A filament or ridge line corresponds to an integral curve with a particular starting value which itself then obviously lies on the filament. However, given observations, the choice of the 'right' starting value is the problem. Instead we use the intuitive geometric idea that many integral curves of the gradient field will ultimately run along ridges lines of the density (i.e. filament) up to a local maximum. Our proposed methodology can be described in brief as follows:

- estimate the integral curves via a (modified) mean shift algorithm; each estimated integral curve is given by a sequence of pairs $(\widehat{m}_1, \widehat{d}_1), (\widehat{m}_2, \widehat{d}_2), \ldots, (\widehat{m}_N, \widehat{d}_N)$, where the $\widehat{m}_j$ denote the local means and the $\widehat{d}_j$ denote the directions in which to move from $\widehat{m}_j$;
- clip mean shift paths once close to ridges;
- use local means $\widehat{m}_j$ and directions $\widehat{d}_j$ from clipped mean shift paths as input for the following steps;
- use input to construct estimates of filamentary pieces as follows:
  - determine a starting value $\widehat{x}_0$;

  - i = 0

  WHILE $\widehat{f}(\widehat{x}_i) > \epsilon_0$ DO
  - estimate directional derivative of filament at $\widehat{x}_i$ by averaging the directions $\widehat{d}_j$ corresponding to means $\widehat{m}_j \in \{x \in \mathbb{R}^2 : \|\widehat{x}_j - \widehat{x}_i\| < \epsilon\}$ for a predetermined value $\epsilon > 0$;
  - move a 'small' step into the estimated directional derivative of the filament to find $\widehat{x}_{i+1}$;
  - set $i = i + 1$

  END
- apply B-spline (combination of interpolation and fitting) to get final smooth estimate of filament.

The algorithm also involves a bias correction at each step, and the determination of candidates for branching points (or points of intersections) of filaments via clustering of directional data. Further details are omitted here.

2. Method II: Ridge estimation via bump hunting

Definition: A point is said to lie on a ridge or filament of a twice differentiable function f if

$$H\,\nabla f = \lambda_1 \nabla f$$
$$\lambda_2 < 0$$

where $\lambda_1 > \lambda_2$ are the two eigenvalues of the Hessian H of f.

Let V denote *second* eigenvector of Hessian H. Notice that on the filament, either $\nabla f = 0$ or $\nabla f \parallel V^{\perp}$, i.e. $\langle \nabla f,\ V \rangle = 0$. The underlying *geometric idea* is the following. Consider the vector field generated by the *second* eigenvectors $V(x)$ of the Hessian H of f. Then,

- *a ridge point corresponds to a local mode of f along the path of the corresponding integral curve generated by $V(x)$.*

Let $W \subset \mathrm{supp}(f)$ open. For each $w \in W$ let $x_w(t), t \in [0, T]$ denote the integral curve corresponding to the vector field $V(x)$ starting at $x_w(0)$ parametrized by $t \in [0, T]$ and define

$$\theta_w = \arg \max_{t \in [0,T]} f(x_w(t)),$$

i.e. $x_w(\theta_w)$ is a point on the ridge line of f.

Estimation. Let $\widehat{f}_n$ denote a standard kernel density estimator of f with kernel K and bandwidth h based on a random sample $X_1, \ldots, X_n \sim f$. Let further $\widehat{V} = \widehat{V}(x)$ denote the second eigenvector of the Hessian of $\widehat{f}_n$. We write

$$\widehat{V}(x) = G(\mathbf{vech}\widehat{H}(x)),$$

where for a symmetric (2×2)-matrix $A = (a_{ij})$ we write $\mathbf{vech}(A) = (a_{11}, a_{12}, a_{22})^T$, and the function $G : \mathbb{R}^3 \to \mathbb{R}^2$ is defined by this identity. We let further $\widehat{X}_w(t), t \in$

$[0,T]$ denote the integral curve corresponding to vector field $\widehat{V}(x)$ starting at $w \in W$ and define

$$\widehat{\theta}_w = \arg\max_{t\in[0,T]} f(\widehat{X}_w(t)),$$

i.e. $\widehat{X}(\widehat{\theta}_w)$ is a point on the ridge line of $\widehat{f}_n$. Of course we assume that both θ_w and $\widehat{\theta}_w$ are well defined. For a vector v we let $v^\perp$ denote the vector orthogonal to v.

Assumptions.
(A0) $W \subset \mathbb{R}^2$ is a bounded open set.
(A1) $supp(f) := \overline{\{x : f(x) \neq 0\}} \subset G$ and f is four times continuously differentiable. All partial derivatives up to 4th-order are bounded and L_1-integrable.
(A2) The kernel K is a symmetric pdf with bounded support, all partial derivatives up to 3-rd order are bounded and $\int_{\mathbb{R}^2} K(x)xx^T dx = \mu_2(K)\mathbf{I}$ with $\mu_2(K) < \infty$.
(A3) $||\mathbf{R}(\mathbf{vech}\nabla^2 K)|| < \infty$ where $\mathbf{R}(g) := \int_{\mathbb{R}^2} g(x)g(x)^T dx$.
(A4) The two eigenvalues of the Hessian of f are different within $supp(f)$.
(A5) There exists a ridge in W defined by f with the corresponding underlying manifold (ridge line) being differentiable and having bounded curvature.

Theorem 1. Suppose that in addition to the above assumptions we have $nh^9 \to \beta \geq 0$, $h_n \to 0$.

(a) If $\nabla f(x(\theta)) \neq 0$, then for any starting point $x \in W$ we have

$$\sqrt{nh^6}V(x_w(\theta))'(\widehat{X}_w(\widehat{\theta}_w) - x_w(\theta)) \to N(0, \sigma_1^2(x_w(\theta_w)))$$
$$\sqrt{nh^5}\big(V(x_w(\theta))^\perp\big)'(\widehat{X}_w(\widehat{\theta}_w) - x_w(\theta_w)) \to N(0, \sigma_2^2(x_w(\theta_w)))$$

(b) If $\nabla f(x(\theta)) = 0$, then for any starting point $w \in W$ we have

$$\sqrt{nh^5}V(x_w(\theta))'(\widehat{X}_w(\widehat{\theta}_w) - x_w(\theta_w)) \to N(0, \sigma_3^2(x_w(\theta_w)))$$
$$\sqrt{nh^5}(\big(V(x_w(\theta))^\perp\big)'(\widehat{X}_w(\widehat{\theta}_w) - x_w(\theta_w)) \to N(0, \sigma_4^2(x_w(\theta_w)))$$

We only provide a heuristic argument for the proof. It is well-known that (for $d = 2$) our smoothness assumptions assume that the kernel estimate of the 1-st derivatives converges at the rate $O_P(1/\sqrt{nh^{d+2}}) = O_P(\sqrt{1/nh^4})$, and the corresponding estimate of the 2-nd derivatives at rate $O_P(1/\sqrt{nh^{d+4}}) = O_P(1/\sqrt{nh^6})$. Integral curves, however, are of the form $x(t) = w + \int_0^t V(x(s))\,ds$, i.e. they are one-dimensional integrals of a smooth function of *second* derivatives. Therefore one can expect that points at the ridge can be estimated at rate $O_P(1/\sqrt{nh^5})$ (gain one power of h in the variance).

Next observe that we can write

$$\widehat{X}_w(\widehat{\theta}_w) - x_w(\theta_w) = \underbrace{\left[\widehat{X}_w(\widehat{\theta}_w) - x_w(\widehat{\theta}_w)\right]}_{\approx O_P(1/\sqrt{nh^5})} + \underbrace{\left[x_w(\widehat{\theta}_w) - x_w(\theta_w)\right]}_{\approx O_P\left(V(x_w(\theta_w))(\widehat{\theta}-\theta)\right)}.$$

We will see below that

$$\widehat{\theta}_w - \theta_w = \begin{cases} O_P(1\sqrt{/nh^6}) \text{ if } \nabla f(x_w(\theta_w)) \neq 0, \\ O_P(1/\sqrt{nh^5}) \text{ if } \nabla f(x_w(\theta_w)) = 0 \end{cases}$$

This then implies the result. Next we present some discussion to understand the rates of convergence of $\widehat{\theta} - \theta$. The following result is crucial. To simplify notation we omit the index w indicating the starting value:

Lemma. If $z(x(\theta)) = \nabla \langle \nabla f(x(\theta)), V(x(\theta)) \rangle \neq 0$, then we have under our smoothness assumptions that

$$\begin{aligned} \hat{\theta} - \theta &= \frac{1}{z(X(\theta))} \langle \nabla \widehat{f}(\widehat{X}(\theta)), \widehat{V}(\widehat{X}(\theta)) \rangle + o_p(\widehat{\theta} - \theta) \\ &= \frac{1}{z(X(\theta))} \Big[\langle H(x(\theta))(\widehat{X}(\theta) - x(\theta)), V(x(\theta)) \rangle \\ &\quad + \langle \nabla f(x(\theta)), \widehat{V}(X(\theta)) - V(x(\theta)) \rangle \Big] + o_p(\widehat{\theta} - \theta) \end{aligned}$$

Considering the leading terms in this expansion, we can see the following. At first order the difference $\widehat{\theta} - \theta$ depends on estimates of the first derivatives, the estimates of the second derivatives, and on the estimates of the integral curve itself. The estimation of $\widehat{V}$ is the hardest, and it determines the rates of $\widehat{\theta} - \theta$ provided that $\nabla f(x(\theta) \neq 0$. Otherwise, the estimation of $V(x(\theta))$ only enters the second order terms, and the rates of convergences of the integral curve determine the rate of $\widehat{\theta} - \theta$.

gradient itself is zero, then one does not need to estimate the second derivatives well in order for $\widehat{\theta}$ to be close to θ. Otherwise, the rate of estimation of $\widehat{V}$ comes into the picture and in fact this rate dominates the other estimation rates.

More precisely, since $\nabla \widehat{f}(X(\theta)) - \nabla f(X(\theta)) = O_P(\frac{1}{\sqrt{nh^4}})$, $\widehat{V}(x(\theta)) - V(x(\theta)) = O_p(\frac{1}{\sqrt{nh^6}})$ and $\widehat{X}(\theta) - x(\theta) = O_p(\frac{1}{\sqrt{nh^5}})$ (for the latter, see below), we see that for $\nabla f(x(\theta)) = 0$, we have

$$\sqrt{nh^5}(\widehat{\theta} - \theta) = \frac{\langle H(x(\theta))\sqrt{nh^5}(\widehat{X}(\theta) - x(\theta)), V(x(\theta)) \rangle}{\lambda_2(x(\theta)) \; ||V(x(\theta))||^2} + o_p(1),$$

and for $\nabla f(x(\theta)) \neq 0$ we have

$$\sqrt{nh^6}(\widehat{\theta} - \theta) = \frac{\langle \nabla f(x(\theta)), \sqrt{nh^6}[\widehat{V}(x(\theta)) - V(x(\theta))] \rangle}{\langle \nabla \langle \nabla f(x(\theta)), V(x(\theta)) \rangle, V(x(\theta)) \rangle} + o_p(1).$$

Conditions that assure the rate of convergence of $\widehat{X}(\theta) - x(\theta)$ used in the arguments above, are formulated in the following theorem, which can be proven by

an adaptation of the arguments given in [3]. Recall that $\widehat{V} = \widehat{V}(x)$ denotes the second eigenvector of the Hessian of $\widehat{f_n}$. We write

$$\widehat{V}(x) = G(\mathbf{vech}\widehat{H}(x)),$$

where the function $G : \mathbb{R}^3 \to \mathbb{R}^2$ is defined by this identity.

Theorem 2. Assume $h_n \to 0$, $nh_n^9 \to \beta \geq 0$ as $n \to \infty$, and $nh^7 \to \infty$ as $n \to \infty$, then

$$\sqrt{nh_n^5}(\hat{X}(t) - x(t)) \to \xi(t) \quad \text{weakly}$$

in $C[0,T] = C([0,T), \mathbb{R}^2)$, the space of $\mathbb{R}^2$-valued continuous functions on $[0,T]$. Here $\xi(t), 0 \leq t \leq T$, is a Gaussian process satisfying

$$d\xi(t) = \frac{\sqrt{\beta}}{2}\nabla G(\mathbf{vech}H(x(t)))\mathcal{B}(x(t))dt + \nabla G(\mathbf{vech}H(x(t)))\xi(t)dt + \left(\nabla G(\mathbf{vech}H(x(t)))\Omega(x(t))(\nabla G(\mathbf{vech}H(x(t))))^T\right)^{1/2} dW(t)$$

with initial condition $\xi(0) = 0$, where $W(t), t \geq 0$, is a standard Brownian sheet in $\mathbb{R}^2$, $\Omega(x(t)) = \int\int \Psi(x(t), \tau, z) f(x(t)) dz d\tau$ where $\Psi(x(s), \tau, z)$ denotes the (3×3)-matrix $(a_{ij}), i, j = 1, 2, 3$ with $a_{ij} = b_i c_j$, where $\mathbf{vech}(\nabla^2 K(z)) = (a_1, a_2, a_3)^T$ and $\mathbf{vech}(\nabla^2 K(G(\mathbf{vech}H(x(s)))\tau + z) = (b_1, b_2, b_3)^T$. Further,

$$\mathcal{B}(x(s)) = \begin{pmatrix} \int K(z) z^T \nabla^2 f^{(2,0)}(x(s)) z dz \\ \int K(z) z^T \nabla^2 f^{(1,1)}(x(s)) z dz \\ \int K(z) z^T \nabla^2 f^{(0,2)}(x(s)) z dz \end{pmatrix}$$

with $(f^{(2,0)}(x), f^{(1,1)}(x), f^{(0,2)}(x))^T = \mathbf{vech}H(x)$.

Uniform convergence of $\widehat{X}_w(\widehat{\theta}_w) - x_w(\theta_w)$. Notice that the difference $\widehat{X}_w(\widehat{\theta}_w) - x_w(\theta_w)$ is the difference of one point at the estimated ridge line and a 'corresponding' point at the true ridge line. By corresponding points we mean that each of them is determined by an integral curve (true and estimated, respectively) with the same starting value. By considering the supremum over all starting values, we obtain a measure for the distance between the true and the estimated ridge line. We have the following result for the supremums distance.

Theorem 3. Under the above assumptions we have

$$P\left(\sup_{w \in W} |\alpha_n^w (\widehat{X}_w(\widehat{\theta}_w) - x_w(\theta_w))^T V(x_w(\theta_w))| < B(h) + z/(2\log\tfrac{1}{h})^{1/2}\right) = \exp\{-2\exp\{-z\}\}$$

where $B(h) = (2\log\frac{1}{h})^{1/2} + (2\log\frac{1}{h})^{-1/2}\log\{(2\pi)^{-1/2} M\}$ with M a constant depending on the filament, and

$$\alpha_n^w = \frac{\langle \nabla \langle \nabla f(x_w(\theta_w)), V(x_w(\theta_w))\rangle, V(x_w(\theta_w))\rangle / \|V(x_w(\theta_w))\|^2}{\sqrt{\nabla f(x_w(\theta_w))^T \nabla G(\mathbf{vech}H(x_w(\theta_w)))\Sigma(x_w(\theta_w))\nabla G(\mathbf{vech}H(x_w(\theta_w)))^T \nabla f(X_w(\theta_w))}}$$

with $\Sigma(x) = \mathbf{R}(\mathbf{vech}\nabla^2 K) f(x)$.

References

[1] M.-Y. Cheng, P. Hall and J.A. Hartigan, Estimating gradient trees, In: *A* Festschrift for Herman Rubin, IMS Lecture Notes -Monograph Series **45** (2004), 237 –249.

[2] E. Arias-Castro, D.L. Donoho, and X. Huo, Adaptive multiscale detection of filamentary structures in a background of uniform random points, *A*nn. Statist. **34** (2006), 326 – 349.

[3] V. Koltchinskii, L. Sakhanenko, and S. Cai, Integral curves of noisy vector fields and statistical problems in diffusion tensor imaging: Nonparametric kernel estimation and hypothesis testing, *A*nn. Statist. **35** (2007), 1576 – 1607.

[4] C. Genovese, M. Perone-Pacifico, I. Verdinelli and L. Wasserman, On the path density of a gradient field, *Ann. Statist.*, **37** (2009), 3236 – 3271.

[5] C. Genovese, M. Perone-Pacifico, I. Verdinelli, and L. Wasserman, Nonparametric Filament Estimation, *arXiv:1003.5536v1*, (2010).

Efficient Estimation of Single Index Models using Adapted Bregman Losses

Pradeep Ravikumar

(joint work with Martin J. Wainwright, Bin Yu)

A *multiple-index* regression model [1] is a semiparametric regression model where the response or output variable $Y \in \mathbb{R}$ depends on the covariates or input variables $X \in \mathbb{R}^p$ as

$$Y = \sum_{j=1}^{m} g_j(\beta_j^T X) + \epsilon, \tag{1}$$

where ϵ is additive zero mean noise, independent of X. The functions $\{g_j\}_{j=1}^m$ are assumed to belong to some given class of functions $\mathcal{G}$, for instance the set of differentiable functions, or monotonically increasing functions. The linear projections $\beta_k^T X$ provide unidimensional summaries of the covariates, and each of these is called an *index* (hence the name multiple-index model). Another term typically used for these components is that of a *ridge* function, since the function $g(\beta^T x)$ is constant over the hyperplane $\beta^T x = c$ (so that its function surface looks like a ridge). Given n i.i.d. samples $S = \{(X^i, Y^{(i)}),\ i = 1, ..., n\}$ from the model (1), the model-estimation task comprises estimating both a parametric component $\{\beta_j\}_{j=1}^m$ as well as the nonparametric components or functions $\{g_j\}_{j=1}^m$.

Single Index Model. A key step in estimating these multiple-index models, for instance via back-fitting, requires the estimation of a *single-index model*, which is the special case of (1) with $m = 1$. In the sequel of this report, we concern ourselves with just the estimation of such a single-index model. Overloading notation, we assume the following model:

$$Y = g(\beta^T X) + \epsilon, \tag{2}$$

where ϵ is zero mean noise independent of X, and we are given samples $\{(X^{(i)}Y^{(i)})\}_{i=1}^n$ drawn iid from this model. A popular approach for estimating the model components (g, β) from these samples is via an alternating procedure

that optimizes over the parameters β and the function g alternately [1]. As [2] and others show, such alternating steps (or even a finite number of them) can be shown to result in good estimator *provided* we are able to obtain the global optima of the corresponding optimization problems. However, this is a significant caveat because estimating the β parameters entails solving a non-convex optimization problem.

Non-convexity. Estimating the parameteric vector of a single index model using the squared error loss function is a non-convex estimation task. The solutions computed using practical methods are thus only suboptimal, and do not enjoy the strong guarantees available to the global minimum [2]. Moreover, they can be unstable, particularly in the presence of multiple local minima. Our goal is thus to obtain a surrogate loss function that is convex, and moreover has a nearly identical minimum as the squared error loss. Towards this, we propose a novel two-stage estimation procedure, where instead of using the squared error, we use a convex loss function applied to β that is *adapted* from the current estimate of g. For the case of monotonic functions g, by using appropriate classes of Bregman divergences, we obtain an overall procedure that involves only tractable convex optimization steps, and is provably Fisher consistent.

Bregman Updates. Consider the population least squares functional, namely

$$\min_{g\in\mathcal{G},\beta\in\mathbb{R}^p} \mathbb{E}(Y - g(\beta^T X))^2. \tag{3}$$

By computing the Hessian with respect to β, it is straightforward to see that this function is not convex in terms of β for general functions g. (It is convex, for instance, for linear g.) Given this non-convexity, we are motivated to consider a larger class of loss functions, in particular the class of Bregman divergences. For any Bregman function F (roughly, a strictly convex differentiable function), the Bregman divergence $D_F(a, b)$ is defined as,

$$D_F(a\|b) := F(a) - F(b) - \nabla F(b)^T(a - b), \tag{4}$$

The Bregman divergence induced by a univariate Bregman function F, between Y and $g(\beta^T X)$ is then given by,

$$D_F(Y\|g(\beta^T x)) = F(Y) - F(g(\beta^T X)) - f(g(\beta^T X))(Y - g(\beta^T X)), \tag{5}$$

where $f = F'$. The squared error loss function is a special case, obtained by setting $F(z) = 1/2z^2$. Of interest to us are alternative choices of Bregman distances; in particular, the following result shows that for any monotonic g, there is always a Bregman divergence for which estimation of β reduces to a convex problem:

Proposition 5. *Consider the single index model* (2) *when g belongs to the class $\mathcal{G}$ of monotonically increasing functions. Then for any $g \in G$, there exists a Bregman divergence $D_F(g)$ for which the estimation of β is a convex problem. In particular, define $G(v) = \int_\infty^v g(t)dt$, and define the function*

$$F(u) = \sup_{v\in\mathbb{R}} v^T u - G(v), \tag{6}$$

The Bregman divergence $D_F(g)$ induced by this choice of F, when applied to the pair y and $g(\beta^T x)$, takes the form

$$D_F(g)(y\|g(\beta^T x)) = G(\beta^T x) - \beta^T xy + F(y), \tag{7}$$

which is a convex function of β whenever g is monotonic.

Note that the function (6) is the Fenchel conjugate of the function G. Overall, this result motivates the following practical scheme. Since G is convex for monotonic g, optimizing the "surrogate" function (7) for β is a convex program. On the other hand, for fixed β, estimation of the function g in the single index model (2) is a standard problem in isotonic regression. Thus, we have the following two-stage procedure for estimating a single index model:

Solving a single-index model: Bregman Updates

Initialize: $\beta = 0$, $g = 0$.
for outer iterations $t = 1, 2, \ldots$ until convergence **do**
Fixing g, obtain β by solving:

$$\beta \in \arg\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2n} \sum_{i=1}^{n} \left(G(\beta^T X^{(i)}) - Y^{(i)}(\beta^T X^{(i)}) \right) \right\}. \tag{8}$$

Fixing β, obtain g by solving

$$g \in \arg\min_{g \in \mathcal{G}} \left\{ \frac{1}{2n} \sum_{i=1}^{n} (Y^{(i)} - g(\beta^T X^{(i)}))^2 \right\}. \tag{9}$$

end for

References

[1] T. J. Hastie, R. J. Tibshirani, and J.H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. 2009.

[2] D. Donoho, I. Johnstone, P. Rousseeuw, and W. Stahel. Discussion: Projection pursuit. *The Annals of Statistics*, 13(2):496–500.

Accuracy of empirical projections in high dimension

Angelika Rohde

As a consequence of the Bai and Yin (1993) law, the maximal singular value $\lambda_{\max}(\varepsilon)$ of a standard Gaussian $M \times M$-matrix ε is equal to $2\sqrt{M}(1 + o(1))$ a.s. With $\hat{\pi}_1$ denoting the projection onto the one-dimensional subspace of $\mathbb{R}^M$ maximizing $\|\tilde{\pi}_1 \varepsilon\|_{S_2}^2$ over all one-dimensional projections $\tilde{\pi}_1$, i.e. $\|\hat{\pi}_1 \varepsilon\|_{S_2}^2 = \lambda_{\max}(\varepsilon)^2$, one can also establish the bound, due to uniform integrability,

$$\mathbb{E}\|\hat{\pi}_1 \varepsilon\|_{S_2}^2 \;=\; 4M(1 + o(1)), \tag{1}$$

with the Hilbert-Schmidt or Frobenius norm $\|\cdot\|_{S_2}$. In contrast, $\mathbb{E}\|\tilde{\pi}_1\varepsilon\|_{S_2}^2 = M$ for every fixed $\tilde{\pi}_1$. Thus, replacing one single projection by the supremum over all projections increases the Hilbert-Schmidt norm by a positive factor. This effect raises the question about the accuracy for empirical reduced-rank projections in general. Here and subsequently, let

$$\hat{\pi}_r \;:=\; \arg\max_{\tilde{\pi}_r\in\mathcal{S}_{M,r}} \|\tilde{\pi}_r X\|_{S_2}^2 \;\text{ and }\; \pi_r \;\in\; \mathrm{Arg}\max_{\tilde{\pi}_r\in\mathcal{S}_{M,r}} \mathbb{E}\|\tilde{\pi}_r X\|_{S_2}^2$$

with $\mathcal{S}_{M,r}$ denoting the set of all orthogonal projections onto r-dimensional subspaces of $\mathbb{R}^M$. How close is $\mathbb{E}\|\hat{\pi}_r X\|_{S_2}^2$ to $\mathbb{E}\|\pi_r X\|_{S_2}^2 = \|\pi_r C\|_{S_2}^2 + \sigma^2 rM$ if the Gaussian matrix $X = C + \varepsilon$, ε_{ij} iid $\mathcal{N}(0,\sigma^2)$, is not centered? Can the above described situation in (1) be improved by an adequate choice of $C = \mathbb{E}X$, or does there exist for any arbitrarily large real number c some unfavorable matrix $C(c)$ such that $\mathbb{E}\|\hat{\pi}_r X\|_{S_2}^2 - \mathbb{E}\|\pi_r X\|_{S_2}^2 \geq c$?
The motivation for considering this problem is manifold. Mainly, since

$$\mathbb{E}\|\hat{\pi}_r X\|_{S_2}^2 \;=\; \big(\mathbb{E}\|\hat{\pi}_r X\|_{S_2}^2 \;-\; \mathbb{E}\|\pi_r X\|_{S_2}^2\big) \;+\; \|\pi_r C\|_{S_2}^2 \;+\; \sigma^2 rM$$

and $\mathbb{E}\|\hat{\pi}_r X\|_{S_2}^2 - \mathbb{E}\|\pi_r X\|_{S_2}^2 \geq 0$, the problem is of theoretical interest as our results complement the bound in (1) for centered Gaussian matrices with a detailed analysis of the noncentered case, extending also to more general rank-r projections. Finite-rank perturbations of random matrices have found recently a lot of attention, see Capitaine et al. (2009), Capitaine et al. (2012), Pizzo et al. (2012), Tao (2012) among others. Tao (2012), Theorem 1.7, studies the the eigenvalue value spectrum of low rank perturbations of an iid complex random matrix and proves, as a special case, that $\gamma_{\max}(C+\varepsilon/(\sigma\sqrt{M})) = \gamma_{\max}(C) + o_p(1)$ as $M\to\infty$ and $\mathrm{rank}(C) = O(1)$ as long as $|\gamma_{\max}(C)| = O(1)$ is sufficiently large, with $\gamma_{\max}(C)$ the eigenvalue of C which is maximal in absolute value. Capitaine et al. (2009) and Pizzo et al. (2012) study Wigner matrices instead of iid random matrices. Somewhat remarkably, the outlier eigenvalues of the perturbed matrix are not close in probability to those of the original matrix C but to some shifted value $\lambda_i(C) + \sigma^2/\lambda_i(C)$, where σ^2 is the common variance of the entries of the Wigner matrix, and $\lambda_i(C)$ the eigenvalues of an Hermitian matrix C. Our results are complementary: We derive non-asymptotic cumulated second moment bounds on the singular values in the deformed (non-Hermitian) iid real Gaussian matrix case for general, not necessarily low or uniformly bounded rank perturbations, and we study the quality of approximation in particular. Our proofs differ significantly from the techniques of the above mentioned results but rely on empirical process techniques without making use of classical random matrix tools.
Subsequently, $c > 0$ denotes some constant which does not depend on the variable parameters in the expressions. It may vary over different places of appearance.
Our first result is an upper bound in the *small amplitude regime* $\|C\|_{S_\infty} \leq \sigma\sqrt{M}$, where $\|\cdot\|_{S_\infty}$ denotes the spectral norm.

Proposition 6 (Small amplitude regime)**.** *For any $M \in \mathbb{N}$, $r < M$:*

$$\sup_{\substack{C\in\mathbb{R}^{M\times M}:\\ \|C\|_{S_\infty}\le\sigma\sqrt{M}}} \mathbb{E} \sup_{\tilde\pi_r\in\mathcal{S}_{M,r}} \Big(\|\tilde\pi_r(C+\varepsilon)\|_{S_2}^2 - \|\pi_r(C+\varepsilon)\|_{S_2}^2\Big) \;\le\; c\sigma^2 rM.$$

The proposition demonstrates that the small amplitude regime resembles the well-known situation in high-dimension for $C = 0$.
The *large amplitude regime* is shown to be substantially different. The next result is a lower bound on the expected squared Hilbert-Schmidt norm of the rank-r-projection in case that the singular value spectrum of C is constant.

Proposition 7. *Let $C_\alpha \in \mathbb{R}^{M\times M}$ with singular value decomposition $U\Lambda_\alpha V'$. Assume that $\Lambda_\alpha = \alpha\mathrm{Id}$ with some real number $\alpha \in \mathbb{R}$. Then for any fixed $\pi_r \in \mathcal{S}_{M,r}$ and for any $\alpha \in \mathbb{R}$, (i)*

$$\mathbb{E}\Big(\sup_{\tilde\pi_r\in\mathcal{S}_{M,r}} \big\|\tilde\pi_r(C_\alpha+\varepsilon)\big\|_{S_2}^2 - \big\|\pi_r(C_\alpha+\varepsilon)\big\|_{S_2}^2\Big) \;\ge\; \mathbb{E}\Big(\sup_{\tilde\pi_r\in\mathcal{S}_{M,r}} \|\tilde\pi_r\varepsilon\|_{S_2}^2\Big) \;-\; \sigma^2 rM,$$

and (ii)

$$\liminf_{|\alpha|\to\infty} \frac{1}{|\alpha|}\,\mathbb{E}\Big(\sup_{\tilde\pi_r\in\mathcal{S}_{M,r}} \big\|\tilde\pi_r(C_\alpha+\varepsilon)\big\|_{S_2}^2 - \big\|\pi_r(C_\alpha+\varepsilon)\big\|_{S_2}^2\Big) \;>\; 0 \quad \textit{for any } r < M.$$

Proposition 7 (i) says that the accuracy of the empirical projection in case $C = U\alpha\mathrm{Id}V'$ is always worse than in case $C = 0$. (ii) complements this message by a lower bound on the accuracy: $\mathbb{E}\|\hat\pi_r X\|_{S_2}^2 - \mathbb{E}\|\pi_r X\|_{S_2}^2$ explodes (at least) linearly in the amplitude $|\alpha|$ for $|\alpha| \to \infty$. Inspection of the difference

$$\begin{aligned}&\mathbb{E}\big\|\hat\pi_r(C+\varepsilon)\big\|_{S_2}^2 \;-\; \mathbb{E}\big\|\pi_r(C+\varepsilon)\big\|_{S_2}^2\\ &= \mathbb{E}\sup_{\tilde\pi_r\in\mathcal{S}_{M,r}} \Big\{\big\|\tilde\pi_r\varepsilon\big\|_{S_2}^2 - \big\|\pi_r\varepsilon\big\|_{S_2}^2 + 2\,\mathrm{tr}\Big(\varepsilon'(\tilde\pi_r-\pi_r)C\Big) - \Big(\big\|\pi_r C\big\|_{S_2}^2 - \big\|\tilde\pi_r C\big\|_{S_2}^2\Big)\Big\}.\end{aligned}$$

shows that there is no *deterministic compensation term* in case $C = U\alpha\mathrm{Id}V'$: $\|\pi_r C\|_{S_2}^2 - \|\tilde\pi_r C\|_{S_2}^2 = 0$ for every $\pi_r \in \mathcal{S}_{M,r}$. Therefore, the case of constant singular value spectrum $C = U\alpha\mathrm{Id}V'$ is some *prototype of weak accuracy.*
The conjecture about the possibility of improvement for a certain type of matrices follows from the fact that, in contrast to the situation in the above Proposition 7, the differences $\|\tilde\pi_r(C+\varepsilon)\|_{S_2}^2 - \|\pi_r(C+\varepsilon)\|_{S_2}^2$ are usually not centered but have expectation less than zero. The next Theorem provides a general upper bound on the approximation of the reduced-rank projection for non-centered Gaussian random matrices $X = C + \varepsilon$ and covers the *large amplitude regime* in particular. It turns out that the characterization of the quality of approximation is expressed in terms of the "signal-to-noise ratio" $\tilde C := C/(\sigma\sqrt{M})$ rather than C. Correspondingly, $\tilde\lambda_1, ..., \tilde\lambda_M$ denote the singular values of $\tilde C$.

Theorem 8. *Let $(\varepsilon_{ij})_{i,j=1}^M$ be a centered matrix of independent Gaussian entries with variance σ^2. Then for any $C \in \mathbb{R}^{M\times M}$ with* $\mathrm{rank}(C) \ge r$, *$r \le M - r$, and*

$\{\tilde{\pi}_r \in \mathcal{S}_{M,r} : \|\tilde{\pi}_r C\|_{S_2}^2 = \|\pi_r C\|_{S_2}^2\} = \{\pi_r\}$, *the following bound holds true:*

$$\begin{aligned}
&\mathbb{E} \sup_{\tilde{\pi}_r \in \mathcal{S}_{M,r}} \Big(\|\tilde{\pi}_r(C+\varepsilon)\|_{S_2}^2 - \ \|\pi_r(C+\varepsilon)\|_{S_2}^2 \Big) \\
&\leq c\sigma^2 rM \Bigg\{ \min\left(\frac{(1+\tilde{\lambda}_1)^2}{\tilde{\lambda}_r^2}, (1+\tilde{\lambda}_1) \right) + \log(1+r) 1\Big(\tilde{\lambda}_r^2 \leq \log(1+r)(\sqrt{r}(1+\tilde{\lambda}_1)+r) \Big) \Bigg\} \\
&+ c\sigma^2 rM \Bigg\{ \left(\frac{\frac{1}{r}\sum_{i=r+1}^{2r} \lambda_i^2}{\lambda_r^2} \right)^{1/2} \cdot (1+\tilde{\lambda}_1) + \min\left[\sqrt{r} \left(\frac{\frac{1}{r}\sum_{i=r+1}^{2r} \lambda_i^2}{\lambda_r^2} \right)^{1/2}, \log(1+r) \right] \Bigg\}.
\end{aligned}$$

It is worth mentioning that the expression remains bounded in order by $\sigma^2 rM$ up to a logarithmic (in r) factor as long as

$$\frac{1+\tilde{\lambda}_1}{\tilde{\lambda}_r} \quad \text{and} \quad \left(\frac{\frac{1}{r}\sum_{i=r+1}^{2r} \tilde{\lambda}_i^2}{\tilde{\lambda}_r^2} \right)^{1/2} (1+\tilde{\lambda}_1)$$

stay uniformly bounded, and this may be possible even if $\|C\|_{S_\infty} \to \infty$. The most tractable situation arises for rank-r-matrices with rectangular singular value spectrum: In this case, the bound is of the order $\sigma^2 rM$ up to some logarithmic term which can be omitted as the amplitude tends to infinity.
The question remains whether this bound in case of rank-r-matrices with rectangular singular value spectrum is optimal. Our last result gives a positive answer:

Theorem 9. *Let $(\varepsilon_{ij})_{i,j=1}^M$ be a centered matrix of independent Gaussian entries with variance σ^2. Let $C_{\alpha,s} \in \mathbb{R}^{M\times M}$ with singular value decomposition $U\Lambda_{\alpha,s}V'$, where $\Lambda_{\alpha,s} = \alpha \mathrm{Id}_s$ and $1 \leq s < M$. Id_s denotes the $s \times s$-identity, canonically embedded into $\mathbb{R}^{M\times M}$. Then*

$$\liminf_{|\alpha|\to\infty} \max_{s\in\{r,M-r\}} \mathbb{E}\left(\sup_{\tilde{\pi}_s \in \mathcal{S}_{M,s}} \big\| \tilde{\pi}_s(C_{\alpha,s}+\varepsilon) \big\|_{S_2}^2 - \big\| \pi_s(C_{\alpha,s}+\varepsilon) \big\|_{S_2}^2 \right) \ \geq \ c\sigma^2 r(M-r).$$

The same result without the max over over $\{r, M-r\}$ is established for $s = r = 1$.

References

[1] Z.D. Bai and Y.Q. Yin, *Limit of the smallest eigenvalue of a large dimensional sample covariance matrix*, Ann. Probab. **21** (1993), 1275–1294.

[2] M. Capitaine, C. Donati-Martin and D. Féral, *The largest eigenvalue of finite rank deformations of large Wigner matrices: convergence and non universality of the fluctuations*, Ann. Probab. **37** (2009), 1–47.

[3] M. Capitaine, C. Donati-Martin and D. Féral, *Central limit theorems for eigenvalues of deformations of Wigner matrices*, Ann. Inst. H. P. **to appear** (2012).

[4] A. Pizzo, D. Renfrew and A. Soshnikov, *On finite rank deformations of Wigner matrices*, Ann. Inst. H. P. **to appear** (2012).

[5] T. Tao, *Outliers in the spectrum of iid matrices with bounded rank perturbations*, Probab. Theory and Relat. Fields **to appear** (2012).

Some remarks on the problem of bias in Bayesian semi-parametrics

Judith Rousseau

There has been an increasing literature in the past ten years on asymptotic properties of Bayesian nonparametric procedures, initiated mostly by the work of [2] on posterior concentration rates for density estimation. Now in many nonparametric models and for quite a large range of families of priors bounds have been obtained on posterior concentration rates when the (pseudo) metric which is considered is "comparable" with the Kullback-Leibler divergence. Let X^n denote the observations where n represents a measure of information brought by the data, such as the sample size when n observations are observed or the inverse of the variance of the noise in a white noise model. Consider a sampling model P_θ^n for the observations X^n conditionally on a parameter value θ. Given a loss (metric) $d(.,.)$ on θ, the posterior concentration rate on θ is defined as the smallest rate ϵ_n such that there exists $M > 0$ with

$$P^\pi[d(\theta, \theta_0) \leq M\epsilon_n | X^n] = 1 + o_p(1) \tag{1}$$

where $P^\pi[.|X^n]$ denotes the posterior distribution given the observations X^n. This type of problems has been studied in the last 10 years for various types of models. Typical models are :

- The density model.
  The observations X^n are independently and identiacally distributed from a distrbution having a density f_θ with respect to some fixed measure μ and $d(\theta, \theta')$ is either the Hellinger or the L_1 distance between the densities.
- The regression or the white noise model : the unknown parameter is the regression function (or the signal) and potentially the variance of the noise. In this case the loss function is usually the L_2 norm, or possibly the empirical L_2 norm over the design points in the case of the regression.
- Stationary gaussian time series : in this case the unknown parameter of interest is the spectral density. In this case the loss function is often

  $$d(\theta, \theta') = \int_{-\pi}^{\pi} (\log f_\theta - \log f_{\theta'})^2(x)dx$$

  where f_θ is the spectral density.

There are many variants of those models that have also been studied, however for the purpose of the present work I will only mention those three. In particular for these models general conditions on the priors have been established to derive an upper bound on the posterior concentration rate defined in (1) and some families of priors have been considered, see for instance [2], [3], [7]. For instance in the case of density estimation under the Hellinger or the L_1 loss adaptive minimax concentration rates have been obtained with nonparametric priors based on Gaussian or Beta mixtures, when the density belongs to a Hölder class with smoothness α, when α is not known in advance, see [4]. Similarly [8] have obtained similar results for scaled Gaussian process priors. This means that in both cases the prior does not depend on the unknown smoothness α (nor on the radius of the Hölder ball

the density is assumed to belong to) and that for each $\alpha > 0$ and each true density f_0 belonging to a Hölder class with smoothness α the posterior concentration rate was proved to be bounded by $(n/\log n)^{-\alpha/(2\alpha+1)}$ which is the minimax rate up to a $\log n$ term. Moreover, adaptation is obtained by very natural priors leading to procedures which are relatively easy to implement. The same featureis true for the other two types of models.

This phenomenon becomes untrue when the pseudo-metric d does not compare well with Kullback-Leibler.

Three examples are considered. First consider the white noise model which we write as the infinite sequence model:

$$Y_i = \theta_i + \epsilon_i/\sqrt{n}, \quad i \in \mathbb{N}, \boldsymbol{\theta} = (\theta_i, i \in \mathbb{N}) \in \ell_2,$$

using an expansion of the observed signal on an orthonormal basis of $L_2[0,1]$, for instance the Fourier basis. Consider the following hierarchical prior on $\boldsymbol{\theta}$: Let $k \in \mathbb{N}$ follow a Poisson or Geometric prior and given k, $\boldsymbol{\theta}$ is distributed as

$$\theta_i/\tau_i \sim \mathcal{N}(0,1), \quad i \leq k, \quad \theta_i = 0 \quad i > k \tag{2}$$

independently, where $\tau_i = i^{-\alpha}$ for some $1 > \alpha > 1/2$. Then for all $\boldsymbol{\theta}_0$ in a Sobolev ball with smoothness $\beta > 1/2$ the posterior distribution on $\boldsymbol{\theta}$ concentrates around $\boldsymbol{\theta}_0$ at the rate $(n/\log n)^{-\beta/(2\beta+1)}$ in terms of the L_2 loss and under $P_{\boldsymbol{\theta}_0}$. However if the parameter of interest is $\psi = \sum_i \theta_i$ which corresponds to the signal function computed at 0 in the case of a Fourier basis, then the posterior concentration rate around $\psi_0 = \sum_i \theta_{i0}$ is of order $n^{-(\beta-1/2)/(2\beta+1)}$ instead of $n^{-(\beta-1/2)/(2\beta)}$ which is the minimax rate in this case. The reason is that the posterior in the truncation parameter k concentrates on $k \leq k_n = O(n^{1/(2\beta+1)})$ which is the optimal (up to $\log n$) threshold for the L_2 norm and not to $k'_n = O(n^{1/(2\beta)})$ which would be optimal for estimating ψ. The reason behind this behaviour is that the posterior is mainly driven by the Kullback-Leibler loss, which in this case corresponds to the L_2 loss. This result is presented in [1].

Interstingly the same phenomenon occures for the estimation of the spectral density in stationary Gaussian long memory processes. Such models can be written as :

$$X^n \sim \mathcal{N}(0, T_n(f)), \quad (T_n(f))_{j_1,j_2} = cov(X_{j_1}, X_{j_2}) = \int_{-\pi}^{\pi} e^{i(j_1-j_2)x} f(x) dx$$

where f is the spectral density and in the case of long-memory f has the form

$$f(x) = (1 - \cos(x))^{-d} g(x) \approx |x|^{-2d} g(x)$$

near 0, where g is a positive and continuous function on $[0,1]$. The asymptotic in such models is slightly different than usual since the autocovariances are not summable and the minimax estimation of d depends on the smoothness of the short memory part g. Consider a prior constructed on the expansion of $\log g$ on

th Fourier basis on $[-\pi, \pi]$:

$$f_{\boldsymbol{\theta},k,d}(x) = (1-\cos x)^{-d} \exp\left(\sum_{j=0}^{k} \theta_j \cos(jx)\right), \quad k \in \mathbb{N}$$

where a Poisson distribution is chosen on k and given k the θ_j's follow (2). If $\boldsymbol{\theta}_0$ belongs to a Sobolev ball with smoothness $\beta > 1$ (in fact $1/2$) then the posterior distribution on f concentrates at the rate $(n/\log n)^{-2\beta/(2\beta+1)}$ in terms of the loss $l(f, f') = \|\log f - \log f'\|_2^2$, and under $P_{\boldsymbol{\theta}_0}$. We thus obtain the adaptive minimax concentration rate (up to a $\log n$ term).However the posterior concentration rate to estimate d in this case is $(n/\log n)^{-(\beta-1/2)/(2\beta+1)}$ instead of optimal minimax rate $n^{-(\beta-1/2)/(2\beta)}$. The reason is exactly the same as in the case of the white noise model above. This result is presented in [6]

These two semi-parametric problems are considered as irregular since the minimax rate for estimating the finite dimensional parameter of interest is slower than the usual $1/\sqrt{n}$. However the problems encountered by some Bayesian semiparametric priors is not restricted to nonregular cases. Indeed, in the case of the density model, if the parameter of interest is the cumulative distribution function at a given point, the posterior distribution can behave strangely although it has avery good behaviour for estimating the whole density under the Hellinger loss. Consider a prior based on the following representation of the density :

$$f_{\boldsymbol{\theta}}(x) = \exp\left(\sum_{j=0}^{k} \theta_j \phi_j(x)\right)$$

where ϕ_j is the Fourier basis on $[0, 1]$, where k follows a Poisson distribution and given k $\boldsymbol{\theta}$ follows the same distribution as in (2). Then for all $\beta > 1/2$, there exists $\boldsymbol{\theta}_0$ belonging to a Sobolev ball with smoothness β such that the posterior distribution of $F_\theta(x_1)$ for a given x_1 and with F_θ denoting the cumulative distribution function concentrates at a rate which is bounded from below by a constant times $\sqrt{\log n}/\sqrt{n}$. Hence although the prior leads to adaptive estimation of the density f_θ in terms of Hellinger distance it is suboptimal for estimating the cumulative distribution function at a given point, which is a smooth functional of the density.

These three examples are all based on the same pattern although they concern very different models. Suboptimality appear in these semi-parametric frameworks when there is a conflict between optimal approximation schemes under losses that are close to Kullback and optimal approximation schemes under losses that are different. The question of existence of optimal procedures that would adapt (up to $\log n$ terms) both in these global and local losses is still open.

References

[1] Arbel, J. and Gayraud, G. and Rousseau, J. *Bayesian optimal adaptive estimation using a sieve prior*, Preprint. arXiv:1204.2392

[2] Ghosal, S. and Ghosh, J.K. and van der Vaart, A.W., *Convergence rates of posterior distributions*, Annals of statistics (2000), **28**, p 500–531.

[3] Ghosal, S. and van der Vaart, A., *Convergence rates of posterior distributions for non iid observations*, Annals of statistics (2007) **35**, p 192–225.
[4] Kruijer, W. and Rousseau, J. and van der Vaart, A., *Adaptive Bayesian Density Estimation with Location-Scale Mixtures*, Electronic journal of statistics (2009)
[5] Rousseau, Judith and Kruijer, Willem, *Adaptive Bayesian estimation of a spectral density* (2011) Preprint.
[6] Kruijer, W. and Rousseau, J. *Bayesian semi-parametric estimation of the long-memory parameter under FEXP-priors* Preprint arXiv:1202.4863
[7] Rousseau, Judith and Chopin, Nicolas and Liseo, Brunero, *Bayesian nonparametric estimation of the spectral density of a long memory Gaussian process*, Annals of Statistics, (2012) To appear.
[8] van der Vaart, A. and van Zanten, J. H., *Adaptive Bayesian estimation using a Gaussian random field with inverse Gamma bandwidth*, Annals of Statistics, **37** (2009), 2655-2675.

Convex Variational Regularization Methods for Inverse Problems

Otmar Scherzer
(joint work with Clemens Kirisits)

In this talk we give an overview on convex variational methods for imaging processing and for the solution of linear inverse problems, such as inversion of the Radon transform.

The first observation is that gray-valued images (we only concentrate on such) can be described in various manners:

(1) Images can be considered as functions from $\mathbb{R}^2$ into $\mathbb{R}$,
(2) as graphs in $\mathbb{R}^3$.
(3) They can also be described via their level sets. These are three *infinite dimensional* formulations.
(4) They can be described as matrices with real valued entries.
(5) Moreover, they can be described *fully discrete* as functions from $\{1,\dots,N\}^2$ (pixels) to $\{0,\dots,255\}$ (discrete intensities),
(6) or as *ordered binary tensors* $\{0,1\}^{N\times N\times 256}$, where for each $(i,j)\in\{1,\dots,N\}^2$ the corresponding subrow of the tensor consists of first ones and second zeros. The sum of this vector defines the intensity of the image.

The talk is concerned with the infinite dimensional setting, where images are considered as functions from $\mathbb{R}^2$ into $\mathbb{R}$, and the fully discrete setting, where they are regarded as ordered binary tensors $\{0,1\}^{N\times N\times 256}$.

In the infinite dimensional setting we give an overview on recent regularization results from our work [11, 8, 9]. Convex variational regularization consists in minimizing the functional

$$\boxed{\mathcal{T}_\alpha(u) := \frac{1}{p}\|Fu - v^\delta\|_V^p + \alpha\mathcal{R}(u)\,,} \tag{1}$$

where $F : U \to V$ is the forward operator mapping between Banach spaces U and V and where we have $1 \le p < \infty$. Moreover, $\mathcal{R} : U \to [0,+\infty]$ is a convex and

proper *stabilizing functional*. Under these conditions it is relatively straightforward to prove convergence and stability. However, convergence rates results for Tikhonov-regularized minimizers with convex minimizers have been established just recently [3, 1, 11], in an abstract setting. The convergence rates results have been applied to concrete applications, such as total variation and sparsity regularization, where $\mathcal{R}(u)$ denotes the total variation of u or the l^1-norm of the coefficients of a series expansion with respect to an orthonormal basis, respectively. One remarkable property of l^1-sparsity regularization is that it allows for optimal convergence rate in a norm of order δ, which can be recasted from the abstract convergence rates results from [11] under additional restricted injectivity conditions, see [11, 8, 9]. For numerical minimization the functionals and the functions u have to be discretized. This introduces discretization errors, which have to be taken into account in a numerical analysis [17]. By solving the discretized system, one recovers a matrix with typically real valued function values.

We go one step further and consider images with a discrete range of intensities. In the case of image denoising with total variation regularization, we aim to find

$$\arg\min \mathcal{T}_\alpha(u), \tag{2}$$

where $u : \{1,\dots,N\}^2 \to \{0,\dots,255\}$ is a discrete image defined on a grid $\{1,\dots,N\}^2$ with discrete intensity values $\{0,\dots,255\}$, and $\mathcal{T}_\alpha$ is a discretized version of the total variation functional, which can be derived in a sound way by utilizing the coarea and Cauchy-Crofton formulae [5, 12]. Combinatorial optimization algorithms, such as *graph cuts*, can then be employed to compute *exact* minima of the discrete functional [10, 5, 4, 7]. For the efficient implementation it is important to notice that, on the one hand, images can be represented as binary tensors, i.e. they are decomposable into binary levels, and that, on the other hand, the discrete total variation functional has the favourable property that its minimization can be formulated as a sequence of minimization problems for the levels of u. As a consequence, optimization algorithms can be applied to each level separately while reusing information in each step. Finally, after "gluing" the solutions together, the result resembles a total variation regularized image.

Basically, graph cut algorithms make use of the fact that certain functions can be interpreted as cut functions of suitably constructed graphs. In other words, the minimization of a given function is mapped to the problem of finding the minimum cut on a flow network. For image processing tasks this usually leads to graphs, where every vertex represents one pixel and proximity of pixels is indicated by weighted edges between them. Due to the max-flow min-cut theorem [6], the minimum cut problem is equivalent to computing the maximum flow through the network, which in turn can be done exactly in low-order polynomial time [2].

Our particular work in this context is concerned with image analysis on a *hexagonal* grid. That a rectangular lattice might not be the best choice for the sampling and processing of two-dimensional signals has been recognized at least half a century ago [15, 13]. Since then, researchers who investigated the use of hexagonally arranged pixel configurations almost unanimously conclude that they should be

preferred over their more common counterparts for a wide variety of applications, such as edge detection, shape extraction or image reconstruction (see e.g. [14]). Total variation denoising on hexagonal grids and its approximation properties are analyzed in [16], and confirm the former observation by some limiting arguments. The finite dimensional implementation and realization with graph cuts is analyzed in [12].

References

[1] Y. Alber and I. Ryazantseva. *Nonlinear Ill-posed Problems of Monotone Type*. Springer Verlag, Dordrecht, 2006.

[2] Y. Boykov and V. Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(9):1124–1137, September 2004.

[3] M. Burger and S. Osher. Convergence rates of convex variational regularization. *Inverse Probl.*, 20(5):1411–1421, 2004.

[4] A. Chambolle and J. Darbon. On Total Variation Minimization and Surface Evolution Using Parametric Maximum Flows. *Int. J. Comput. Vision*, 84(3):288–307, April 2009.

[5] J. Darbon and M. Sigelle. Image restoration with discrete constrained total variation. Part I: Fast and exact optimization. *J. Math. Imaging Vision*, 26(3):261–276, 2006.

[6] L. R. Ford and D. R. Fulkerson. Maximal flow through a network. *Canad. J. Math.*, 8(3):399–404, 1956.

[7] D. Goldfarb and W. Yin. Parametric Maximum Flow Algorithms for Fast Total Variation Minimization. *SIAM J. Sci. Comput.*, 31(5):3712–3743, October 2009.

[8] M. Grasmair, M. Haltmeier, and O. Scherzer. Sparse regularization with l^q penalty term. *Inverse Probl.*, 24(5):055020, 13, 2008.

[9] M. Grasmair, M. Haltmeier, and O. Scherzer. Necessary and sufficient conditions for linear convergence of l^1-regularization. *Comm. Pure Appl. Math.*, 64(2):161–182, 2011.

[10] D. S. Hochbaum. An efficient algorithm for image segmentation, Markov random fields and related problems. *J. ACM*, 48(4):686–701 (electronic), 2001.

[11] B. Hofmann, B. Kaltenbacher, C. Pöschl, and O. Scherzer. A convergence rates result for Tikhonov regularization in Banach spaces with non-smooth operators. *Inverse Probl.*, 23(3):987–1010, 2007.

[12] C. Kirisits, and O. Scherzer. Total Variation Denoising on Hexagonal Grids. work in progress, University of Vienna, Austria, 2012.

[13] R. M. Mersereau. The processing of hexagonally sampled two-dimensional signals. *Proc. IEEE*, 67(6):930–949, 1979.

[14] L. Middleton and J. Sivaswamy. *Hexagonal Image Processing: A Practical Approach*. Advances in Pattern Recognition. Springer, 2005.

[15] D. P. Petersen and D. Middleton. Sampling and reconstruction of wave-number-limited functions in N-dimensional Euclidean spaces. *Inform.and Control*, 5:279–323, 1962.

[16] C. Pöschl, E. Resmerita, and O. Scherzer. Finite dimensional approximation of total variation regularization with hexagonal pixels. work in progress, University of Vienna, Austria, 2012.

[17] C. Pöschl, E. Resmerita, and O. Scherzer. Discretization of variational regularization in Banach spaces. *Inverse Probl.*, 26(10):105017, 2010.

Obtaining Qualitative Statements in Deconvolution Models

Johannes Schmidt-Hieber
(joint work with Axel Munk and Lutz Dümbgen)

Introduction and model. Whereas pointwise estimation in density deconvolution is nowadays a well-studied problem, there has been some recent interest in construction of uniform confidence bands (cf. Bissantz et *al.* [1] and Lounici and Nickl [5]). Pointwise estimates are known to have very slow convergence rates and are highly sensitive to the choice of the bandwidth parameter. Moreover, in applications, pointwise reconstructions are usually not necessary, since one is often rather interested in qualitative features of the underlying density, such as the number of modes, confidence intervals for the modes, regions of increase and decrease and so on (for a real data example see [1], p.500). Therefore, it is natural to ask for an analysis of confidence statements. We do this, by extending the idea of multiscale inference, introduced by Dümbgen and Spokoiny [2] and Dümbgen and Walther [3] in regression and density estimation, to deconvolution problems. Suppose that we observe

$$Y_i = X_i + \epsilon_i, \quad i = 1, \dots, n,$$

where $\epsilon_1, \epsilon_2, \dots$ and $X_1, X_2, \dots$ are independent sequences of i.i.d. random variables. The densities of Y_1, X_1, and ϵ_1 are denoted by g, f, and f_ϵ, respectively.

We consider the moderately ill-posed case, where the decay of the Fourier transform of the error density, denoted by $\widehat{f_\epsilon}$, is polynomial. Together with the assumptions on the tail decay of the derivatives of $\widehat{f_\epsilon}$ from Fan [4], we can show that the inversion operator which maps g to f is essentially pseudo-differential. Since a large class of shape constraints can be expressed as pseudo-differential inequalities, this shows that deconvolution and shape analysis are very similar and can therefore nicely be composed. For simplicity, let us consider confidence statements for qualitative features which are related to monotonicity, i.e. regions of increase/decrease and modes. In this case, it is convenient to derive a multiscale result for f'. For that suppose that for all $(t,h) \in [0,1] \times (0,1]$ we can construct a (kernel-type) estimator $\widehat{f}'_{nh}(t)$ with expectation equals $\frac{1}{h}\int K\big(\frac{s-t}{h}\big) f'(s) ds$, where the kernel K is itself a pdf supported on $[0,1]$. In the following we focus on the statistical implications. In particular we show how multiscale statistics appears in this context.

Multiscale inference: If we can find a (random) function $b_h(t)$, that is measurable with respect to our observation vector, and satisfies with probability $1-\alpha$ for all $t \in [0,1]$, $|\widehat{f}'_{nh}(t) - \mathbb{E}\widehat{f}'_{nh}(t)| \le b_h(t)$, then

$$t \mapsto [\widehat{f}'_{nh}(t) - b_h(t), \widehat{f}'_{nh}(t) + b_h(t)] \tag{1}$$

is a uniform $1-\alpha$ confidence band for the mean function $t \mapsto \mathbb{E}\widehat{f}'_{nh}(t)$. Now, suppose we know how to choose a bandwidth h such that we slightly undersmooth, then, inflating the band (1) by a small number yields an asymptotically uniform

confidence band for f'. Such a confidence band already allows us to construct simultaneous confidence statements. For instance, we might conclude that f is monotone increasing on all points t, where the confidence band for f' lies completely in the upper half plane. The strength of a confidence band lies in the simple visual interpretability. However, such a procedure has also a number of drawbacks. It depends on a proper bandwidth selection and even if we could choose the bandwidth locally adaptive, this does not guarantee a high power of the test. To overcome these difficulties it could be more convenient to use a multiscale approach, which means to test on *all* scales simultaneously. A multiscale object does neither require a bandwidth selection step nor a distinction between undersmoothing and oversmoothing scales. In the following, let us describe the construction in more detail. Define the multiscale statistic

$$T_n := \sup_{(t,h)\in B_n} \frac{\sqrt{\log \frac{e}{h}}}{\log\left(e \log \frac{e}{h}\right)} \Big(\frac{|\widehat{f}'_{nh}(t) - \mathbb{E}\widehat{f}'_{nh}(t)|}{\widehat{\mathrm{Std}}\big(\widehat{f}'_{nh}(t)\big)} - \sqrt{2 \log \tfrac{e}{h}} \Big), \tag{2}$$

with B_n a subset of $[0,1]\times(0,1]$, $\widehat{\mathrm{Std}}\big(\widehat{f}'_{nh}(t)\big)$ an estimate of the standard deviation of $\widehat{f}'_{nh}(t)$, and e the Euler number. Now, suppose that we can find a distribution-free approximation of T_n, denoted by T_n^∞, such that $|T_n - T_n^\infty| = o_P(1)$ (uniformly over f, with f in a certain function space) and $\sup_n T_n^\infty < \infty$ a.s. Then, we can approximate the $(1-\alpha)$-quantile of T_n by the one from T_n^∞ and thus, in view of (1), we obtain, with probability $1-\alpha$, for all $(t,h) \in B_n$ (computable) bounds $b_h(t)$ such that $|\widehat{f}'_{nh}(t) - \mathbb{E}\widehat{f}'_{nh}(t)| \le b_h(t)$. This extends (1) to all scales h. Note that the calibration factors in (2) are motivated by Lévy's modulus of continuity of Brownian motion and ensure finiteness of T_n and T_n^∞ as well as that the supremum is attained uniformly over different scales.

To give a straightforward statistical application of the result, note that whenever $(t,h) \in B_n$ and $\widehat{f}'_{nh}(t) > b_h(t)$ we can conclude that with probability $1-\alpha$, $f(s_1) < f(s_2)$ for some points $s_1, s_2 \in [t, t+h]$. Hence, if we find such a tuple, we can reject the hypothesis that f is constant. To give another application, one can show that the multiscale approach leads to simultaneous confidence intervals for modes and inflection points of f, with length of the optimal order up to a $\log(n)$-factor.
As described above, the construction of $\widehat{f}'_{nh}$ depends on a kernel K. By using variational calculus, one can easily identify the optimal kernel as the density of a beta distributed random variable. Hence, the proposed method depends finally only on the choice of the confidence level α.

For a more general treatment of shape constraints in deconvolution models, let us refer to the article [6].

References

[1] N. Bissantz, L. Dümbgen, H. Holzmann, and A. Munk, (2007), *Nonparametric confidence bands in deconvolution density estimation*, J. Royal Statist. Society Ser. B **69**, 483–506.

[2] L. Dümbgen and V. Spokoiny (2001), *Multiscale testing of qualitative hypothesis*, Ann. Statist. **29**, 124–152.
[3] L. Dümbgen and G. Walther (2008), *Multiscale inference about a density*, Ann. Statist. **26**, 1758–1785.
[4] J. Fan (1991), *On the optimal rates of convergence for nonparametric deconvolution problem*, Ann. Statist. **19**, 1257–1272.
[5] K. Lounici and R. Nickl (2011), *Global uniform risk bounds for wavelet deconvolution estimators*, Ann. Statist. **39**, 201–231.
[6] J. Schmidt-Hieber, A. Munk, and L. Dümbgen (2012), *Multiscale methods for shape constraints in deconvolution: confidence statements for qualitative features*, Preprint arxiv.org/abs/1107.1404.

Bernstein von Mises Theorem for quasi-posterior

Vladimir Spokoiny

Let Π be a prior measure on the parameter set Θ. Recall that the posterior is the random measure on Θ describing the conditional distribution of $\boldsymbol{\vartheta}$ given $\boldsymbol{Y}$ and obtained by normalization of the product $\exp\{L(\boldsymbol{\theta})\}\Pi(d\boldsymbol{\theta})$. This relation is usually written as

$$\boldsymbol{\vartheta} \mid \boldsymbol{Y} \quad \propto \quad \exp\{L(\boldsymbol{\theta})\}\,\Pi(d\boldsymbol{\theta}). \tag{1}$$

Now we study the properties of the posterior measure. An important feature of our analysis is that $L(\boldsymbol{\theta})$ is not assumed to be the true log-likelihood. This means that a model misspecification is possible and the underlying data distribution can be beyond the considered parametric family. In this sense, the Bayes formula (1) describes a *quasi posterior*.

The *Bernstein - von Mises Theorem* states a result which is similar to the Fisher Theorem: the posterior being centered at $\tilde{\boldsymbol{\theta}}$ and properly standardized is nearly standard normal. This is very useful for constructing Bayesian credible sets.

However, practical applications of all mentioned results are limited: they require true parametric distribution, large samples and a fixed parameter dimension. Modern applications stimulate a further extension of the classical theory beyond the classical parametric assumptions. [15] offers a general approach which appears to be very useful for such an extension.

This paper also discusses the Bernstein - von Mises (BvM) Theorem for Gaussian priors. The Bayes approach with Gaussian priors is effectively equivalent to roughness penalization in the frequentist approach. In particular, the credible sets based on the posterior distribution are nearly equivalent to the confidence sets in the penalized maximum likelihood estimation and the size is determined by the total Fisher information; see e.g. [10, 11] and references therein. Both confidence and credible sets are asymptotically centered at the penalized MLE and both suffer from the bias. If the bias term is larger than the width of the confidence or credible set, the true parameter will not be included in any of these two sets with a large probability; cf. [5] or [13, 14].

The Bayesian nonparametrics is being intensively developed in the last years. There is a number of papers recently appeared. We mention [7], [8], [9] for high

dimensional linear models, [1], [12] for non-Gaussian models, [4] for the semiparametric version of the BvM result, [13], [3], [14] for the misspecified parametric case, among many others. However, all the mentioned results require some special parametric structure, mainly model linearity w.r.t. the parameter, as well as large samples. An extension to the case of a large parameter dimension relative to the sample size require to change the main tools and methods: asymptotic expansions have to be replaced by nonasymptotic bounds.

Here the main result of the paper:

Theorem 10. *Assume* (ED_0G), (ED_1G), *and* $(\mathcal{L}_0G)$ *on* $\Theta_{0,G}$. *Then for any* $\mathcal{A} \subseteq D_{\boldsymbol{\epsilon},G}\Theta_{0,G}$ *and* $\underline{\mathcal{A}} \subseteq D_{\underline{\boldsymbol{\epsilon}},G}\Theta_{0,G}$, *it holds on* $\Omega_G(\mathrm{x})$

$$\begin{aligned}
I\!P\big\{D_{\boldsymbol{\epsilon},G}(\boldsymbol{\vartheta}-\boldsymbol{\theta}^*_G) \in \mathcal{A} \mid \boldsymbol{Y}\big\} &\leq \exp\big\{\Delta_{\boldsymbol{\epsilon},G} + \kappa_{\boldsymbol{\epsilon},G} + \tau_{\underline{\boldsymbol{\epsilon}},G}\big\}\Phi\big(\boldsymbol{\xi}_{\boldsymbol{\epsilon},G}, \mathcal{A}\big), \\
I\!P\big\{D_{\underline{\boldsymbol{\epsilon}},G}(\boldsymbol{\vartheta}-\boldsymbol{\theta}^*_G) \in \underline{\mathcal{A}} \mid \boldsymbol{Y}\big\} &\geq \exp\big\{-\Delta_{\boldsymbol{\epsilon},G} - \kappa_{\boldsymbol{\epsilon},G} - \tau_{\boldsymbol{\epsilon},G} - \delta_{\boldsymbol{\epsilon},G}\big\}\Phi\big(\boldsymbol{\xi}_{\underline{\boldsymbol{\epsilon}},G}, \underline{\mathcal{A}}\big), \\
\frac{\int_{\Theta\setminus\Theta_{0,G}} \exp\big\{L_G(\boldsymbol{\theta},\boldsymbol{\theta}^*_G)\big\}d\boldsymbol{\theta}}{\int_{\Theta_{0,G}} \exp\big\{L_G(\boldsymbol{\theta},\boldsymbol{\theta}^*_G)\big\}d\boldsymbol{\theta}} &\leq \exp\{\diamondsuit_{\underline{\boldsymbol{\epsilon}},G}(\mathrm{r}_0) + \tau_{\underline{\boldsymbol{\epsilon}},G}\}\,\delta_{\boldsymbol{\epsilon},G}
\end{aligned}$$

where

$$\begin{aligned}
\Delta_{\boldsymbol{\epsilon},G} &\stackrel{\mathrm{def}}{=} \diamondsuit_{\varrho,G} + \diamondsuit_{\underline{\boldsymbol{\epsilon}},G} + \big(\|\boldsymbol{\xi}_{\boldsymbol{\epsilon},G}\|^2 - \|\boldsymbol{\xi}_{\underline{\boldsymbol{\epsilon}},G}\|^2\big)/2, \\
\kappa_{\boldsymbol{\epsilon},G} &\stackrel{\mathrm{def}}{=} \log\det D_{\underline{\boldsymbol{\epsilon}},G} - \log\det D_{\boldsymbol{\epsilon},G}, \\
\tau_{\boldsymbol{\epsilon},G} &\stackrel{\mathrm{def}}{=} -\log\Phi\big(\boldsymbol{\xi}_{\boldsymbol{\epsilon},G}, D_{\boldsymbol{\epsilon},G}\Theta_{0,G}\big) \\
\delta_{\boldsymbol{\epsilon},G} &\stackrel{\mathrm{def}}{=} \frac{\det(D_{\underline{\boldsymbol{\epsilon}},G})}{(2\pi)^{p/2}} \int_{\Theta\setminus\Theta_{0,G}} \exp\big\{-\mathrm{u}(\boldsymbol{\theta})\big\}d\boldsymbol{\theta}
\end{aligned}$$

The result involves some terms like $\Delta_{\boldsymbol{\epsilon},G}$, $\kappa_{\boldsymbol{\epsilon},G}$, $\tau_{\boldsymbol{\epsilon},G}$, $\delta_{\boldsymbol{\epsilon},G}$, which can be shown to be relatively small under standard conditions. The main message is the upper and lower Gaussian approximation of the posterior measure which is the nonasymptotic version of the Bernstein - von Mises result.

References

[1] Boucheron, S. and Gassiat, E. (2009). A Bernstein-von Mises theorem for discrete probability distributions. *Electron. J. Stat.*, 3:114–148.

[2] Boucheron, S. and Massart, P. (2011). A high-dimensional wilks phenomenon. *Probability Theory and Related Fields*, 150:405–433. 10.1007/s00440-010-0278-7.

[3] Bunke, O. and Milhaud, X. (1998). Asymptotic behavior of Bayes estimates under possibly incorrect models. *Ann. Statist.*, 26(2):617–644.

[4] Castillo, I. (2012). A semiparametric bernstein–von mises theorem for gaussian process priors. *Probability Theory and Related Fields*, 152:53–99. 10.1007/s00440-010-0316-5.

[5] Cox, D. D. (1993). An analysis of Bayesian inference for nonparametric regression. *Ann. Stat.*, 21(2):903–923.

[6] Fan, J., Zhang, C., and Zhang, J. (2001). Generalized likelihood ratio statistics and Wilks phenomenon. *Ann. Stat.*, 29(1):153–193.

[7] Freedman, D. (1999). On the Bernstein-von Mises theorem with infinite-dimensional parameters. *Ann. Stat.*, 27(4):1119–1140.

[8] Ghosal, S. (1999). Asymptotic normality of posterior distributions in high-dimensional linear models. *Bernoulli*, 5(2):315–331.

[9] Ghosal, S. (2000). Asymptotic normality of posterior distributions for exponential families when the number of parameters tends to infinity. *J. Multivariate Anal.*, 74(1):49–68.
[10] Ghosal, S., Ghosh, J. K., and van der Vaart, A. W. (2000). Convergence rates of posterior distributions. *Ann. Statist.*, 28(2):500–531.
[11] Ghosal, S. and van der Vaart, A. (2007). Convergence rates of posterior distributions for noniid observations. *Ann. Statist.*, 35:192.
[12] Kim, Y. (2006). The Bernstein-von Mises theorem for the proportional hazard model. *Ann. Statist.*, 34(4):1678–1700.
[13] Kleijn, B. J. K. and van der Vaart, A. W. (2006). Misspecification in infinite-dimensional Bayesian statistics. *Ann. Statist.*, 34(2):837–877.
[14] Kleijn, B. J. K. and van der Vaart, A. W. (2012). The bernstein-von-mises theorem under misspecification. *Electronic J. Statist.*, 6:354–381.
[15] Spokoiny, V. (2011). Parametric estimation. finite sample theory.

Separable regularization penalties and structured sparsity

Sara van de Geer

We consider the linear model

$$Y = X\beta^0 + \epsilon,$$

where $Y \in \mathbb{R}^n$ is a response variable, X is an $n \times p$ matrix of covariables, $\beta^0 \in \mathbb{R}^p$ is an unknown vector of coefficients, and $\epsilon \in \mathbb{R}^n$ is a noise vector.

Let Ω be some norm on $\mathbb{R}^p$, and let $\hat{\beta}$ be the norm-penalized estimator

$$\hat{\beta} := \hat{\beta}_\Omega := \arg\min_{\beta \in \mathbb{R}^p} \left\{ \|Y - X\beta\|_2^2/n + 2\lambda\Omega(\beta) \right\}.$$

The parameter $\lambda > 0$ is a tuning parameter. Our aim is now to show that the estimator mimics an oracle which knows the sparsity structure of the unknown vector β^0.

For an index set $S \subset \{1, \ldots, p\}$, we use the notation

$$\beta_{j,S} := \beta_j 1\{j \in S\},\ j = 1, \ldots, p.$$

Definition *Fix some set S. We say that the norm Ω is separable for S if there exists a norm Ω^{S^c} on $\mathbb{R}^{p-|S|}$ such that for all $\beta \in \mathbb{R}^p$,*

$$\Omega(\beta) \geq \Omega(\beta_S) + \Omega^{S^c}(\beta_{S^c}).$$

We then take Ω^{S^c} as large as possible:

$$\nexists\ \tilde{\Omega}^{S^c} :\ \left\{ \tilde{\Omega}^{S^c}(\beta_{S^c}) \geq \Omega^{S^c}(\beta_{S^c})\ \forall\ \beta_{S^c} \right\}.$$

Examples of separable norms are the ℓ_1-norm and the group Lasso norm, more generally, the structured sparsity norm, as introduced by [2], which is defined as

$$\Omega(\beta) := \Omega(\beta; \mathbf{A}) := \min_{a \in \mathbf{A}} \frac{1}{2} \sum_{j=1}^p \left(\frac{\beta_j^2}{a_j} + a_j \right),$$

where $\mathbf{A} \subset [0, \infty)^p$ be a given convex cone, satisfying $\mathbf{A} \cap (0, \infty)^p \neq \emptyset$.

The following definition extends the notion of compatibility constant ([3]) or of restricted eigenvalue ([1]).

Definition *Suppose Ω is separable for S. Let $L > 0$ be some constant. The Ω-eigenvalue (for S) is*

$$\delta^2_\Omega(L,S) := \min\Big\{ \|X\beta_S - X\beta_{S^c}\|_2^2/n : \ \Omega(\beta_S) = 1, \ \Omega^{S^c}(\beta_{S^c}) \leq L \Big\}.$$

The Ω-effective sparsity is $\Gamma^2_\Omega(L,S) := 1/\delta^2_\Omega(L,S)$.

The dual norm of Ω is denoted by Ω_*, that is

$$\Omega_*(w) := \sup_{\Omega(\beta)\leq 1} |w^T\beta|, \ w \in \mathbb{R}^p.$$

We moreover let $\Omega_*^{S^c}$ be the dual norm of Ω^{S^c}.

Theorem *Let $S_0 \supset \{j : \beta^0_j \neq 0\}$ and let Ω be separable for S_0. Define*

$$\lambda^{S_0} := \Omega_*\Big((\epsilon^T X)_{S_0}/n\Big), \ \lambda^{S_0^c} := \Omega_*^{S_0^c}\Big((\epsilon^T X)_{S_0^c}/n\Big).$$

Suppose

$$\lambda > \lambda^{S_0^c}.$$

Define

$$L := \frac{\lambda + \lambda^{S_0}}{\lambda - \lambda^{S_0^c}}.$$

Then

$$\|X(\hat\beta - \beta^0)\|_2^2/n \leq 4(\lambda + \lambda^{S_0})^2\Gamma^2_\Omega(L,S_0),$$

$$\Omega(\hat\beta_{S_0} - \beta^0_{S_0}) \leq 2(\lambda + \lambda^{S_0})\Gamma^2_\Omega(L,S_0),$$

and

$$\Omega^{S_0^c}(\hat\beta) \leq 2L(\lambda + \lambda^{S_0})\Gamma^2_\Omega(L,S_0).$$

The above theorem is a generalization of the sparsity oracle inequalities for the Lasso and group Lasso.

References

[1] P. Bickel, P. Y. Ritov, Y. and A. Tsybakov, *Simultaneous analysis of Lasso and Dantzig selector*, Annals of Statistics **37** (2009), 1705–1732.

[2] C.A. Micchelli, J.M. Morales and M. Pontil, *A family of penalty functions for structured sparsity*, Advances in Neural Information Processing Systems, NIPS **23** (2010), 1612–1732.

[3] S.A. van de Geer, *The deterministic Lasso*, JSM proceedings, **140** (2007).

Gaussian priors and Credible Sets

Aad van der Vaart
(joint work with Bartek Knapik, Suzanne Sniekers, Botond Szabo, Harry van Zanten)

We model a function or surface a-priori as the sample path of a Gaussian process, and next by the usual Bayesian machine combine this with the likelihood to produce a *posterior distribution* for the function given the data. This posterior distribution can be visualized by plotting the posterior mean, and/or a number of realizations, and/or *posterior credible bands*. Pointwise versions of the latter are defined by computing for each argument lower and upper quantiles of the marginal posterior distribution of the function value at that argument.

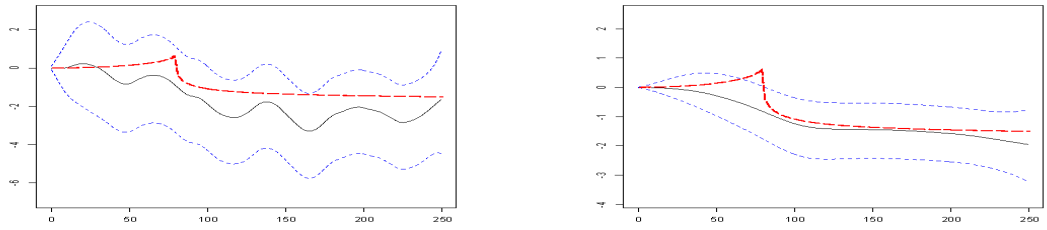
We ask the questions:

- Does this give good constructions?
- Do credible sets give a correct sense of remaining uncertainty?

As an illustration consider *nonparametric logistic regression* with integrated Brownian motion as a prior. In the Bayesian model unknown function and data are generated according to

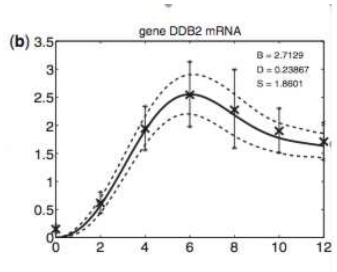
$$\begin{cases} \theta \sim \text{ scaled integrated Brownian motion,} \\ (X_1, Y_1), \ldots, (X_n, Y_n) | \theta \sim \text{ i.i.d. : } \Pr(Y_i = 1 | X_i = x) = 1/(1 + e^{-\theta(x)}). \end{cases}$$

The *posterior distribution* is the law of the function θ given $(X_1, Y_1), \ldots, (X_n, Y_n)$. Integrated Brownian motion is just one example of many possible priors: Brownian motion is an obvious example of a Gaussian process, but it might be considered to rough as a model for the unknown function, whence it is integrated once. The figure illustrates its application on a simulated data set.



Simulation experiment ($n = 250$). Two realisations of the posterior mode (black, solid) and 95 % posterior credible bands (blue, dotted), overlaid with true curve θ_0 (red, dashed). Two different scalings of IBM. Computations by the INLA package.

The bands in the figure supposedly give an impression of the remaining uncertainty in the estimate. Such bands are standard output of any Bayesian analysis. The following figure gives a real data example.



Nonparametric Bayesian analysis in *genomics*. Estimated abundance of a transcription factor as function of time: posterior mean curve and 95% credible bands. From Gao et al. *Bioinformatics*, 2008, 70–75.

Our interest is to investigate the validity of doing this. For this we leave the Bayesian model, which assumes that θ is random, and assume that the data are generated according to a given true function θ_0. We then investigate whether the posterior distribution puts most of its mass around this function, and whether the bands cover the function with a given probability.

In an asymptotic framework the data Y_n has a density $y \mapsto p_n(y|\theta)$ that depends on a parameter n, and the posterior distribution is given by

$$d\Pi_n(\theta|Y_n) \propto p_n(Y_n|\theta)\, d\Pi(\theta).$$

The *rate of contraction* of the posterior distribution is defined to be (at least) $\epsilon_n = \epsilon_n(\theta_0)$ if, for every $M_n \to \infty$,

$$\mathrm{E}_{\theta_0}\Pi_n\big(d(\theta,\theta_0) > M_n\epsilon_n|Y_n\big) \to 0.$$

A *credible set* is a a set $C(Y_n)$ with $\Pi_n\big(C(Y)|Y_n\big) = 0.95$. The *coverage* of the credible region $C_n(Y_n)$ is

$$\Pr_{\theta_0}\big(C_n(Y_n) \ni \theta_0\big).$$

Does it tend to 95 %?

We established a number of results showing that the rate of contraction depends on the fine properties of the Gaussian prior. One result is a randomly time-scaled Gaussian process with analytic sample paths yields a posterior distribution that adapts its rate of contraction to the smoothness of the true function.

Surprisingly the question of coverage of credible sets has so far hardly been studied. One of the few authors to consider the second question, Cox (1993) wrote

Non-Bayesians often find such Bayesian procedures attractive because as $n \to \infty$, the frequentist coverage probability of the Bayesian regions tends to the posterior coverage probability in "typical" cases. It was my hope that this would also hold in the nonparametric setting $[\cdots]$ Unfortunately, the hoped for result is false in about the worst possible way, viz.,

$$\liminf_{n\to\infty} P\big[\|\hat\theta - \theta\|^2 \le \Delta_n|\theta\big] = 0, \quad \text{a.s.}.$$

The notation $\hat\theta$ refers here to the posterior mean, in a special conjugate Gaussian problem, and $\{\theta : \|\hat\theta - \theta\|^2 \le \Delta_n\}$ is a credible ball in this problem. The assertion is that for almost every θ_0 according to the prior distribution, the coverage is zero, which is a disturbing message.

We were able to establish that the problem is actually one of a failing bias-variance-spread trade-off. If the (Gaussian) prior charges only functions that are smoother than the true function, then the posterior mean tends to be relatively far from the truth, while at the same time the spread of the posterior is small, as the set of smooth functions is relatively small. This creates coverage zero. On the other hand if the prior undersmoothes the truth, then the coverage is one. Furthermore, credible balls have the same order of magnitude as correct frequentist confidence balls.

We established these findings so far only for a number of special models. Future research effort is focused on proving these results in general, and to illuminate the role of Bayesian adaptation of the prior, by hierarchical or empirical Bayes methods.

References

[1] Cox, D. D. An analysis of Bayesian inference for nonparametric regression. *Ann. Statist. 21*, 2 (1993), 903–923.

[2] Knapik, B. T., van der Vaart, A. W., and van Zanten, J. H. Bayesian inverse problems with Gaussian priors. *Ann. Statist. 39*, (2011).

[3] A. W. van der Vaart and J. H. van Zanten. Adaptive Bayesian estimation using a Gaussian random field with inverse gamma bandwidth. *Ann. Statist. 37(5B)*, 2655–2675, 2009.

Statistical inference in high dimensions: From parametric to non-parametric

Martin J. Wainwright

(joint work with Alekh Agarwal, Sahand Negahban, Pradeep Ravikumar, Bin Yu)

We discuss some recent results in high-dimensional inference, which range from results on parametric estimation non-parametric regression. After providing an overview of various high-dimensional models, we present some theory for noisy matrix decomposition. The problem of noisy matrix decomposition is to recover a pair of matrices (Θ^*, Γ^*) based on observations of the form $Y = \mathfrak{X}(\Theta^* + \Gamma^*) + W$, where $\mathfrak{X}$ is a linear observation operator, and W is a noise matrix. This problem has applications in robust PCA, multitask regression, robust covariance estimation, and security-aware forms of matrix completion. We derive non-asymptotic bounds on the performance of a natural convex relaxation, and show that they are minimax-optimal for Gaussian additive noise [1]. We then describe some of the general theory that underlies results of this type, including the notion of a decomposable regularizer and restricted strong convexity [2]. Time permitting, we sketch some extensions of these techniques to obtain optimal rates for high-dimensional non-parametric regression [3].

References

[1] A. Agarwal, S. Negahban, and M. J. Wainwright. Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions. To appear in *Annals of Statistics*.
[2] S. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. Full length version at http://arxiv.org/abs/1010.2731v1.
[3] G. Raskutti, M. J. Wainwright, and B. Yu. Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *Journal of Machine Learning Research*, 12:389–427, 2012.

The Average Likelihood Ratio for Large-scale Multiple Testing and Detecting Sparse Mixtures

Günther Walther

Large-scale multiple testing problems require the simultaneous assessment of many p-values. This talk compares several methods to assess the evidence in multiple binomial counts of p-values: the maximum of the binomial counts after standardization (the 'higher-criticism statistic'), the maximum of the binomial counts after a log-likelihood ratio transformation (the 'Berk-Jones statistic'), and a newly introduced average of the binomial counts after a likelihood ratio transformation. Simulations show that the higher criticism statistic has a superior performance to the Berk-Jones statistic in the case of very sparse alternatives (sparsity coefficient $\beta \gtrapprox 0.75$), while the situation is reversed for $\beta \lessapprox 0.75$. This is due to the heavy long tail of the binomial distribution, which results in the higher criticism statistic assigning a large weight to the evidence contained in the smallest p-values, and these smallest p-values are known to be the relevant statistic in the very sparse case.

On the other hand, the log-likelihood transformation results in an equal weighting of the evidence contained in the various p-values, and this yields a better performance in the less sparse case where the evidence of mixing is most pronounced in p-values that are not among the smallest ones.

Finally, the average likelihood ratio is motivated via a minimax approach to obtain good detection power uniformly in the sparsity parameter. Simulations show that the average likelihood ratio does indeed combine the favorable performance of higher criticism in the very sparse case with that of the Berk-Jones statistic in the less sparse case and thus appears to dominate both statistics.

Adaptive Covariance Matrix Estimation Through Block Thresholding

Ming Yuan
(joint work with T. Tony Cai)

Covariance matrix estimation is of fundamental importance in multivariate analysis. Driven by a wide range of applications in science and engineering, the high dimensional setting, where the dimension p can be much larger than the sample

size n, is of particular current interest. In such a setting, conventional methods and results based on fixed p and large n are no longer applicable and in particular the commonly used sample covariance matrix and normal maximum likelihood estimate perform poorly. A number of regularization methods, including banding, tapering, thresholding, and ℓ_1 minimization, have been developed in recent years for estimating a large covariance matrix or its inverse.

One of the most commonly considered class of covariance matrices is the "bandable" matrices, where the entries of the matrix decay as they move away from the diagonal. More specifically, consider the following class of covariance matrices introduced in Bickel and Levina (2008):

$$\mathcal{C}_\alpha = \mathcal{C}_\alpha(M_0, M) := \Big\{ \Sigma : \max_j \sum_i \{|\sigma_{ij}| : |i-j| \geq k\} \leq M k^{-\alpha} \quad \forall k, \tag{1}$$
$$\text{and } 0 < M_0^{-1} \leq \lambda_{\min}(\Sigma), \lambda_{\max}(\Sigma) \leq M_0 \Big\}.$$

Several regularization methods have been introduced for estimating a bandable covariance matrix $\Sigma \in \mathcal{C}_\alpha$. In particular, Cai, Zhang and Zhou (2010) established the minimax rate of convergence for estimation over $\mathcal{C}_\alpha$ and introduced a tapering estimator $\bar{\Sigma} \circ T_k$ where the tapering matrix T_k is given by

$$T_k = \left(\frac{2}{k} \{(k - |i-j|)_+ - (k/2 - |i-j|)_+\} \right)_{1 \leq i,j \leq p},$$

with $(x)_+ = \max(x, 0)$ and $\circ$ stands for the Schur or element-wise product. It was shown that the tapering estimator $\bar{\Sigma} \circ T_k$ with $k \asymp n^{1/(2\alpha+1)}$ achieves the minimax optimal rate of convergence

$$\|\bar{\Sigma} \circ T_k - \Sigma\| = O_p\left(n^{-\frac{\alpha}{2\alpha+1}} + \left(\frac{\log p}{n} \right)^{\frac{1}{2}} \right) \tag{2}$$

uniformly over $\mathcal{C}_\alpha$.

The minimax rate of convergence in (2) provides an important benchmark for the evaluation of the performance of covariance matrix estimators. It is, however, evident from its construction that the rate optimal tapering estimator constructed in Cai, Zhang and Zhou (2010) requires explicit knowledge of the decay rate α which is typically unknown in practice. This naturally leads to the arguably more practically important question of adaptive estimation: Is it possible to construct a single estimator, not depending on the decay rate α, that achieves the optimal rate of convergence simultaneously? We shall show in this paper that the answer is affirmative. A fully data-driven adaptive estimator $\hat{\Sigma}$ is constructed and is shown to be simultaneously rate optimal over the collection of the parameter spaces $\mathcal{C}_\alpha$ for all $\alpha > 0$.

$\bar{\Sigma}$ into blocks and then simultaneously estimate the entries of Σ in a block by thresholding.

The adaptive covariance matrix estimator achieves its adaptivity through block thresholding of the sample covariance matrix $\bar{\Sigma}$. The idea of adaptive estimation

through block thresholding can be traced back to nonparametric function estimation using Fourier or wavelet series. However, the application of block thresholding to covariance matrix estimation poses new challenges. One of the main difficulties in dealing with covariance matrix estimation as opposed to function estimation or sequence estimation problems is the fact that the spectral norm is not separable in its entries. Another practical challenge is due to the fact that the covariance matrix is "two-directional" where one direction is along the rows and another along the columns. The blocks of different sizes need to be carefully constructed so that they fit well in the sample covariance matrix and the risk can be assessed based on their joint effects rather than their individual contributions. There are two main steps in the construction of the adaptive covariance matrix estimator. The first step is the construction of the blocks. Once the blocks are constructed, the second step is to estimate the entries of the covariance matrix Σ in groups and make simultaneous decisions on all the entries within a block. This is done by thresholding the sample covariance matrix block by block. The threshold level is determined by the location, block size and corresponding spectral norms.

We shall show that the proposed block thresholding estimator $\hat{\Sigma}$ is simultaneously rate-optimal over every $\mathcal{C}_\alpha$ for all $\alpha > 0$. The theoretical analysis of the estimator $\hat{\Sigma}$ requires some new technical tools that can be of independent interest. One is a concentration inequality which shows that although the sample covariance matrix $\bar{\Sigma}$ is not a reliable estimator of Σ, its submatrices could still be a good estimate of the corresponding submatrices of Σ. Another useful tool is a so-called Norm Compression Inequality which reduces the analysis on the whole matrix to a matrix of much smaller dimensions, whose entries are the spectral norms of the blocks.

References

[1] P. Bickel and E. Levina, *Regularized estimation of large covariance matrices*, The Annals of Statistics, **36** 2008, 199–227.

[2] T. Cai, C. Zhang and H. Zhou, *Optimal rates of convergence for covariance matrix estimation*, The Annals of Statistics, **38** 2010, 2118–2144.

Calibrated elastic regularization in matrix completion

Cun-Hui Zhang
(joint work with Tingni Sun)

Matrix completion concerns the estimation of a large matrix when a small fraction of it is observed. Consider an unknown matrix $\Theta \in \mathbb{R}^{d_1 \times d_2}$. Let $\Omega^* = \{1, \ldots, d_1\} \times \{1, \ldots, d_2\}$. Suppose we observe iid vectors (ω_i, y_i),

$$y_i = \Theta_{\omega_i} + \varepsilon_i,$$

where ω_i is uniformly distributed in Ω^* and $\varepsilon_i \sim N(0, \sigma^2)$.

Suppose Θ is of low rank. It seems natural to consider the nuclear penalized least squares estimator [5]. This estimator can be written as the minimizer of

$$\sum_{i=1}^{n} M_{\omega_i}^2/2 - \sum_{i=1}^{n} y_i M_{\omega_i} + \lambda \|M\|_{(N)}, \tag{1}$$

where $\|M\|_{(N)}$, the nuclear norm, is the sum of the singular values of M . However, analytical properties of the minimizer of (1) is unclear.

Error bounds for two modifications of minimizing (1) have been provided in [6] and [4]. Let r =rank(Θ) and $d = d_1 + d_2$. Consider $d_1 \le d_2$ without loss of generality. Define $\alpha_{(sp)}(M) = \|M\|_\infty \sqrt{d_1 d_2}/\|M\|_{(F)}$ as the spikiness of a matrix M, where $\|M\|_\infty = \max_{jk} |M_{jk}|$ is the vectorized supreme norm and $\|M\|_{(F)}$ is the Frobenius norm. For $\|\Theta\|_{(F)} \le 1$ and $\alpha_{(sp)}(\Theta) \le \alpha^*$, [6] proved

$$\|\widehat{\Theta}^{(NW)} - \Theta\|_{(F)}^2 \le C_0 \max(d_1 d_2 \sigma^2, 1)(\alpha^*)^2 r d (\log d)/n$$

with large probability, where $\widehat{\Theta}^{(NW)}$ is the minimizer of (1) under the constraint $\|M\|_\infty \le \alpha^*/\sqrt{d_1 d_2}$. Here and in the sequel, C_0 denotes a numerical constant. In [4], the quadratic term $\sum_{i=1}^n M_{\omega_i}^2/2$ in (1) is replaced by its expectation $\pi_0 \|M\|_{(F)}^2$, with $\pi_0 = n/(d_1 d_2)$, and the resulting minimizer is proved to satisfy

$$\|\widehat{\Theta}^{(KLT)} - \Theta\|_{(F)}^2/(d_1 d_2) \le C_0 \max(\sigma^2, \|\Theta\|_\infty^2) r d (\log d)/n$$

with large probability. We note that for $\|\Theta\|_{(F)} \le 1$, $d_1 d_2 \|\Theta\|_\infty^2 \le \alpha_{(sp)}^2(\Theta)$. The penalty level λ is of the order $\sigma\sqrt{\pi_0 d \log d}$ in [6] and $\max(\sigma, \|\Theta\|_\infty)\sqrt{\pi_0 d \log d}$ in [4]. In both cases, the sample size requirement is $n \ge C^* r d \log d$ with a factor C^* depending on the spikiness $\alpha_{(sp)}(\Theta)$. These results provide (nearly) optimal error bounds when σ and $\|\Theta\|_\infty$ are of the same order, but not for smaller noise level σ.

In [3], an error bound proportional to the noise level was obtained for the result $\widehat{\Theta}^{(KMO)}$ of a non-convex recursive algorithm based on the knowledge of the rank r. For $n \ge C_1^* r d \log d + C_2^* r^2 d \sqrt{d_2/d_1}$ with certain $\{C_1^*, C_2^*\}$ dependent on several coherence factors,

$$\|\widehat{\Theta}^{(KMO)} - \Theta\|_{(F)}^2/(d_1 d_2) \le C_0 (s_1/s_r)^4 \sigma^2 r d (\log d)/n,$$

where s_j is the j-th largest singular value of Θ. This provides continuity in the matrix completion theory between the noisy and noiseless cases [1]. However, the required sample size is large for large d_2/d_1.

The main difficulty with matrix completion theory is that the quadratic $\sum_{i=1}^n M_{\omega_i}^2$ is ill-posed. This has led to the modifications of (1) in [6] and [4]. We propose to consider calibrated elastic regularization as follows. In linear regression, the elastic net is the lease squares estimator with a sum of the ℓ_1 and ℓ_2 penalties [7]. For completing a low-rank Θ, the corresponding elastic regularized estimator is

$$\widetilde{\Theta} = \arg\min_M \Big\{ \sum_{i=1}^{n} M_{\omega_i}^2/2 - \sum_{i=1}^{n} y_i M_{\omega_i} + \lambda_1 \|M\|_{(N)} + (\lambda_2/2)\|M\|_{(F)}^2 \Big\} \tag{2}$$

Let $\Theta = USV^\top$ be the singular value decomposition of Θ with $S \in \mathbb{R}^{r \times r}$. Under proper coherence conditions on $\{\Theta, U, V, S\}$ and proper choices of penalty levels,

$\widetilde{\Theta}$ is approximately $\overline{\Theta} = (\pi_0 + \lambda_2)^{-1}(\pi_0\Theta - \lambda_1 UV^\top)$. This suggests correcting the bias of $\widetilde{\Theta}$ with the following calibrated elastic penalized least squares estimator

$$\widehat{\Theta} = (1 + \lambda_2/\pi_0)\widetilde{\Theta}. \tag{3}$$

We briefly describe our analytical result in the following theorem.

Theorem 1 *For proper choices of λ_1 and λ_2, the calibrated elastic penalized least squares estimator (3) satisfies*

$$\|\widehat{\Theta} - \Theta\|_{(F)}^2/(d_1 d_2) \le 10\sigma^2 rd(\log d)/n$$

with at least probability $1 - 1/d^2$, provided that $n \ge C^* r^2 d \log d$, where C^* is a constant depending only on the following coherence factors: $\alpha_{sp}(\Theta)$, $\alpha_{(sp)}(U)$, $\alpha_{(sp)}(V)$, and $\|\Theta\|_{(F)}/(r^{1/2}s_r)$.

Compared with [6] and [4], the error bound in Theorem 1 is proportional to the noise level. Compared with [3], Theorem 1 replaces the root aspect ratio $\sqrt{d_2/d_1} \ge 1$ with a log factor in the sample size requirement.

Let $T = \{UU^\top M_1 + M_2 VV^\top : M_j \in \mathbb{R}^{d_1 \times d_2}\}$ be the tangent space of the nuclear norm at Θ. Define $\mathcal{H}$ by $\langle \mathcal{H}M, M\rangle = \sum_{i=1}^n M_{\omega_i}^2$ with the trace inner product. Let $\mathcal{P}_T$ be the orthogonal projection from $\mathbb{R}^{d_1 \times d_2}$ to T. The key element in our analysis is to prove

$$\max_{\|\Delta\|_{(N)}=1} \langle (\mathcal{H} - \pi_0)(\mathcal{P}_T \mathcal{H} \mathcal{P}_T)^{-1}(\mathcal{H}/\pi_0 - 1)(\lambda_1 UV^\top + \lambda_2 \Theta), \mathcal{P}_T^\perp \Delta\rangle \le \lambda_1/2$$

with large probability. For $\lambda_2 = 0$, this was considered in [2] where the sample size requirement is $n \ge C_0 \min\{\mu^2 r^2 (\log d)^2 d, \mu^2 r (\log d)^6 d\}$ for a certain coherence factor μ stronger than $\max\{\alpha_{(sp)}(U), \alpha_{(sp)}(V)\}$. We are able to remove a log factor in the r^2 bound, resulting in the sample size requirement in Theorem 1. If the second bound in [2] (linear in r) or its proof is also applicable, our calculation can be modified to achieve the sample size requirement $n \ge C^* rd(\log d)^6$.

References

[1] E. Candes and B. Recht. (2009) Exact matrix completion via convex optimization. *Found. Comput. Math.* **9** 717-772.

[2] E. J. Candès and T. Tao. (2009). The Power of Convex Relaxation: Near-Optimal Matrix Completion. *IEEE Trans. Inform. Theory* **56**(5), 2053-2080.

[3] R. H. Keshavan, A. Montanari, and S. Oh. (2010). Matrix completion from noisy entries. *Journal of Machine Learning Research* **11** 2057-2078.

[4] V. Koltchinskii, K. Lounici, and A. B. Tsybakov (2011). Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics* **39** 2302-2329.

[5] R. Mazumder, T. Hastie, and R. Tibshirani. (2010). Spectral Regularization Algorithms for Learning Large Incomplete Matrices. *Journal of Machine Learning Research* **11** 2287-2322.

[6] S. Negahban and M. J. Wainwright. (2010). Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. http://arxiv.org/abs/1009.2118v2.

[7] H. Zou and T. Hastie (2005). Regularization and variable selection via the elastic net. *J. R. Statist. Soc. B*, **67** 301-320.

A New Minimax Lower Bound for Matrices Estimation

Huibin Zhou

Minimax risk is one of the most widely used benchmarks for optimality and substantial efforts have been made on developing minimax theories in the statistics literature. A key step in establishing a minimax theory is the derivation of minimax lower bounds and several effective lower bound arguments based on hypothesis testing have been introduced in the literature. Well known techniques include Le Cam's method, Assouad's Lemma and Fano's Lemma. See Le Cam (1986) and Tsybakov (2009) for more detailed discussions on minimax lower bound arguments.

Driven by a wide range of applications in high dimensional data analysis, estimation of large covariance matrices has drawn considerable recent attention. See, for example, Bickel and Levina (2008a, b), El Karoui (2008), Ravikumar, Wainwright, Raskutti and Yu (2008), Lam and Fan (2009), Cai and Zhou (2009), Cai, Zhang and Zhou (2010), and Cai and Liu (2011). Many theoretical results, including consistency and rates of convergence, have been obtained. However, the optimality question remains mostly open in the context of covariance matrix estimation under the spectral norm, mainly due to the technical difficulty in obtaining good minimax lower bounds.

We develop a minimax lower bound technique that is particularly well suited for treating "two-directional" problems such as estimating sparse covariance matrices. The result can be viewed as a simultaneous generalization of Le Cam's method in one direction and Assouad's Lemma in another. This general technical tool is of independent interest and is useful for solving several matrix estimation problems such as optimal estimation of sparse covariance, precision and volatility matrices.

We now introduce our new lower bound technique. Again, let $X \sim \mathbb{P}_\theta$ where $\theta \in \Theta$. The parameter space Θ of interest has a special structure which can be viewed as the Cartesian product of two components Γ and Λ. For a given positive integer r and a finite set $B \subset \mathbb{R}^p \setminus \{\mathbf{0}_{1\times p}\}$, let $\Gamma = \{0,1\}^r$ and $\Lambda \subseteq B^r$. Define

$$\Theta = \Gamma \otimes \Lambda = \{\theta = (\gamma, \lambda) : \gamma \in \Gamma \text{ and } \lambda \in \Lambda\}. \tag{1}$$

In comparison, the standard lower bound arguments work with either Γ or Λ alone. For example, Assouad's Lemma considers only the parameter set Γ and the Le Cam's method typically applies to a parameter set like Λ with $r = 1$. For $\theta = (\gamma, \lambda) \in \Theta$, denote the projection of θ to Γ by $\gamma(\theta) = \gamma$ and to Λ by $\lambda(\theta) = \lambda$.

It is important to understand the structure of the parameter space Θ. One can view an element $\lambda \in \Lambda$ as an $r \times p$ matrix with each row coming from the set B and view Γ as a set of parameters along the rows indicating whether a given row of λ is present or not. Let $D_\Lambda = \text{Card}(\Lambda)$. For a given $a \in \{0,1\}$ and $1 \le i \le r$, denote $\Theta_{i,a} = \{\theta \in \Theta : \gamma_i(\theta) = a\}$ where $\theta = (\gamma, \lambda)$ and $\gamma_i(\theta)$ is the $i-$th coordinate of of the first component of θ. It is easy to see that $\text{Card}(\Theta_{i,a}) = 2^{r-1} D_\Lambda$. Define the mixture distribution $\bar{\mathbb{P}}_{i,a}$ by

$$\bar{\mathbb{P}}_{i,a} = \frac{1}{2^{r-1} D_\Lambda} \sum_{\theta \in \Theta_{i,a}} \mathbb{P}_\theta \tag{2}$$

So $\bar{\mathbb{P}}_{i,a}$ is the mixture distribution over all $\mathbb{P}_\theta$ with $\gamma_i(\theta)$ fixed to be a while all other components of θ vary over all possible values in Θ.

The following result gives a lower bound for the maximum risk over the parameter set Θ of estimating a functional $\psi(\theta)$ belonging to a metric space with metric d.

For any $s > 0$ and any estimator T of $\psi(\theta)$ based on an observation from the experiment $\{\mathbb{P}_\theta, \theta \in \Theta\}$ where Θ is given in (1),

$$\max_{\Theta} 2^s \mathbb{E}_{\mathbf{X}|\theta} d^s\left(T, \psi\left(\theta\right)\right) \geq \alpha \frac{r}{2} \min_{1 \leq i \leq r} \left\| \bar{\mathbb{P}}_{i,0} \wedge \bar{\mathbb{P}}_{i,1} \right\| \tag{3}$$

where $\bar{\mathbb{P}}_{i,a}$ is defined in Equation (2) and α is given by

$$\alpha = \min_{\{(\theta,\theta'): H(\gamma(\theta),\gamma(\theta')) \geq 1\}} \frac{d^s(\psi(\theta), \psi(\theta'))}{H(\gamma(\theta), \gamma(\theta'))}. \tag{4}$$

The idea behind this new lower bound argument is similar to the one for Assouad's Lemma, but in a more complicated setting. Based on an observation $X \sim \mathbb{P}_\theta$ where $\theta = (\gamma, \lambda) \in \Theta = \Gamma \otimes \Lambda$, we wish to test whether $\gamma_i = 0$ or 1 for each $1 \leq i \leq r$. The first factor α in the lower bound (3) is the minimum cost of making an error per comparison. The second factor $r/2$ is the expected number of errors one makes to estimate γ when $\mathbb{P}_\theta$ and $\mathbb{P}_{\theta'}$ are indistinguishable from each other in the case $H\left(\gamma(\theta), \gamma(\theta')\right) = r$, and the last factor is the lower bound for the total probability of making type I and type II errors for each comparison. A major difference is that in this third factor the distributions $\bar{\mathbb{P}}_{i,0}$ and $\bar{\mathbb{P}}_{i,1}$ are both complicated mixture distributions instead of the typically simple ones in Assouad's Lemma. This makes the lower bound argument more generally applicable, while the calculation of the affinity becomes much more difficult.

In applications of the result, for a $\gamma = (\gamma_1, ..., \gamma_r) \in \Gamma$ where γ_i takes value 0 or 1, and a $\lambda = (\lambda_1, ..., \lambda_r) \in \Lambda$ where each $\lambda_i \in B$ is a p-dimensional nonzero row vector, the element $\theta = (\gamma, \lambda) \in \Theta$ can be equivalently viewed as an $r \times p$ matrix

$$\begin{pmatrix} \gamma_1 \cdot \lambda_1 \\ \gamma_2 \cdot \lambda_2 \\ \vdots \\ \gamma_r \cdot \lambda_r \end{pmatrix} \tag{5}$$

where the product $\gamma_i \cdot \lambda_i$ is taken elementwise: $\gamma_i \cdot \lambda_i = \lambda_i$ if $\gamma_i = 1$ and the ith row of θ is the zero vector if $\gamma_i = 0$. The term $\|\bar{\mathbb{P}}_{i,0} \wedge \bar{\mathbb{P}}_{i,1}\|$ of Equation (3) is then the lower bound for the total probability of making type I and type II errors for testing whether or not the ith row of θ is zero.

Note that the lower bound (3) reduces to the classical Assouad's Lemma when Λ contains only one matrix for which every row is nonzero, and becomes a two-point argument of Le Cam with one point against a mixture when $r = 1$. The technical argument is an extension of that of Assouad's Lemma. See Assouad (1983), Yu (1997) and van der Vaart (1998).

The advantage of this method is to break down the lower bound calculations for the whole matrix estimation problem into calculations for individual rows so

that the overall analysis is simplified and more tractable. Although the tool is introduced here for the purpose of estimating a sparse covariance matrix, it is of independent interest and is expected to be useful for solving other matrix estimation problems as well.

References

[1] T. Cai and H. Zhou, *Optimal Rates of Convergence for Sparse Covariance Matrix Estimation*, Annals of Statistics, to appear.

Regularized Semiparametric Estimation for Ordinary Differential Equations

Ji Zhu

(joint work with Yun Li, Naisyin Wang)

In engineering, physics and bio-medical sciences, dynamic systems are often modeled through a set of ordinary differential equations (ODEs). Most ODE dynamic systems are fully determined by the parameters and initial values. They usually have non-linear structures and non-trivial analytic solutions. Given the parameters and initial values, there exist various numerical methods to solve non-linear ODEs, including the well known family of Runge-Kutta methods. In reality, the parameters of an ODE system are often unknown and need to be estimated using the observed data.

Suppose that an ODE dynamic model has the following general structure:

$$\frac{dX}{dt} = F\{X(t), \theta, t\} \tag{1}$$

where $X(t) = \{X_1(t), \cdots, X_m(t)\}^T$ is the state vector (also referred as ODE curves) to describe the dynamic system, $\theta = (\theta_1, \cdots, \theta_d)^T$ denotes the unknown parameters to be estimated, and $F(\cdot) = \{F_1(\cdot), \cdots, F_m(\cdot)\}^T$ is a known force functional structure, which is usually highly non-linear. Instead of directly observing the true state vector $X(t)$, we assume that we observe the surrogate $Y(t)$ at discrete time points

$$Y_{ij} = Y_j(t_{ij}) = X_j(t_{ij}) + \varepsilon_{ij},\ i = 1, \ldots, n_j; j = 1, \ldots, m. \tag{2}$$

In most of the current statistics literature, the parameters θ are assumed as constants, and there are mainly two categories of methods for estimating the constant θ. The first category consists of various two-stage methods: one estimates the ODE curves $X(t)$ and their first derivatives in stage-one by a nonparametric smoothing fit to the data, and then, in the second stage, finds the parameter estimates through the classical least-square optimization with $X(t)$ and $dX(t)/dt$ replaced by the nonparametric estimates obtained from the first stage. For example, Varah (1982) estimated $X(t)$ and $dX(t)/dt$ using a spline smoothing technique in stage-one. Liang and Wu (2008) extended the work of Varah (1982) by using the local polynomial regression as the smoothing approach and they further provided statistical properties of the estimator. The use of non-parametric kernel

estimation was proposed and studied in Brunel (2008). These approaches can be easily implemented and can perform very well with moderate to large data sets with densely observed data points. However, if the level of observation noise is relatively high and/or the sample size is relatively small, the two-stage method may not be able to obtain sufficiently precise estimates of $dX(t)/dt$ in the first stage and consequently the estimation of parameters in the second stage also suffers.

The second category of methods are built on profile estimation. The approach was introduced by Ramsay et al. (2007), and it has been referred to as the parameter cascade method. Instead of estimating the ODE curves directly from the data, one first constructs the ODE curves as functions of the parameters in the inner step. These estimated functions are then included into the outer step which minimizes a loss function between the observed data and the esimated ODE curves. In Ramsay et al. (2007) and the follow-up papers, a penalty term is included in the inner step with the intention of balancing the goodness of fit between the observed data and the estimated ODE curves and the faithfulness of the estimated ODE curves towards the assumed system.

Recently, a variation of the parameter cascade method was investigated in Li et al. (2011). Their theoretical and numerical findings all suggest that, for the variance reduction purpose, one should remove the additional penalty term in the inner step. It turns out that the resulting simplified estimator is the most efficient one for the larger family of estimators considered by Ramsay et al. (2007). Furthermore, by considering the ODE initial values as part of the parameters to be estimated and reconstructing the optimization criterion in the inner step, the simplified parameter cascade method achieves smaller estimation standard errors and the results are much less affected by tuning parameters within the nonparametric estimation in the inner step, such as the choice of B-spline knots.

We note that the above methods all assume the parameters θ as constants. In reality, however, the parameters θ may not always remain constants as the system evolves with time. For example, Chen and Wu (2008) noticed that the ODE parameters in the HIV/AIDS dynamics could vary with time and they applied a two-stage method to estimate the time-varying ODE parameters. In this paper, we consider a different modeling approach that allows the ODE parameters to vary with time, but at the same time retains the interpretation advantage of a parametric ODE system. Taking the Lotka-Volterra dynamic model as an example, which is widely used to study the population evolution of predator and prey in ecological sciences. When the two components of the Lotka-Volterra model are dynamically balanced with each other, the parameters of the model are constants. However, when certain unpredictable human factors or unusual natural phenomena strike, such as earthquake, forest fire or environmentally unsound logging practice, the balance of the system may be broken and the ODE parameter values will change. If the perturbation does not last long, after a certain time period, another balanced system may be re-established and the parameters would again become constants, usually at different values from before. Note that in this situation, the estimation methods by treating ODE parameters as constants will not be

suitable, while assuming time varying parameters through out the whole time domain will result in loosing the understandings of the system provided by the constant parameters.

With this setup in mind, we wish to achieve a comprise between the two. Specifically, we propose a semi-parametric method that encourages the ODE parameters to stay as constants (for interpretability) and at the same time also allows the ODE parameters to vary with time when needed (for flexibility). The proposed method extends the framework of Li et al. (2011); the new contribution comes from a penalty term that we propose to add in the outer step of the parameter cascade method. We also show that, under certain regularity conditions, the difference between the estimated ODE curves by the proposed method and the truth is bounded at a certain rate as the sample size grows.

References

[1] N. Brunel, *Parameter estimation of ODEs via nonparametric estimators*, Electron. J. Stat. **2** (2008), 1242-1267.

[2] J. Chen and H. Wu, *Efficient local estimation for time-varying coefficients in deterministic dynamic models with applications to HIV-1 dynamics*, J. Amer. Statist. Assoc. **103** (2008), 369–384.

[3] H. Liang, H. and H. Wu, *Parameter Estimation for Differential Equation Models Using a Framework of Measurement Error in Regression Models*, Journal of the American Statistical Association **103** (2008), 1570–1583.

[4] J. Ramsay, G. Hooker, D. Campbell and J. Cao, *Parameter Estimation for Differential Equations: A Generalized Smoothing Approach*, Journal of the Royal Statistical Society, Ser. B **69** (2007), 741–796.

[5] J. Varah, *A Spline Least Squares Method for Numerical Parameter Estimation in Differential Equations*, SIAM Journal on Scientific Computing **3** (1982), 131–141.

Sparse Precision Matrix Estimation via Positive Definite Constrained Minimization of ℓ_1 Penalized D-Trace Loss Penalized D-Trace Loss

Hui Zou

(joint work with Teng Zhang)

We introduce a new collection of convex loss functions for estimating precision matrices, including the likelihood function of Gaussian graphical model as a special case. Another interesting special case gives rise to a simple loss function called the D-Trace loss which is expressed as the difference of two trace operators. We then introduce a new sparse precision matrix estimator defined as the minimizer of the ℓ_1 penalized D-Trace loss under a positive definite constraint. We develop a very efficient algorithm based on alternating direction methods for computing the positive definite constrainedℓ_1 penalized D-Trace loss estimator. Under a new irrepresentable condition our estimator has the sparse recovery property. Our irrepresentable condition is different from the irrepresentable condition for the ℓ_1 penalized MLE. An example is given to show that our irrepresentable condition can hold while the irrepresentable condition for the ℓ_1 penalized MLE fails. We establish rates of convergence of our estimator under the element-wise

maximum norm, Frobenius norm and operator norm for distributions with sub-Gaussian and polynomial tails. Simulated and real data are used to demonstrate the computational efficiency of our algorithm and the finite sample performance of our estimator. It is shown that our estimator compares favorably with the ℓ_1 penalized MLE.

Reporter: Axel Munk

Participants

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

Prof. Dr. James O. Berger
Institute of Statistics and
Decision Sciences
Duke University
112B Old Chemistry Building
Durham NC 27708-0251
USA

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

Prof. Dr. Peter Bühlmann
Seminar für Statistik
ETH Zürich
HG G 17
Rämistr. 101
CH-8092 Zürich

Prof. Dr. T. Tony Cai
Department of Statistics
The Wharton School
University of Pennsylvania
3730 Walnut Street
Philadelphia , PA 19104-6340
USA

Diego Colombo
Seminar für Statistik
ETH Zürich
HG G 17
Rämistr. 101
CH-8092 Zürich

Prof. Dr. Rainer Dahlhaus
Institut für Angewandte Mathematik
Universität Heidelberg
Im Neuenheimer Feld 294
69120 Heidelberg

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

Prof. Dr. Lutz Dümbgen
Institut für mathematische Statistik &
Versicherungslehre
Universität Bern
Alpeneggstr. 22
CH-3012 Bern

Prof. Dr. Noureddine El Karoui
Department of Statistics
University of California, Berkeley
367 Evans Hall
Berkeley CA 94720-3860
USA

Prof. Dr. Jianqing Fan
Department of Operations Research
and Financial Engineering
Princeton University
Princeton NJ 08544
USA

Dr. Klaus Frick
Institut f. Mathematische Stochastik
Georg-August-Universität Göttingen
Goldschmidtstr. 7
37077 Göttingen

Prof. Dr. Sara van de Geer
Seminar für Statistik
ETH Zürich
HG G 17
Rämistr. 101
CH-8092 Zürich

PD Dr. Markus Haltmeier
Max Planck Institute for
Biophysical Chemistry
Am Faßberg 11
37077 Göttingen

Alain Hauser
ETH Zürich
Seminar für Statistik
Rämistraße 101
CH-8049 Zürich

Rebecca von der Heide
Institut f. Mathematische Stochastik
Georg-August-Universität Göttingen
Goldschmidtstr. 7
37077 Göttingen

Prof. Dr. Chris Holmes
Department of Statistics
University of Oxford
1 South Parks Road
GB-Oxford OX1 3TG

Prof. Dr. Mark Low
University of Pennsylvania
Department of Statistics
The Wharton School
Philadelphia PA 19104-6302
USA

Dr. Zongming Ma
Department of Statistics
The Wharton School
University of Pennsylvania
3730 Walnut Street
Philadelphia , PA 19104-6340
USA

Prof. Dr. Marloes Maathuis
ETH Zürich
Seminar für Statistik
Rämistrasse 101, HG G 24.2
CH-8092 Zürich

Prof. Dr. Enno Mammen
Abteilung f. Volkswirtschaftslehre
Universität Mannheim
L 7, 3-5
68131 Mannheim

Prof. Dr. Nicolai Meinshausen
Department of Statistics
Oxford University
University Offices
Wellington Square
GB-Oxford OX1 2JD

Prof. Dr. Hans-Georg Müller
Department of Statistics
University of California
469 Kerr Hall
Davis , CA 95616-8705
USA

Prof. Dr. Axel Munk
Institut f. Mathematische Stochastik
Georg-August-Universität Göttingen
Goldschmidtstr. 7
37077 Göttingen

Prof. Dr. Susan A. Murphy
Department of Statistics
University of Michigan
439 West Hall
1085 South University
Ann Arbor MI 48109-1107
USA

Prof. Dr. Richard Nickl
Statistical Laboratory
Centre for Mathematical Sciences
Wilberforce Road
GB-Cambridge CB3 OWB

Prof. Dr. Wolfgang Polonik
Department of Statistics
University of California, Davis
One Shields Avenue
Davis CA 95616
USA

Garvesh Raskutti
2115 1/2 Ashby Ave.
Berkeley , CA 94705
USA

Prof. Dr. Pradeep Ravikumar
Department of Computer Science
University of Texas at Austin
Austin , TX 78712
USA

Prof. Dr. Yaacov Ritov
Department of Statistics
The Hebrew University of Jerusalem
Mount Scopus
Jerusalem 91905
ISRAEL

Prof. Dr. James M. Robins
Department of Biostatistics
Harvard School of Public Health
677 Huntington Ave.
Boston , MA 02115
USA

Prof. Dr. Angelika Rohde
Department Mathematik
Universität Hamburg
Bundesstr. 55
20146 Hamburg

Prof. Dr. Judith Rousseau
Universite Paris Dauphine
Place du Marechal DeLattre de Tassigny
F-75016 Paris

Till Sabel
Institut f. Mathematische Stochastik
Georg-August-Universität Göttingen
Goldschmidtstr. 7
37077 Göttingen

Dr. Richard Samworth
Statistical Laboratory
Centre for Mathematical Sciences
Wilberforce Road
GB-Cambridge CB3 OWB

Prof. Dr. Otmar Scherzer
Computational Science Center
Universität Wien
Nordbergstr. 15
A-1090 Wien

Dr. Johannes Schmidt-Hieber
Ecole Nationale de la Statistique
e de l'Adm. Economique
ENSAE
3, avenue Pierre-Larousse
F-92245 Malakoff

Hannes Sieling
Institut f. Mathematische Stochastik
Georg-August-Universität Göttingen
Goldschmidtstr. 7
37077 Göttingen

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

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

Prof. Dr. Aad W. van der Vaart
Mathematisch Instituut
Universiteit Leiden
Postbus 9512
NL-2300 RA Leiden

Prof. Dr. Martin Wainwright
Department of Statistics
University of California, Berkeley
367 Evans Hall
Berkeley CA 94720-3860
USA

Prof. Dr. Günther Walther
Department of Statistics
Stanford University
Sequoia Hall
Stanford , CA 94305-4065
USA

Prof. Dr. Lie Wang
Department of Mathematics
MIT
Cambridge , MA 02139
USA

Yin Xia
Department of Statistics
The Wharton School
University of Pennsylvania
3730 Walnut Street
Philadelphia , PA 19104-6340
USA

Prof. Dr. Bin Yu
Department of Statistics
University of California, Berkeley
367 Evans Hall
Berkeley CA 94720-3860
USA

Prof. Dr. Ming Yuan
Industrial and Systems Engineering
Georgia Institute of Technology
Atlanta , GA 30332-0205
USA

Prof. Dr. Cun-Hui Zhang
Department of Statistics
Rutgers University
110 Frelinghuysen Road
Piscataway , NJ 08854-8019
USA

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

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

Prof. Dr. Ji Zhu
Department of Statistics
University of Michigan
439 West Hall
1085 South University
Ann Arbor MI 48109-1107
USA

Prof. Dr. Hui Zou
School of Statistics
313 Ford Hall
224 Church Street S.E.
Minneapolis , MN 55455
USA