Sparse Statistics, Optimization and Machine Learning.
January 16-21 2011.

MEALS

*Breakfast (Buffet): 7:00–9:30 am, Sally Borden Building, Monday–Friday
*Lunch (Buffet): 11:30 am–1:30 pm, Sally Borden Building, Monday–Friday
*Dinner (Buffet): 5:30–7:30 pm, Sally Borden Building, Sunday–Thursday
Coffee Breaks: As per daily schedule, 2nd floor lounge, Corbett Hall
***Please remember to scan your meal card at the host/hostess station in the dining room for each meal.**

MEETING ROOMS

All lectures will be held in Max Bell 159 (Max Bell Building accessible by walkway on 2nd floor of Corbett Hall). LCD projector, overhead projectors and blackboards are available for presentations. Note that the meeting space designated for BIRS is the lower level of Max Bell, Rooms 155–159. Please respect that all other space has been contracted to other Banff Centre guests, including any Food and Beverage in those areas.

SCHEDULE

Sunday

16:00	Check-in begins (Front Desk - Professional Development Centre - open 24 hours)
	Lecture rooms available after 16:00 (if desired)
17:30–19:30	Buffet Dinner, Sally Borden Building
20:00	Informal gathering in 2nd floor lounge, Corbett Hall (if desired)
	Beverages and a small assortment of snacks are available on a cash honor system.

Monday

7:00–8:45	Breakfast
8:45–9:00	Introduction and Welcome by BIRS Station Manager, Max Bell 159
9:00–10:00	**Robert Nowak**, *Adaptive Sensing and Sparse Interactions.*
10:00–10:30	Coffee Break, 2nd floor lounge, Corbett Hall
10:30–11:30	**Philippe Rigollet**, *Prediction in misspecified high-dimensional generalized linear models.*
11:30–13:00	Lunch
13:00–14:00	Guided Tour of The Banff Centre; meet in the 2nd floor lounge, Corbett Hall
14:00–14:15	Group Photo; meet on the front steps of Corbett Hall
14:15–15:15	**Guillaume Obozinski**, *Structured Sparse Coding: efficient algorithms and applications.*
15:15–15:30	Coffee Break, 2nd floor lounge, Corbett Hall
15:30–16:30	**Ben Recht**, *The Convex Geometry of Linear Inverse Problems.*
16:30–17:30	**Nathan Srebro**, *Sparse and Low-Rank Prediction: The Generalization Bound Approach*
17:30–19:30	Dinner

Tuesday

7:00–9:00	Breakfast
9:00–10:00	**Peter Bühlmann**, *High-dimensional causal inference, DAGs and intervention DAGs.*
10:00–10:30	Coffee Break, 2nd floor lounge, Corbett Hall
10:30–11:30	**Marten Wegkamp**, *Minimax Adaptive and Consistent Effective Rank Estimation of Matrices in High Dimensional Multivariate Response Regression.*
11:30–13:30	Lunch
13:30–14:30	**Tong Zhang**, *Convex Relaxation Formulations for Structured Sparsity.*
14:30–15:30	Coffee Break, 2nd floor lounge, Corbett Hall
15:30–16:30	**Alexander Rakhlin**, *From Statistical Learning to Online Learning.*
16:30–17:30	**Constantine Caramanis**, *Robust PCA and Collaborative Filtering: Rejecting Outliers, Identifying Manipulators.*
17:30–19:30	Dinner

Wednesday

7:00–9:00	Breakfast
9:00–10:00	**Sara van de Geer**, *A Note on Variable Selection with Concave Penalty.*
10:00–10:30	Coffee Break, 2nd floor lounge, Corbett Hall
10:30–11:30	**Donald Goldfarb**, *Alternating Direction Augmented Lagrangian Algorithms for Convex Optimization.*
11:30–13:30	Lunch
	Free afternoon
17:30–19:30	Dinner

Thursday

7:00–9:00	Breakfast
9:00–10:00	**Richard Baraniuk**, *Go With The Flow: A New Manifold Modeling and Learning Framework for Image Ensembles*
10:00–10:30	Coffee Break, 2nd floor lounge, Corbett Hall
10:30–11:30	**Inderjit Dhillon**, "Title TBD"
11:30–13:30	Lunch
13:30–14:30	**Cun-Hui Zhang**, *Matrix estimation based on simultaneous confidence intervals.*
14:30–15:30	Coffee Break, 2nd floor lounge, Corbett Hall
15:30–16:30	**Noureddine El Karoui**, "Title TBD"
16:30–17:30	**Joel Tropp**, *User-Friendly Tail Bounds for Sums of Random Matrices*
17:30–19:30	Dinner

Friday

7:00–9:00	Breakfast
9:00–10:00	**Lieven Vandenberghe**, *Sparse topology selection in graphical models of autoregressive time series.*
10:00–10:30	Coffee Break, 2nd floor lounge, Corbett Hall
10:30–11:30	**Stephen Becker**, *A General Framework for Solving Constrained Optimization.*
11:30–13:30	Lunch

Checkout by 12 noon.

** 5-day workshops are welcome to use BIRS facilities (2nd Floor Lounge, Max Bell Meeting Rooms, Reading Room) until 3 pm on Friday, although participants are still required to checkout of the guest rooms by 12 noon. **

Sparse Statistics, Optimization and Machine Learning.
January 16-21 2011.

ABSTRACTS
(in alphabetic order by speaker surname)

Speaker: **Richard Baraniuk** (Rice U.)
Title: *Go With The Flow: A New Manifold Modeling and Learning Framework for Image Ensembles*
Abstract: Image processing in the internet age benefits immensely from massive databases of images and videos such as FlickR and Youtube. For instance, a large class of difficult computer vision and image understanding problems (including video summarization, automated image annotation, and localization) can be solved by learning information from image ensembles via crowd-sourcing. When an image ensemble is generated by varying a small number of imaging parameters or camera articulations, it is endowed with an additional geometric structure, namely that it can be modeled as a low dimensional image articulation manifold (IAM). The geometrical properties of the IAM encode the physical attributes of the scene under view; presumably these attributes can be accessed using differential geometry on the IAM.

A host of manifold-based processing and learning algorithms have been developed around this presumption, but unfortunately they suffer from one or more of three major shortcomings. First, current manifold processing methods assume that the image manifold is isometric to the parameter space; this assumption is violated by realistic images with textures. Second, locally linear methods that are based on a tangent space approximation to the manifold generally apply only in an extremely small neighborhood around each point on the manifold and thus can fail to capture important curvature properties of the IAM. Third, algebraic methods, while powerful in theory, require that the manifold possess an unrealistic algebraic (e.g., Lie group) structure.

In this talk, we leverage recent advances in the theory and practice of sparse and dense image correspondences to circumvent these shortcomings. We propose a new framework for modeling IAMs based on the notion of a transport operator that maps one image point an IAM to another. We observe that the optical flow between pairs of images on an IAM serves a natural and well-behaved transport operator. We establish that the space of optical flows is itself a low-dimensional smooth manifold, which enables new analytical tools for modeling, navigating, and processing IAMs. A key hallmark of our approach is that it applies to images with complex textures as well as articulations that are modeled by non-rigid and/or unstructured deformations. Numerous experiments involving novel-view synthesis, spatial and temporal super-resolution, geometric clustering, and manifold charting validate that our new framework offers significantly superior performance to existing methods.

Speaker: **Stephen Becker** (Cal. Tech.)
Title: *A General Framework for Solving Constrained Optimization*
Abstract: There are many efficient algorithms to solve an unconstrained problems like l1-regularized least-squares, but few efficient algorithms to solve variations that have constraints. We propose a technique that efficiently solves constrained l1-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 technique is based on adding a strongly convex perturbation to the primal objective which allows us to solve the dual problem. To eliminate the effect of the perturbation, we use a novel "accelerated continuation" scheme. In this talk, we will also present some recent strong convergence results.

Speaker: **Peter Bühlmann** (ETH Zürich)
Title: *High-dimensional causal inference, DAGs and intervention DAGs*
Abstract: Understanding cause-effect relationships between variables is of interest in many fields of science.

It is desirable to obtain causal information from observational data obtained by observing the system of interest without subjecting it to interventions (randomized experiments). When assuming no information about causal influence diagrams, the problem in its full generality is ill-posed. We will describe two directions addressing the difficulties: (i) infer reasonable lower bounds for causal effects; (ii) make use of additional intervention data which enables inference of causal effects (instead of lower bounds only).

A major statistical problem is estimation of high-dimensional directed acyclic graphs (DAG): either of the "usual" equivalence class of DAGs (CPDAG) based on observational i.i.d. data, or of the intervention DAG based on intervention data (or a mix of observational and intervention data) where each observation has a different distribution. We will present statistical theory and algorithmic results, illustrate real applications and discuss open problems.

Speaker: **Constantine Caramanis** (U. Texas at Austin)
Title: *Robust PCA and Collaborative Filtering: Rejecting Outliers, Identifying Manipulators*
Abstract: Principal Component Analysis is one of the most widely used techniques for dimensionality reduction. Nevertheless, it is plagued by sensitivity to outliers; finding robust analogs, particularly for high-dimensional data, is critical. We present an efficient convex optimization-based algorithm we call Outlier Pursuit that recovers the exact optimal low-dimensional subspace, and identifies the corrupted points. We extend this to the partially observed setting, significantly extending matrix completion results to the setting of corrupted rows or columns.

An interesting feature of the analysis is that unlike problems that have recently been looked at, such as decomposing the sum of a low-rank and a sparse matrix, the natural convex optimization algorithms fail to recover the exact low-rank and column-sparse matrices. Fortunately, we do not require exact recovery: discovering the identity of the outliers, and the column-space of the low-rank matrix, is enough. The consequence is that the standard proof "roadmap" for this sort of problem breaks down, requiring a new ingredient.

Speaker: **Sara van de Geer** (ETH Zürich)
Title: *A Note on Variable Selection with Concave Penalty*
Abstract: For the high-dimensional linear model $\mathbf{Y} = \mathbf{X}\beta + \epsilon$, $\mathbf{Y} \in \mathbb{R}^n$, $\beta \in \mathbb{R}^p$, $p \gg n$, we consider the penalized least squares estimator, with ℓ_r-penalty

$$\text{pen}(\beta) = \lambda^{2-r}\|\beta\|_r^r,$$

where $\lambda > 0$ is a regularization parameter, $0 \le r \le 1$ is fixed and $\|\beta\|_r^r = \sum_{j=1}^p |\beta_j|^r$. When $r = 1$, this is the Lasso, and for $r = 0$ the estimator penalizes the number of non-zero coefficients. The case $r = 0$ is often seen as the "ideal" case, where one can prove oracle inequalities without assuming restricted eigenvalue conditions, and where the number of false positives is of the same order as the size s_* of the active set of an oracle. However, with $r = 0$ the estimator is very hard to compute. The same can be said for any $0 < r < 1$. Nevertheless, it is of theoretical interest to see to what extent the estimator with $0 < r < 1$ has similar properties as the "ideal" estimator, and whether there is actually a gain in terms detection of large coefficients or in terms of the number of false negatives.

For index sets $S \subset \{1, \ldots, p\}$ and constants $L \ge 0$, we introduce the so-called ℓ_r-compatibility condition, which requires that

$$\phi_r^2(L, S) := \min\{\beta^T\hat{\Sigma}\beta|S|^{\frac{2-r}{r}} : \ \|\beta_S\|_r = 1, \ \|\beta_{S^c}\|_r \le L\}$$

is strictly positive. Here $\hat{\Sigma} := \mathbf{X}^T\mathbf{X}/n$ is the Gram matrix, and $\beta_{j,S} = \beta_j \mathrm{l}\{j \in S\}$. We show that the ℓ_r-penalized least squares estimator mimics an oracle β^* which trades off approximation error using only the coefficients in S_* and an estimation error order $\lambda^2 s_*^2/\phi^{\frac{2r}{2-r}}(3, S_*)$. Moreover, for $0 < r \le 1$ the ℓ_r-error $\|\hat{\beta} - \beta^*\|_r$ is of order $\lambda s_*^{1/r}/\phi_r^{\frac{2}{2-r}}(3, S_*)$, implying that the method will detect all coefficients $|\beta^*|$ sufficiently larger than $\lambda^r s_*/\phi_r^{\frac{2r}{2-r}}(3, S_*)$. Finally, for $0 \le r < 1$, the number of false positives is no more than

$$\left[\frac{\Lambda(s_*)}{\phi_r(3, S_*)}\right]^{\frac{r}{1-r}} \mathcal{O}(s_*) \wedge \left[\frac{1}{\phi_r(3, S_*)}\right]^{\frac{r}{1-r}} \mathcal{O}(s_*^{1+\frac{r}{2(1-r)}}).$$

Here $\Lambda(s_*)$ is a so-called maximal sparse eigenvalue. This result shows that under sparse eigenvalue conditions, the number of false positives of the ℓ_r-penalized estimator with $r < 1$ is comparable to the case $r = 0$. Without sparse eigenvalue conditions, one sees the gain as $r \downarrow 0$. The results are presented on the set

$$\mathcal{T} := \left\{ \max_{1 \le j \le p} |\mathbf{X}_j^T \epsilon|/n \le \lambda/4 \right\}.$$

We will also show that when $r = 1$ the (usual Lasso) estimator can select as many as $p - s_*$ false positives. We finally compare the results with those for the adaptive Lasso.

Speaker: **Don Goldfarb** (Columbia U.)
Title: *Alternating Direction Augmented Lagrangian Algorithms for Convex Optimization*
Abstract: Alternating direction augmented Lagrangian methods can facilitate the minimization of a convex function that is the sum of several functions subject to convex constraints, where each function is relatively easy to minimize separately subject to the constraints. In this talk, we present both Gauss-Seidel-like and Jacobi-like algorithms that compute an epsilon-optimal solution in $O(1/\epsilon)$ iterations. Nesterov-like accelerated versions that have an $O(1/\sqrt(\epsilon))$ iteration complexity are also given. For the case where the sum only involves two functions, our complexity results only require one of the functions to have a Lipschitz continuous gradient.

We present extensive numerical results on a varied set of problem classes, including matrix completion, robust principal component analysis (PCA), sparse PCA, sparse inverse covariance for graphical model selection and overlapping group LASSO problems. Some of the problems solved have tens of millions of variables and constraints.

Speaker: **Robert Nowak** (U. of Wisconsin, Madison)
Title: *Adaptive Sensing and Sparse Interactions*
Abstract: Tremendous progress has been made in high-dimensional inference problems by exploiting intrinsic low-dimensional structure. Sparsity is perhaps the simplest model for low-dimensional structure. It is based on the assumption that the object of interest can be represented as a combination of a small number of elementary components. The specic components needed in the representation are assumed to belong to a large collection or dictionary, but are otherwise unknown. Sparse recovery is the problem of determining which components are needed in the representation based on measurements of the object. I will discuss two issues related to this line of research.

1. Most theory and methods for sparse recovery are based on non-adaptive measurements. I will discuss the advantages of sequential measurement schemes that adaptively focus sensing using information gathered throughout the measurement process. In particular, I will show that adaptive sensing can be more powerful when the measurements are contaminated with additive noise.

2. The standard sparse recovery problem involves inferring sparse linear functions. I will discuss generalizations of the standard problem 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. This problem is motivated by the study of interactions between processes in complex networked systems (e.g., among genes and proteins in living cells). Our results 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.

Speaker: **Guillaume Obozinski** (INRIA - LIENS, Paris)
Title: *Structured Sparse Coding: efficient algorithms and applications.*
Abstract: One of the most successful applications of sparse methods is it use in the context of dictionary learning for image denoising and inpainting. In this talk, I will show how this application and others can benefit from regularization with structured sparsity inducing norms. To solve the corresponding

formulations, I will then present efficient and scalable algorithms for structured sparse coding based on proximal methods.

Speaker: **Alexander Rakhlin** (U. of Pennsylvania)
Title: *From Statistical Learning to Online Learning*
Abstract: Statistical Learning Theory studies the problem of estimating (learning) an unknown function given a class of hypotheses and an i.i.d. sample of data. Classical results show that combinatorial parameters (such as Vapnik-Chervonenkis and scale-sensitive dimensions) and complexity measures (such as covering numbers, Rademacher averages) govern learnability and rates of convergence. In contrast to the i.i.d. case, in the online learning framework the learner is faced with a sequence of data appearing at discrete time intervals, where data are chosen by the adversary. Unlike statistical learning, where the focus has been on complexity measures, the online learning research has been predominantly algorithm-based. That is, an algorithm with a non-trivial guarantee provides a certificate of learnability. Many known online algorithms are optimization-based methods, such as Mirror Descent or Follow the Regularized Leader.

We develop tools for analyzing learnability in the online setting even for non-convex problems. We define complexity measures which capture the difficulty of learning in a sequential manner. Among these measures are analogues of Rademacher complexity, covering numbers and fat shattering dimension from statistical learning theory. These can be seen as temporal generalizations of the classical results. A further generalization replaces the notion of regret with a general function which the algorithm and the adversary are trying to sequentially minimize/maximize. Our results cover a vast array of known frameworks, such as internal and Phi-regret, Blackwell's approachability, calibration of forecasters, global non-additive notions of cumulative loss, and more.

Speaker: **Ben Recht** (U. Wisconsin, Madison)
Title: *The Convex Geometry of Linear Inverse Problems*
Abstract: Building on the success of generalizing compressed sensing to matrix completion, this talk discusses progress on further extending the catalog of objects and structures that can be recovered from partial information. I will focus on a suite of data analysis algorithms designed to decompose signals into sums of atomic signals from a simple but not necessarily discrete set. These algorithms are derived in a convex optimization framework that encompasses previous methods based on l1-norm minimization and nuclear norm minimization for recovering sparse vectors and low-rank matrices. I will discuss general recovery guarantees and implementation schemes for this suite of algorithms and will describe several example classes of atoms and applications.

Speaker: **Philippe Rigollet** (Princeton U.)
Title: *Prediction in misspecified high-dimensional generalized linear models.*
Abstract: In her seminal paper, van de Geer (2008), investigates the prediction properties of penalized maximum likelihood estimators under rather general assumptions. The goal of this talk is to extend these results in two directions by (i) establishing sparse oracle inequalities under no assumption on the design matrix and (ii) introducing an alternative to cross-validation for the selection of the regularization parameter with provable performance.

Speaker: **Nati Srebro** (TTI Chicago)
Title: *Sparse and Low-Rank Prediction: The Generalization Bound Approach*
Abstract: In this talk I will consider sparse prediction and low-rank reconstruction problems from a generalization error analysis perspective. I will discuss the guarantees that can be obtained by combining generic generalization error bounds, bounds on the Radamacher complexity of regularized prediction classes, and bounds on the regularizer in terms of the sparsity or rank. Especially for the matrix case, I will discuss how the generalization error approach can yield more direct results, with much fewer assumptions.

Speaker: **Joel Tropp** (Cal. Tech.)
Title: *User-Friendly Tail Bounds for Sums of Random Matrices*

Abstract: We introduce a new methodology for studying the maximum eigenvalue of a sum of independent, symmetric random matrices. This approach results in a complete set of extensions to the classical tail bounds associated with the names Azuma, Bennett, Bernstein, Chernoff, Freedman, Hoeffding, and McDiarmid. Results for rectangular random matrices follow as a corollary. This research is inspired by the work of Ahlswede–Winter and Rudelson–Vershynin, but the new methods yield essential improvements over earlier results. We believe that these techniques have the potential to simplify the study a large class of random matrices.

Speaker: **Lieven Vandenberghe** (UCLA)
Title: *Sparse topology selection in graphical models of autoregressive time series*
Abstract: In a Gaussian graphical model, the topology of the graph specifies the sparsity pattern of the inverse covariance matrix. Several topology selection methods based on convex optimization and 1-norm regularization have been proposed recently. In this talk we discuss extensions of these methods to graphical models of autoregressive Gaussian time series. We discuss the problem of maximum likelihood estimation of autoregressive models with conditional independence constraints and convex techniques for topology selection via nonsmooth regularization.

Speaker: **Marten Wegkamp** (Florida State University)
Title: *Minimax Adaptive and Consistent Effective Rank Estimation of Matrices in High Dimensional Multivariate Response Regression*
Abstract: We introduce a new criterion, the Rank Selection Criterion (RSC), for selecting the optimal reduced rank estimator of the coefficient matrix in multivariate response regression models. The corresponding RSC estimator minimizes the Frobenius norm of the fit plus a regularization term proportional to the number of parameters in the reduced rank model. The rank of the RSC estimator provides a consistent estimator of the rank of the coefficient matrix; in general the rank of our estimator is a consistent estimate of the effective rank, which we define as the number of singular values of the target matrix that are appropriately large. The consistency results are valid not only in the classic asymptotic regime, when n, the number of responses, and p, the number of predictors, stay bounded, and m, the number of observations, grows, but also when either, or both, n and p grow, possibly much faster than m. We establish minimax optimal bounds on the mean squared errors of our estimators. Our finite sample performance bounds for the RSC estimator show that it achieves the optimal balance between the approximation error and the penalty term. Furthermore, our procedure has very low computational complexity, linear in the number of candidate models, making it particularly appealing for large scale problems. We contrast our estimator with the nuclear norm penalized least squares estimator (NNP), which has an inherently higher computational complexity than RSC. We show that NNP has estimation properties similar to those of RSC, albeit under stronger conditions. However, it is not as parsimonious as RSC. We offer a simple correction of the NNP estimator which leads to consistent rank estimation. We verify and illustrate our theoretical findings via an extensive simulation study.

Speaker: **Cun-Hui Zhang** (Rutgers University)
Title: *Matrix estimation based on simultaneous confidence intervals*
Abstract: We derive methods of learning population covariance and correlation matrices and their inverses based on simultaneous confidence intervals. The performance of these methods will be compared with the Lasso in selection consistency and estimation under the spectrum norm.

Speaker: **Tong Zhang** (Rutgers University)
Title: *Convex Relaxation Formulations for Structured Sparsity*
Abstract: Recent interests in structured sparsity regularization have resulted in challenges for convex relaxation methods. Most existing approaches rely on additive compositions of simpler regularization terms that lead to either under or over penalization. In fact, as we will illustrate in the talk, even for simple structures, this approach does not perform much better (in terms of sparse recovery) than simply

ignoring the structures. To remedy the problem, we investigate a different scheme for convex relaxation that does not suffer from the above mentioned issue. Our analysis implies that it is possible to derive principled convex relaxation methods for structured sparsity problems that are provably effective.