# MMS 2017 Luminy, December 18-22

**In honor of 60th birthday of**
**Oleg Lepski and Alexandre Tsybakov**

## Monday 18 December

**Marten Wegkamp**

Sparse Latent Factor Models with Pure Variables for Overlapping Clustering

*Joint with Xin Bing, Florentina Bunea, Yang Ning*

The problem of overlapping variable clustering, ubiquitous in data science, is that of finding overlapping sub-groups of a p-dimensional random vector X, from a sample of size n of observations on X. Typical solutions are algorithmic in nature, and little is known about the statistical guarantees of the estimated clusters, as most algorithms are not model-based. This work introduces a novel method, LOVE, based on a sparse Latent factor model, with correlated factors, and with pure variables, for OVErlapping clustering with statistical guarantees. The model is used to define the population level clusters as groups of those components of X that are associated, via a sparse allocation matrix, with the same unobservable latent factor, and multi-factor association is allowed. Clusters are respectively anchored by components of X, called pure variables that are associated with only one latent factor. We prove that the existence of pure variables is a sufficient, and almost necessary, assumption for the identifiability of the allocation matrix, in sparse latent factor models. Consequently, model-based clusters can be uniquely defined, and provide a bona fide estimation target. LOVE estimates first the set of pure variables, and the number of clusters, via a novel method that has low computational complexity of order p2. Each cluster, anchored by pure variables, is then further populated with components of X according to the sparse estimates of the allocation matrix. The latter are obtained via a new, computationally efficient, estimation method tailored to the structure of this problem. The combined procedure yields rate-optimal estimates of the allocation matrix and consistent estimators of the number of clusters.

**Vladimir Spokoiny**

Large ball probability and applications

We derive the bounds on the Kolmogorov distance between probabilities of two Gaussian elements to hit a ball in a Hilbert space. The key property of these bounds is that they are dimensional-free and depend on the nuclear (Schatten-one) norm of the difference between the covariance operators of the elements. We are also interested in the anticoncentration bound for a squared norm of a non-centered Gaussian element in a Hilbert space. All bounds are sharp and cannot be improved in general. We provide a list of motivation examples and applications in statistical inference for the derived results as well.

**Axel Munk**

Statistical inference for Wasserstein transport

The Wasserstein distance is an attractive tool for data analysis but statistical inference is hindered by the lack of distributional limits and fast computational schemes. To overcome the first obstacle, for discrete probability measures, we derive the asymptotic distribution of empirical Wasserstein distances as the optimal value of a linear program with random objective function. This facilitates statistical inference, e.g. hypotheses tests and confidence intervals for sample based Wasserstein distances. For the second task we provide a simple subsampling scheme for fast approximate computation of the Wasserstein-distance which can be applied in combination with any exact solver. Theory is illustrated by several data sets and numerical experiments.

**Natalia Bochkina**

Rates of convergence in semi-parametric problems with heterogeneous variance

I consider a semi-parametric estimation problem where independent observations have mean $f(x_i)$ and variance $f(x_i)^{2\theta}$, $i=1,..,n$. The aim is to estimate the unknown parameter $\theta$ that controls the variance when function $f$ is unknown. I show that if $f$ is smooth enough then it is possible to estimate $\theta$ at the parametric rate, and derive the rate of convergence in the alternative case. Asymptotic confidence intervals for $\theta$ are constructed, and the approach is illustrated on simulated data.

**Angelika Rohde**

Locally adaptive confidence bands

*Joint with Tim Patschkowski*

We develop honest and locally adaptive confidence bands for probability densities. They provide substantially improved confidence statements in case of inhomogeneous smoothness, and are easily implemented and visualized. The article contributes conceptual work on locally adaptive inference as a straightforward modification of the global setting imposes severe obstacles for statistical purposes. Among others, we introduce a statistical notion of local H\"older regularity and prove a correspondingly strong version of local adaptivity. We substantially relax the straightforward localization of the self-similarity condition in order not to rule out prototypical densities. The set of densities permanently excluded from the consideration is shown to be pathological in a mathematically rigorous sense. On a technical level, the crucial component for the verification of honesty is the identification of an asymptotically least favorable stationary case by means of Slepian's comparison inequality.

**Nicolas Verzelen**

Adaptive Estimation of Functionals in the Gaussian vector model

TBA

**Anatoli Juditsky**

Estimate aggregation from indirect observations

*Joint with Goldenshluger, A. (Haifa University) and A. Nemirovski (Georgia Tech)*

We consider the problem of point aggregation and adaptive estimation from indirect observations. The approach we promote relies upon near-optimal testing of convex hypotheses.

We show that in the classical problem of $\ell_2$-aggregation the proposed algorithms are near-optimal in different observation settings (e.g, indirect Gaussian observations, Poisson observation model and sampling from discrete distributions).

We also discuss the closely related problem of adaptive estimation. The construction of aggregation procedures reduces to convex optimization problems and can be implemented efficiently.

**Dominique Picard**

Clustering high dimensional data with sparsity

We begin by showing the practical example of finding homogeneous regions in a particular geographical domain (France metropolitan territory) on a climatological basis.

This example, is interesting by itself and shows at the same time the forces and weaknesses of the classification methods. It allows for instance to pose several important theoretical questions.

Among these, we propose to investigate the smoothing problem in the case of high dimensional data.

In particular, we try to give answers to the following questions.

1. For clustering high dimensional data, what is better: keeping the raw data or smoothing? What does smoothing means?
2. What conditions are relevant? in terms of sparsity of the data, in terms of separation of the clusters...
3. How to smooth? Does usual adaptation methods work as well to detect clusters?
4. Does on-line (signal by signal smoothing) performs as well as off-line smoothing (using a pre-process involving all the signals)?
5. What are the rates of convergence?

**Massimiliano Pontil**

Consistent Multitask Learning with Nonlinear Output Relations

*Joint with Carlo Ciliberto, Alessandro Rudi and Lorenzo Rosasco*

Key to multitask learning is exploiting relationships between different tasks to improve prediction performance. If the relations are linear, regularization approaches can be used successfully. However, in practice assuming the tasks to be linearly related might be restrictive, and allowing for nonlinear structures is a challenge. In this paper, we tackle this issue by casting the problem within the framework of structured prediction. Our main contribution is a novel algorithm for learning multiple tasks which are related by a system of nonlinear equations that their joint outputs need to satisfy. We show that the algorithm is consistent and can be efficiently implemented. Experimental results show the potential of the proposed method.

# Tuesday, Dec 19

**Quentin Berthet**

Link prediction with Matrix Logistic Regression

We consider the problem of link prediction, based on partial observation of a large network and on covariates associated to its vertices. The generative model is formulated as matrix logistic regression. The performance of the model is analysed in a high-dimensional regime under structural assumption. The minimax rate for the Frobenius norm risk is established and a combinatorial estimator based on the penalised maximum likelihood approach is shown to achieve it. Furthermore, it is shown that this rate cannot be attained by any algorithm computable in polynomial time, under a computational complexity assumption.

**Guillaume Lecué**

Learning from MOM's principles

*Joint with M. Lerasle*

We obtain estimation error rates for estimators obtained by aggregation of regularized median-of-means tests, following a construction of Le Cam. The results hold with exponentially large probability, under only weak moments assumptions on data. Any norm may be used for regularization. When it has some sparsity inducing power we recover sparse rates of convergence. The procedure is robust since a large part of data may be corrupted, these outliers have nothing to do with the oracle we want to reconstruct. Our general risk bound is of order

$$\max\left(\text{minimax rate in the i.i.d. setup}, \frac{\text{number of outliers}}{\text{number of observations}}\right) .$$

In particular, the number of outliers may be as large as *(number of data)* $\times$ *(minimax rate)* without affecting this rate. The other data do not have to be identically distributed but should only have equivalent $L^1$ and $L^2$ moments. For example, the minimax rate $s\log(ed/s)/N$ of recovery of a $s$-sparse vector in $\mathbb{R}^d$ is achieved with exponentially large probability by a median-of-means version of the LASSO when the noise has $q_0$ moments for some $q_0 > 2$, the entries of the design matrix should have $C_0\log(ed)$ moments and the dataset can be corrupted up to $C_1 s\log(ed/s)$ outliers.

**Iain Johnstone**

Eigenvalues and Variance Components

*Joint work with Mark Blows, Zhou Fan and Yi Sun*

Motivated by questions from quantitative genetics, we consider high dimensional versions of some common variance component models. We focus on quadratic estimators of 'genetic covariance' and study the behavior of both the bulk of the estimated eigenvalues and the largest estimated eigenvalues in some plausible asymptotic models.

**Richard Nickl**

Efficient nonparametric inference for a nonlinear inverse problems with the Schrödinger equation

The inverse problem of determining the unknown potential $f > 0$ in the partial differential equation (PDE) $\Delta u - fu = 0$ on $O$ s.t. $u = g$ on $\partial O$, where $O$ is a bounded $C^\infty$-domain in $R^d$, $\Delta$ is the Laplacian, and $g > 0$ is a given source function, is considered. The data consist of the solution u of the PDE, corrupted by additive Gaussian noise. For such non-linear statistical inverse problems, Bayesian recovery algorithms have been proposed by A. Stuart. As point estimators these algorithms are closely related to penalised least squares methods, but beyond that they are also useful for uncertainty quantification (the construction of confidence sets). No theory supporting the performance of such algorithms in non-linear problems is currently available, and we present some first results in this direction. In particular a Bernstein - von Mises theorem is proved which entails that the posterior distribution given the observations is approximated by an infinite-dimensional Gaussian measure that has a minimal covariance structure in an information-theoretic sense (characterised as the image of standard Gaussian white noise under a Schrödinger type operator). The function space in which this approximation holds true is shown to carry the finest topology permitted for such a result to be possible. As a consequence the posterior distribution provides valid and optimal frequentist confidence sets for f in the small noise limit. The proof techniques extend to some other PDE-type inverse problems of transport and parabolic type.

**Vladimir Koltchinskii**

Estimation of functionals of high-dimensional covariance

We will discuss a bias reduction method in the problem of estimation of smooth functionals of high-dimensional covariance operators that yields asymptotically efficient estimators of such functionals.

**Philippe Rigollet**

A biased random walk through Sasha Tsybakov's work

In this talk I will attempt to give an overview of Sasha's work spanning 35 years since his PhD. Sasha's work is rich of many topics, travels, collaborations, mentorships and friendships that I will try to illustrate in a few examples. A fluent knowledge of "Introduction to nonparametric estimation" (Springer 2009) is recommended but not mandatory.

**Enno Mammen**

Statistical inference in sparse high-dimensional nonparametric models

*The talk reports on joint work with Karl Gregory and Martin Wahl*

In this talk we consider a model that contains two components: a nonparametric component that is of interest and an additional high-dimensional nuisance nonparametric component. We assume that an initial estimator for the nuisance component is available and we apply this estimator to get an estimate of the nonparametric component of interest. This is done by applying the debiasing approach that recently has been used to update LASSO-estimators. In our first result we state finite-sample bounds for the difference between our estimator and an oracle estimator that makes use of the knowledge of the nuisance component. We show that in first order the estimators only differ by bias terms. Thus, in case of undersmoothing the asymptotic distribution theory for the oracle estimator carries over to the two-step estimator. This result will also be applied for an optimality theory for estimation. We show for a large class of smoothing methods that in the model with high-dimensional nuisance parameter estimators can be constructed that are asymptotically equivalent to the smoothers in the oracle model where the nuisance component is known.

Our leading example is estimation of a nonparametric component $f_1$ in a nonparametric additive model $Y=f_1(X_1) + ...+ f_q(X_q) + \varepsilon$. Here, $f_2+...+f_q$ is the high-dimensional nuisance component. We allow the number $q$ of additive components to grow to infinity and we make sparsity assumptions about the number of nonzero additive components. In a first step the summands in the additive model are estimated using a group-Lasso estimator. In a second step a desparsified modification of the estimator of $f_1$ is constructed. Our main mathematical work centers primarily on establishing properties of the group-Lasso estimator that enable us to apply the results of the general theorem.

**Rui Castro**

Are there needles in a moving haystack? Adaptive sensing for detection of dynamically evolving signals

*Joint with Ervin Tánczos*

We investigate the problem of detecting dynamically evolving signals. We model the signal as an n dimensional vector that is either zero or has s non-zero components. At each time step t\in N the non-zero components change their location independently with probability p. The statistical problem is to decide whether the signal is a zero vector or in fact it has non-zero components. This decision is based on m noisy observations of individual signal components collected at times t=1,...,am. We consider two different sensing paradigms, namely adaptive and non-adaptive sensing. For non-adaptive sensing the choice of components to measure has to be decided before the data collection process started, while for adaptive sensing one can adjust the sensing process based on observations collected earlier. We characterize the difficulty of this detection problem in both sensing paradigms in terms of the aforementioned parameters, with special interest to the speed of change of the active components. In addition we provide an adaptive sensing algorithm for this problem and contrast its performance to that of non-adaptive detection algorithms.

# Wednesday, Dec 20

**Mariana Pensky**

Dynamic Stochastic Block Model

*Joint with Teng Zhang*

We studied a Dynamic Stochastic Block Model (DSBM) under the assumptions that the connection probabilities, as functions of time, are smooth and that at most s nodes can switch their class memberships between two consecutive time points. We estimate the edge probability tensor by a kernel-type procedure and extract the group memberships of the nodes by spectral clustering. The procedure is computationally viable, adaptive to the unknown smoothness of the functional connection probabilities, to the rate s of membership switching and to the unknown number of clusters. In addition, it is accompanied by non-asymptotic guarantees for the precision of estimation and clustering.

**Ildar Ibragimov**

Estimation of functions depending on a parameter observed in Gaussian noise

**Alexandre Belloni**

**TBA**

**Alexander Rakhlin**

Online Prediction: Rademacher Averages via Burkholder's Functions

*Joint with D. Foster and K. Sridharan*

We develop a new family of algorithms for the online learning setting with regret against any data sequence bounded by the empirical Rademacher complexity of that sequence. To develop a general theory of when this type of adaptive regret bound is achievable we establish a connection to the theory of decoupling inequalities for martingales in Banach spaces. When the hypothesis class is a set of linear functions bounded in some norm, such a regret bound is achievable if and only if the norm satisfies certain decoupling inequalities for martingales. Donald Burkholder's celebrated geometric characterization of decoupling inequalities (Burkholder, 84) states that such an inequality holds if and only if there exists a special function called a Burkholder function satisfying certain restricted concavity properties. Our online learning algorithms are efficient in terms of queries to this function.

We realize our general theory by giving new efficient and adaptive algorithms for classes including l_p norms, group norms, and reproducing kernel Hilbert spaces. The empirical Rademacher complexity regret bound implies --- when used in the i.i.d. setting --- a data-dependent complexity bound for excess risk after online-to-batch conversion. To showcase the power of the empirical Rademacher complexity regret bound, we derive improved rates for a supervised learning generalization of the online learning with low rank experts task and for the online matrix prediction task.

In addition to obtaining tight data-dependent regret bounds, our algorithms enjoy improved efficiency over previous techniques based on Rademacher complexity, automatically work in the infinite horizon setting, and adapt to scale. To obtain such adaptive methods, we introduce novel machinery, and the resulting algorithms are not based on the standard tools of online convex optimization. We conclude with a number of open problems and new directions, both algorithmic and information-theoretic.

**Markus Reiss**

Adaptivity for partial least squares via early stopping

*Joint work with Gilles Blanchard, Potsdam, and Marc Hoffmann, Paris*

For linear inverse problems $Y = \mathsf{A}\mu + \xi$, it is classical to recover the unknown function $\mu$ by an iterative scheme $(\widehat{\mu}^{(m)}, m = 0, 1, \ldots)$ and to provide $\widehat{\mu}^{(\tau)}$ as a result, where $\tau$ is some stopping rule. Stopping should be decided adaptively, that is in a data-driven way independently of the true function $\mu$. For deterministic noise $\xi$ the discrepancy principle is usually applied to determine $\tau$. In the context of stochastic noise $\xi$, we study oracle adaptation (that is, compared to the best possible stopping iteration). For a stopping rule based on the residual process, oracle adaptation bounds within a certain domain are established. For Sobolev balls, the domain of adaptivity matches a corresponding lower bound. The proofs use bias and variance transfer techniques from weak prediction error to strong $L^2$-error, as well as convexity arguments and concentration bounds for the stochastic part. The performance of our stopping rule for Landweber and spectral cutoff methods is illustrated numerically.

# Thursday, Dec 21

**Alexander Goldenshluger**

Nonparametric density estimation from observations with multiplicative measurement errors

*Joint with D. Belomestny*

In this paper we study the problem of pointwise density estimation from observations with multiplicative measurement errors. We elucidate the main feature of this problem: the influence of the estimation point on the estimation accuracy. In particular, we show that, depending on whether this point is separated away from zero or not, there are two different regimes in terms of the rates of convergence of the minimax risk. In both regimes we develop kernel–type density estimators and prove upper bounds on their maximal risk over suitable nonparametric classes of densities. We show that the proposed estimators are rate–optimal by establishing matching lower bounds on the minimax risk. Finally we test our estimation procedure on simulated data.

**Cristina Butucea**

Estimation of linear functionals in inverse problems with errors in the operator

*This is joint work in progress with Jan Johannes and Martin Kroll*

We consider an inverse problem in a Gaussian sequence model where the multiplication operator is not known but only available via noisy observations. Our aim is to reconstruct a given linear functional of the solution. In our setup the optimal rate depends on two different noise levels, the noise level concerning the observation of the transformed solution and the noise level concerning the errors in the operator. We give minimax and adaptive estimation procedures in the given model using the Goldenshluger-Lepski method.

**Stanislav Minsker**

Robust modifications of U-statistics and estimation of the covariance structure of heavy-tailed distributions (*Joint with Xiaohan Wei)*

We propose and analyze a new estimator of the covariance matrix that admits strong theoretical guarantees under weak assumptions on the underlying distribution, such as existence of moments of only low order. While estimation of covariance matrices corresponding to sub-Gaussian distributions is well-understood, much less in known in the case of heavy-tailed data. As K. Balasubramanian and M. Yuan write, "data from real-world experiments oftentimes tend to be corrupted with outliers and/or exhibit heavy tails. In such cases, it is not clear that those covariance matrix estimators .. remain optimal" and ".. what are the other possible strategies to deal with heavy tailed distributions warrant further studies." We make a step towards answering this question and prove tight deviation inequalities for the proposed estimator that depend only on the parameters controlling the "intrinsic dimension" associated to the covariance matrix (as opposed to the dimension of the ambient space); in particular, our results are applicable in the case of high-dimensional observations.

**Natalia Stepanova**

On application of weighted Kolmogorov-Smirnov statistics to the problems of classification, signal detection, and estimation in sparse models

*Joint work with Tatjana Pavlenko (KTH Royal Institute of Technology), Yibo Wang (University of Alberta), and Lee Thompson (Carleton University)*

In this talk, we show how the goodness-of-fit test statistics based on sup-functionals of weighted empirical processes can be effectively applied to various problems of signal detection, classification, and estimation in sparse models.

The weight functions employed are Chibisov-O'Reilly functions and Erdös-Feller-Kolmogorov-Petrovski upper-class functions of a Brownian bridge.

The obtained results demonstrate the advantage of our approach over a common approach that utilizes regularly varying weight functions.

**Maxim Raginsky**

Compositional properties of statistical decision procedures: an information-theoretic view

From an information-theoretic viewpoint, randomized statistical decision procedures are channels (or Markov kernels) that map observations to probability distributions over actions. Any sufficiently complex statistical decision procedure is a composition of simpler procedures, and it is of both theoretical and practical interest to obtain a precise characterization of the overall procedure from local descriptions of the constituent subprocedures. In this talk, I will show how this problem can be addressed using information-theoretic methods.

**Marc Hoffmann**

Some memories and facts about the work of Oleg Lepski: beyond a ``discourse on method"

In this talk, I will attempt to review some of the results of Oleg Lepski and his co-authors that influenced the course of mathematical statistics over the last thirty years.

It is hard to do fair justice to the origins of ideas that circulate among a vast community of scientists, and instead of taking the route of an illegitimate historian, I will follow byways and give a personal account of what I know and understand of Oleg's influence and personality. In particular I will try not to talk too much about ``Lepski's method", but rather focus on others of his many contributions.

**Felix Abramovich**

Sparse logistic regression: model selection, goodness-of-fit and classification

*Joint with V. Grinshtein*

In the first part of the talk we consider model selection in high-dimensional logistic regression by penalized maximum likelihood estimation with a complexity penalty on the model size extending the existing results for Gaussian regression. We derive non-asymptotic upper bounds for the Kullback-Leibler risk of the resulting estimator and the corresponding minimax lower bounds. The results can be extended to a general GLM. We discuss also several alternative model selection procedures (Lasso, Slope) computationally feasible for high-dimensional data.

In the second part we apply the obtained results for model/feature selection to high-dimensional classification by sparse logistic regression. We derive the misclassification excess risk of the resulting

plug-in classifier and discuss its optimality among the class of sparse linear classifiers.

**Ekaterina Krymova**

On estimation of noise variance in high-dimensional linear models

We consider a problem of estimation of an unknown noise level in the high-dimensinal linear model. The noise variance estimators are constructed with the help of ordered spectral regularisations, which are applied to the maximum likelihood estimator of the unknown vector of linear coefficients. To select the parameter of regularisation, we optimise the deviation of the noise variance estimator from the estimator in the zero-signal case. We prove an oracle inequality for the resulting adaptive procedure. We apply our results to the minimax adaptive estimation of the noise variance in a non-linear regression model.

# Friday, Dec 22

**Pierre Alquier**

Concentration of tempered posteriors and of their variational approximations

*Joint with J. Ridgway*

While Bayesian methods are extremely popular in statistics and machine learning, their application to massive datasets is often challenging, when possible at all. Indeed, the classical MCMC algorithms are prohibitively slow when both the model dimension and the sample size are large. Variational Bayesian methods aim at approximating the posterior by a distribution in a tractable family. Thus, MCMC are replaced by an optimization algorithm which is orders of magnitude faster. VB methods have been applied in such computationally demanding applications as including collaborative filtering, image and video processing, NLP and text processing... However, despite very nice results in practice, the theoretical properties of these approximations are usually not known. In this paper, we propose a general approach to prove the concentration of variational approximations of fractional posteriors. We apply our theory to two examples: matrix completion, and Gaussian VB.

**Harrison Zhou**

Computational and Statistical Guarantees of EM for Gaussian Mixtures

*Joint with Yu Lu*

Clustering is a fundamental problem in statistics and machine learning. Lloyd's algorithm, proposed in 1957, is still possibly the most widely used clustering algorithm in practice due to its simplicity and empirical performance. However, there has been little theoretical investigation on the statistical and computational guarantees of Lloyd's algorithm. This paper is an attempt to bridge this gap between practice and theory. We investigate the performance of Lloyd's algorithm on clustering sub-Gaussian mixtures. Under an appropriate initialization for labels or centers, we show that Lloyd's algorithm converges to an exponentially small clustering error after an order of logn iterations, where n is the sample size. The error rate is shown to be minimax optimal. For the two-mixture case, we only require the initializer to be slightly better than random guess.

In addition, we extend the Lloyd's algorithm and its analysis to community detection and crowdsourcing, two problems that have received a lot of attention recently in statistics and machine learning. Two variants of Lloyd's algorithm are proposed respectively for community detection and crowdsourcing. On the theoretical side, we provide statistical and computational guarantees of the two algorithms, and the results improve upon some previous signal-to-noise ratio conditions in literature for both problems. Experimental results on simulated and real data sets demonstrate competitive performance of our algorithms to the state-of-the-art methods.

**Sara van de Geer**

## Sharp oracle inequalities for non-convex loss

We consider a parameter of interest $\beta^0 \in \mathbb{R}^p$ that is to be estimated from random data $X_1, \ldots, X_n$. Our focus is on the "high-dimensional" case where $p$ can be much larger than $n$.

For $b \in \mathcal{B} \subset \mathbb{R}^p$, let $\hat{R}_n(b)$ be a given risk function depending on the data. The set $\mathcal{B}$ is assumed to be convex. Consider for a tuning parameter $\lambda > 0$ the minimization problem

$$\min\left\{ \hat{R}_n(b) + \lambda \|b\|_1 : \ b \in \mathcal{B} \right\}.$$

Let $\hat{\beta} \in \mathcal{B}$ be a solution of the KKT conditions

$$\dot{\hat{R}}_n(\hat{\beta}) + \lambda \hat{z} = 0, \ \hat{z} \in \partial \|\hat{\beta}\|_1$$

or more generally, suppose $\hat{\beta}$ satisfies

$$\dot{\hat{R}}_n^T(\hat{\beta})(\hat{\beta} - b) + \lambda \|\hat{\beta}\|_1 - \lambda \|b\|_1 \leq 0 \ \forall \ b \in \mathcal{B}.$$

Let the theoretical risk be

$$R(b) := \mathbb{E}\hat{R}_n(b), \ b \in \mathcal{B}.$$

Let $\beta$ be some "oracle", possibly $\beta = \beta^0$ or some approximation thereof. We provide conditions for high probability sharp oracle inequalities of the form

$$R(\hat{\beta}) \leq R(\beta) + \text{error},$$

where the error depends on $\beta$ via the "curvature" of $R$ at $\beta$ and the "effective sparsity" of $\beta$. The results allow for cases where $\hat{R}_n$ is not convex (but $R$ is). Illustrations include sparse principal components and the estimation of an inverse Fisher information matrix in high dimensions. We also present some sharp oracle results for estimators based on non-differentiable but convex risk $\hat{R}_n$ such as the least absolute deviations estimator in a high-dimensional linear model. Finally, we briefly sketch the extension to sparsity inducing norms other than $\ell_1$.

**Yuri Golubev**

## On multi-channel signal detection

The talk deals with testing the simple hypothesis $H_0 : S = 0$ versus the compound alternative $H_1 : S \neq 0$ based on the observations

$$Y_i = S_i \times \mathbf{1}(i \in \tau) + \sigma\xi_i, \quad i = 1, 2, \ldots,$$

where $S = (S_1, \ldots)^\mathsf{T} \in \mathbb{R}^\infty$ is an unknown vector, $\xi_i$ are i.i.d. $\mathcal{N}(0,1)$, $\tau = \{\tau_1, \ldots, \tau_S\}$ is the unknown multi-index with i.i.d. components with a known distribution $\mathbf{P}\{\tau_k = i\} = \bar{\pi}_i$, $i \in \mathbb{Z}^+$. It is assumed that the entropy $H(\bar{\pi})$ of the a priory law $\bar{\pi}$ is large. We compute limit distributions of the maximum a posteriori probability (MAP) test statistics and the Bayes one as $H(\bar{\pi}) \to \infty$. These results permit to find maximal undetectable signal sets for MAP and Bayes tests and thus to explain in what sense the Bayes test over-performs the MAP one. It is shown also that i.i.d. random variables $\zeta_i$ which follow the limit law of the Bayes test statistics are characterized by the following property:

$$\sum_{i=1}^{\infty} \lambda_i \zeta_i \stackrel{\mathbf{P}}{=} H(\lambda) + \zeta_1,$$

where $\lambda_i \geq 0, \sum_{i=1}^{\infty} \lambda_i = 1$, and $H(\lambda) = -\sum_{i=1}^{\infty} \lambda_i \log(\lambda_i)$.

**Pierre Bellec**

## How to generalize bias and variance to convex regularized estimators ?

Convex estimators such as the Lasso, the matrix Lasso and the group Lasso have been studied extensively in the last two decades, demonstrating great success in both theory and practice. However, there are still simple open questions about these estimators, even in the simple linear regression model. We are particularly interested in the following open questions.

1) The bias and variance of linear estimators is easy to define and provide precise insights on the performance of linear estimators. How can bias and variance be generalized to nonlinear convex estimators?

2) The performance guarantees of these estimators require the tuning parameter to be larger than some universal threshold, but the literature is mostly silent about what happens if the tuning parameter is smaller than this universal threshold. How bad is the performance when the tuning parameter is below the universal threshold?

3) The correlations in the design can significantly deteriorate the empirical performance of these nonlinear estimators. Is it possible to quantify this deterioration explicitly? Is there a price to pay for correlations; in particular, is the performance for correlated designs always worse than that for orthogonal designs?

4) Most theoretical results on the Lasso and its variants rely on conditions on the design matrix. These conditions greatly simplify the proofs and our understanding of these estimators, but it is still unclear whether these conditions are truly necessary of whether they are an artifact of the proofs. Are these conditions actually necessary?

We will provide some general properties of norm-penalized estimators and propose a generalization of the bias and the variance for these nonlinear estimators. These generalizations of bias/variance will hopefully let us answer the above questions.