Sparse Statistics, Optimization and Machine Learning

Alexandre d'Aspremont (Princeton University),
Francis Bach (ENS/INRIA),
Martin Wainwright (U.C. Berkeley).

January 16 - January 21 2011.

The workshop was designed to bring together scientists from three disciplines: statistics, computer science and optimization. Several recent directions of research, among them sparse model selection, matrix completion, robust PCA and graphical model estimation, have successfully exploited the interplay between these three disciplines. However, the standard conferences in these fields do not provide a concentrated opportunity for interaction. Bringing together researchers from various backgrounds for a week at BIRS was an ideal way of making connections.

Overall, the workshop was a great success. Many of the talks provoked lively and engaging discussions, which continued on after the talk through the coffee breaks and meals. Several of the participants actively commented how this workshop was genuinely a working environment, in that the buzz of intellectual activity, discussions and collaboration were continuous.

1 Overview of the Fields

Some recent results in model selection and signal processing have recently received a considerable amount of publicity. All of these results all have one point in common: they emerged from a successful collaborations between statisticians, computer scientists, and applied mathematicians, and they relied on mathematical programming in their consistency proofs and implementation performance. Some of the talks in this workshop were part of an effort to push this recipe a bit further and identify broad classes of problems that share common structure. Other talks focused on the problem of deriving sharp model selection conditions, or that of transposing the classical linear regression-model selection results to more exotic settings, including the gene expression networks in discussed in the kick-off talk by Robert Nowak.

2 Recent Developments and Open Problems

Several key themes of research following the directions outlines above are listed below.

2.1 Fast algorithms for structured learning problems

Several talks discussed customized algorithms focused on solving large-scale (or very large scale as in the social networks of I. Dhillon's talk) problems with explicit, polynomial and often accurate complexity bounds and excellent numerical performance. Classical optimization results are focus on a certain number of standard form problems (linear programs, or more generally conic programs, geometric programming, black-box models with various smoothness assumptions, etc.). Most of the problems discussed during the workshop have

a very specific structure which allows significant efficiency gains when properly exploited in the algorithm. Given the popularity and the importance of some of these problems (e.g. NETFLIX challenge, ℓ_1-decoding, etc.) developing specialized algorithms for this class of problems has become a significant direction of work, which was covered by several talks during the workshop.

2.2 Learning with structured penalties

Beyond the classical ℓ_1 or trace norm penalties used for simultaneous variable selection in linear regression or matrix completion, designing structured penalties to account for some prior information on the problem is an active direction of research. These typically arise in imaging for example where relevant variables are often organized in simple connected sets. Structured penalties pose new modeling and computational challenges, and several talks presented efficient results when the structure is modeled as a group.

2.3 Consistency analysis and convergence rates

A number of authors presented results about the statistical consistency, and associated convergence rates of estimators based on solving optimization problems, both of the convex and non-convex variety. There are various theoretical questions at play here. How fast do the estimates obtained by solving mathematical programs converge to the true but unknown parameters? What conditions on the statistical model are required to ensure convergence? Are the rates obtained optimal, when compared to "oracle" procedures that have no computational limits? What are the differences between methods based on convex formulations, such as ℓ_1-relaxation, and those based on non-convex constraints, such as ℓ_q-constraints? A number of researchers described their recent work in addressing questions of this flavor.

3 Presentation Highlights

We summarize some of the workshop presentations in what follows, roughly following the main research directions outlined above.

3.1 Fast first-order algorithms for structured learning problems

Don Goldfarb discussed alternating direction augmented Lagrangian methods to minimize the sum of several functions subject to convex constraints, in the case where each function is relatively easy to minimize separately subject to the constraints. The algorithm followed both Gauss-Seidel-like and Jacobi-like iterations to compute an epsilon-optimal solution in $O(1/\epsilon)$ iterations. The talk also discussed accelerated versions that have an $O(1/\sqrt{\epsilon})$ iteration complexity. For the case where the sum only involves two functions and one of the functions to have a Lipschitz continuous gradient (which is typical in learning applications). These algorithms have a range of applications in matrix completion, robust PCA, covariance selection and linear regression problems with structured penalties. In these large-scale applications, scaling properties of the algorithm are key to practical performance. Numerical experiments were described on problems with tens of millions of variables and constraints.

Inderjit Dhillon presented an algorithm for compressed sensing (or ℓ_1-penalized linear regression) and matrix completion which use hard thresholding to reduce memory usage at each iteration. Convergence results are derived using the same *restricted isometry parameters* that control model selection performance. Numerical performance could be significantly improved by computing only an approximate gradient at each iteration. There is ongoing work to adapt this algorithm to the robust PCA problem (reconstruct a sparse + low rank matrix). Dhillon also discussed specialized matrix factorization techniques that could handle very large scale data sets coming from social networks, where the data tends to be naturally clustered.

Stephen Becker presented recent efficient algorithms solving constrained ℓ_1-minimization, as well as many variants such as the Dantzig Selector, nuclear norm minimization, SVM problems, and composite problems such as minimizing a combination of the TV norm and a weighted l1 norm. The main argument is the addition of a strongly convex perturbation to the primal objective which allows to efficiently solve the dual problem. An accelerated continuation scheme is used to eliminate the effect of the perturbation. In parallel,

Stephen Becker also developed a toolbox for implementing several types of efficient first order algorithms which is specifically designed to handle large-scale learning problems. Efficiently solving learning problems usually meant writing customized code for each specific class, and this toolbox aims to significantly reduce the implementation burden. In the long term- the goal is to integrate into high-level modeling solvers like the package CVX.

3.2 Linear inverse problems: beyond Lasso and matrix completion

Robert Nowak discussed "active learning" approaches. Most theory and methods for sparse recovery are based on non-adaptive measurements. The talk was focused on sequential measurement schemes that adaptively focus sensing using information gathered throughout the measurement process. It showed in particular that adaptive sensing can be more powerful when the measurements are contaminated with additive noise. While the standard sparse recovery setup involves inferring sparse linear functions, the talk discussed generalizations to the recovery of sparse multilinear functions. Such functions are characterized by multiplicative interactions between the input variables, with sparsity meaning that relatively few of all conceivable interactions are present (a setting akin to multidimensional spin glass models). This problem is motivated by the study of interactions between processes in complex networked systems (e.g., among genes and proteins in living cells). The results presented at the workshop extend the notion of compressed sensing from the linear sparsity model to notions of sparsity encountered in nonlinear systems. In contrast to linear sparsity models, in the multilinear case the pattern of sparsity can significantly affect sensing requirements. Here, the combinatorial dimension is used to characterize the complexity of the multivariate dependence pattern, but making a precise link between the sparsity pattern and the recovery rate remains a challenging problem.

In a similar vein, Ben Recht's talk focused on further extending the catalog of objects and structures that can be recovered from partial information. It discussed data analysis algorithms designed to decompose signals into sums of atomic signals from a simple set. These algorithms are derived in a convex optimization framework that encompasses previous methods based on ℓ_1-norm minimization and nuclear norm minimization for recovering sparse vectors and low-rank matrices. The talk discussed general recovery guarantees and implementation schemes for this suite of algorithms as well as several example classes of atoms and applications. Beyond simply casting existing recovery result in a new light, and simplifying their generalization, the results also apply to cases which were not covered by the existing literature, such as recovery on the permutation polytope.

3.3 Estimation and convergence rates

Lieven Vandenberghe described a new penalized regression algorithm for autoregressive Gaussian processes. In a Gaussian model, the topology of the dependence graph specifies the sparsity pattern of the inverse covariance matrix. Several topology selection methods based on convex optimization and ℓ_1-norm regularization have been proposed recently. The talk discussed extensions of these methods to graphical models of autoregressive Gaussian time series (AR processes) and focused on the problem of maximum likelihood estimation of autoregressive models with conditional independence constraints and convex techniques for topology selection via nonsmooth regularization. The maximum likelihood estimation problem for AR processes is nonconvex whenever the lag is larger than zero but the talk showed that when the sample estimates have a block-Toepliz (which is the case for the windowed estimate) a convex relaxation of the ML problem is tight, i.e. yields the global optimum. Beyond its statistical implications, this results perfectly illustrate how classical results from linear algebra and mathematical programming can find direct, and unexpected, applications in statistics (guiding the choice of estimator here).

Peter Bühlmann discussed the need for methods for performing causal inference. His talk highlighted the distinction between problems of regression type, versus those of causal or intervention type. He discussed the use of directed acyclic graphs for characterizing causality, and described an algorithm for estimating the DAG equivalence class. This algorithm is computationally efficient and can be shown to be consistent under high-dimensional scaling, including settings when the sample size n is much smaller than the number of variables p. He then discussed an even more challenging problem of inferring causality using a combination of observational and interventional data, and discussed the non-trivial optimization problems that arise from a likelihood-based formulation.

Tong Zhang discussed the use of various types of group-sparse norms for multivariate regression problems. He first demonstrated that the ordinary group Lasso does not achieve information-theoretically optimal ℓ_2-rates for a certain class of group sparse problems. He then discussed a different type of group regularization, which turned out to be related to work discussed earlier by G. Obozinski, for this problem. He proved that it has lower estimation error, in particular matching the information-theoretic limits for a particular type of group structure.

Nati Srebro discussed the use of Rademacher complexity and related quantities for proving bounds on the error of M-estimators, including the Lasso and also closely-related estimators for matrix regression. He showed that various results can be obtained quite directly by reducing certain empirical process quantities to the Rademacher complexity, and then upper bounding it appropriately. He illustrated this line of attack in application to both sparse linear models, as well as versions of matrix completion using both the nuclear norm and a related "max"-norm on matrix space, both designed to encourage low-rank behavior.