Contributions Abstracts
UCL-Duke U. Workshop on Sensing and Analysis of High-Dimensional Data

Featured Talks

Beyond stochastic gradient descent for large-scale machine learning

Francis Bach, INRIA

Many machine learning and signal processing problems are traditionally cast as convex optimization problems. A common difficulty in solving these problems is the size of the data, where there are many observations ("large n") and each of these is large ("large p"). In this setting, online algorithms such as stochastic gradient descent which pass over the data only once, are usually preferred over batch algorithms, which require multiple passes over the data. Given n observations/iterations, the optimal convergence rates of these algorithms are $O(1/\sqrt{n})$ for general convex functions and reaches $O(1/n)$ for strongly-convex functions. In this talk, I will show how the smoothness of loss functions may be used to design novel algorithms with improved behavior, both in theory and practice: in the ideal infinite-data setting, an efficient novel Newton-based stochastic approximation algorithm leads to a convergence rate of $O(1/n)$ without strong convexity assumptions, while in the practical finite-data setting, an appropriate combination of batch and online algorithms leads to unexpected behaviors, such as a linear convergence rate for strongly convex problems, with an iteration cost similar to stochastic gradient descent. Joint work with Nicolas Le Roux, Eric Moulines and Mark Schmidt.

NuMax: A convex approach for learning near-isometric linear embeddings

Richard Baraniuk, Rice University

We propose a novel framework for the deterministic construction of linear, near-isometric embeddings of finite sets of data points. Given a set of training points $x \subset \mathbb{R}^N$, we consider the secant set $\mathcal{S}(\mathcal{X})$ that consists of all pairwise difference vectors of $\mathcal{X}$, normalized to lie on the unit sphere. We formulate an affine rank minimization problem to construct a matrix Ψ that preserves norms of all the vectors in $\mathcal{S}(\mathcal{X})$ up to a distortion parameter δ. While affine rank minimization is NP-hard, we show that this problem can be relaxed to a convex program program that can be solved using a tractable semidefinite program (SDP). To enable scalability of the proposed SDP to very-large-scale problems, we adopt a two-stage approach. First, in order to reduce compute time, we develop a novel algorithm based on the Alternating Direction Method of Multipliers (ADMM) that we call Nuclear norm minimization with Max-norm constraints (NuMax). Second, we develop a greedy, approximate version of NuMax based on the column generation method commonly used to solve large-scale linear programs. We demonstrate that our framework is useful for a number of applications in machine learning and signal processing via a range of experiments on large-scale synthetic and real datasets.

Building an automatic statistician

Zoubin Ghahramani, University of Cambridge

We will live an era of abundant data and there is an increasing need for methods to automate data analysis and statistics. I will describe the "Automatic Statistician", a project which aims to automate the exploratory analysis and modelling of data. Our approach starts by defining a large space of related probabilistic models via a grammar over models, and then uses Bayesian marginal likelihood computations to search over this space for one or a few good models of the data. The aim is to find models which have both good predictive performance, and are somewhat interpretable. Our initial work has focused on the learning of unknown nonparametric regression functions, and on learning models of time series data, both using Gaussian processes. Once a good model has been found, the Automatic Statistician generates a natural language summary of the analysis, producing a 10-15 page report with plots and tables describing the analysis. I will discuss challenges such as: how to trade off predictive performance and interpretability, how to translate complex statistical concepts into natural language text that is understandable by a numerate non-statistician, and how to integrate model checking.

This is joint work with James Lloyd and David Duvenaud (Cambridge) and Roger Grosse and Josh Tenenbaum (MIT).

Breaking the coherence barrier – A new theory for compressed sensing

Anders Hansen, University of Cambridge

Compressed sensing is based on the three pillars: sparsity, incoherence and uniform random subsampling. In addition, the concepts of uniform recovery and the Restricted Isometry Property (RIP) have had a great impact. Intriguingly, in an overwhelming number of inverse problems where compressed sensing is used or can be used (such as MRI, X-ray tomography, Electron microscopy, Reflection seismology etc.) these pillars are absent. Moreover, easy numerical tests reveal that with the successful sampling strategies used in practice one does not observe uniform recovery nor the RIP. In particular, none of the existing theory can explain the success of compressed sensing in a vast area where it is used. In this talk we will demonstrate how real world problems are not sparse, yet asymptotically sparse, coherent, yet asymptotically incoherent, and moreover, that uniform random subsampling yields highly suboptimal results. In addition, we will present easy arguments explaining why uniform recovery and the RIP is not observed in practice. Finally, we will introduce a new theory that aligns with the actual implementation of compressed sensing that is used in applications. This theory is based on asymptotic sparsity, asymptotic incoherence and random sampling with different densities. This theory supports two intriguing phenomena observed in reality: 1. the success of compressed sensing is resolution dependent, 2. the optimal sampling strategy is signal structure dependent. The last point opens up for a whole new area of research, namely the quest for the optimal sampling strategies.

Optimal compressive imaging for Fourier data

Gitta Kutyniok, Technical University of Berlin

One fundamental problem in applied mathematics is the issue of recovery of continuous data from specific samples. Of particular importance is the case of pointwise samples of the associated Fourier transform, which are, for instance, collected in Magnetic Resonance Imaging (MRI). Strategies to reduce the number of samples required for reconstruction with a prescribed accuracy have thus a direct impact on such applications – which in the case of MRI will, for instance, shorten the time a patient is forced to lie in the scanner.

In this talk, we will present a sparse subsampling strategy of Fourier samples which can be shown to perform optimally for multivariate functions, which are typically governed by anisotropic features. For this, we will introduce a dualizable shearlet frame for reconstruction, which provides provably optimally sparse approximations of cartoon-like images, a class typically regarded as a suitable model for images. The sampling scheme will be based on compressed sensing ideas combined with a coherence-adaptive sampling density considering the coherence between the Fourier basis and the shearlet frame.

This is joint work with W.-Q Lim (TU Berlin).

Variable selection in high dimensional convex regression

John Lafferty, University of Chicago

Deep Gaussian processes

Neil Lawrence, Sheffield University

In this talk we describe how deep neural networks can be modified to produce deep Gaussian process models. The framework of deep Gaussian processes allow for unsupervised learning, transfer learning, semi-supervised learning, multi-task learning and principled handling of different data types (count data, binary data, heavy tailed noise distributions). The main challenge is to solve these models efficiently for massive data sets. That challenge is in reach through a new class of variational approximations known as variational compression. The underlying variational bounds are very similar to the objective functions for deep neural networks, giving the promise of efficient approaches to deep learning that are constructed from components with very well understood analytical properties.

TBA

Yann LeCun, Facebook and New York University

High-dimensional learning with deep network contractions

Stéphane Mallat, Ecole Normale Superieure

Learning suffers from the curse of dimensionality resulting from the huge volume of high-dimensional signal spaces. We show that deep neural networks yield remarkable classification results, by reducing the volume of the signal space, with adapted contractions. These contractions compute invariants relatively to uninformative data modifications.

Geometric invariants are obtained with deep scattering networks implemented with wavelet transforms. Supervised and unsupervised learning of invariants will be explained, with applications to image and audio classification, and to learning of energy functionals in physical systems.

Visual pattern encoding on the Poincaré sphere

Aleksandra Pizurica, Ghent University

In this talk we present a new theory of encoding visual patterns as constellations of signal points in a space spanned by psychophysical features, such as randomness, granularity, directionality, etc. The idea is to represent elementary visual patterns, like image patches or learned image atoms, as codewords in a space with well-defined and psychophysically intuitive structure. This new pattern-encoding scheme is inspired by the graphical representation of polarization states of a light wave on the Poincaré sphere, commonly used in optics and in digital optical communications. Defining similar representations for visual patterns is challenging. We illustrate some of the possible applications in visualizing the properties of learned dictionaries of image atoms and in patch-based image processing.

Conjugate gradient iterative hard thresholding for compressed sensing and matrix completion

Jared Tanner, University of Oxford

Compressed sensing and matrix completion are techniques by which simplicity in data can be exploited for more efficient data acquisition. The design and analysis of computationally efficient algorithms for these problems has been extensively studied over the last 8 years. In this talk we present a new algorithm that balances low per iteration complexity with fast asymptotic convergence. This algorithm has been shown to have faster recovery time than any other known algorithm, both for small scale problems and massively parallel GPU implementations. The new algorithm adapts the classical nonlinear conjugate gradient algorithm and shows the efficacy of a linear algebra perspective to compressed sensing and matrix completion. This work is joint with Jeffrey D. Blanchard (Grinnell College) and Ke Wei (University of Oxford).

Mondrian forests: Efficient random forests for streaming data via Bayesian nonparametrics

Yee Whye Teh, University of Oxford

Ensembles of randomized decision trees are widely used for classification and regression tasks in machine learning and statistics. They achieve competitive predictive performance and are computationally efficient to train (batch setting) and test, making them excellent candidates for real world prediction tasks. However, the most popular variants (such as Breiman's random forest and extremely randomized trees) work only in the batch setting and cannot handle streaming data easily. In this talk, I will present Mondrian Forests,

where random decision trees are generated from a Bayesian nonparametric model called a Mondrian process (Roy and Teh, 2009). Making use of the remarkable consistency properties of the Mondrian process, we develop a variant of extremely randomized trees that can be constructed in an incremental fashion efficiently, thus making their use on streaming data simple and efficient. Experiments on real world classification tasks demonstrate that Mondrian Forests achieve competitive predictive performance comparable with existing online random forests and periodically retrained batch random forests, while being more than an order of magnitude faster, thus representing a better computation vs accuracy tradeoff. Joint work with Balaji Lakshminarayanan and Daniel Roy.

Living on the edge: Phase transitions in convex programs with random data

Joel Tropp, California Institute of Techology

Recent research indicates that many convex optimization problems with random constraints exhibit a phase transition as the number of constraints increases. For example, this phenomenon emerges in the ℓ_1 minimization method for identifying a sparse vector from random linear measurements. Indeed, the ℓ_1 approach succeeds with high probability when the number of measurements exceeds a threshold that depends on the sparsity level; otherwise, it fails with high probability.

This talk summarizes a rigorous analysis that explains why phase transitions are ubiquitous in random convex optimization problems. It also describes tools for making reliable predictions about the quantitative aspects of the transition, including the location and the width of the transition region. These techniques apply to convex methods for denoising, for regularized linear inverse problems with random measurements, and to demixing problems under a random incoherence model. Joint with Dennis Amelunxen, Martin Lotz, and Michael B. McCoy.

Tracking dynamic point processes on networks

Rebecca Willett, University of Wisconsin-Madison

Cascading chains of interactions are a salient feature of many real-world social, biological, and financial networks. In social networks, social reciprocity accounts for retaliations in gang interactions, proxy wars in nation-state conflicts, or Internet memes shared via social media. Neuron spikes stimulate or inhibit spike activity in other neurons. Stock market shocks can trigger a contagion of jumps throughout a financial network. In these and other examples, we only observe individual events associated with network nodes, usually without knowledge of the underlying dynamic network structure. This talk addresses the challenge of tracking how events within such networks stimulate or influence future events. We adopt an online learning framework well-suited to streaming data, using a multivariate Hawkes model to encapsulate autoregressive features of observed events within the social network. Recent work on online learning in dynamic environments is leveraged not only to exploit the dynamics within the underlying network, but also to track that network

structure as it evolves. Regret bounds and experimental results demonstrate that the proposed method (with no prior knowledge of the network) performs nearly as well as would be possible with full knowledge of the network. This is joint work with Eric Hall.

Poster Presentations

Sparse inverse covariance estimation with hierarchical matrices

Jonas Ballani, EPFL

In statistics, a frequent task is to estimate the covariance matrix of a set of normally distributed random variables from given samples. In a high-dimensional regime, this problem becomes particularly challenging as the number of samples is typically much smaller than the number of variables. If the inverse covariance matrix happens to be sparse, an ℓ_1-based sparsity constraint often leads to a successful identification of the underlying correlations. The resulting convex optimisation problem can then efficiently be solved by Newton-like techniques with super-linear or even quadratic convergence rates (QUIC). A drawback of the associated quadratic model is the required computation of (dense) inverses of sparse matrices. We address this problem by the usage of hierarchical matrices which allow for an (approximate) data-sparse representation of large dense matrices. This explicit representation enables us to further exploit the simultaneous treatment of groups of variables in a block-wise manner and to easily ensure positive definiteness of each iterate. Numerical examples indicate an at most quadratic scaling in time in the number of variables under moderate storage consumptions even for high-dimensional problems.

On the absence of the RIP in practical CS and the RIP in levels

Alexander Bastounis, University of Cambridge

The success of compressed sensing in various fields has frequently been attributed to the Restricted Isometry Property (RIP). Significant research effort has shown that if a sensing matrix has the RIP then significant subsampling is possible with good results. However, we shall demonstrate that for a variety of practical problems (e.g. MRI scanning, Computerised Tomography, Flourescence Microscopy) the matrices involved do not exhibit the RIP. We propose an alternative which we term the 'RIP in levels'.

Efficient inference for joint models of LPF and spiking data

David Carlson, Duke University

One of the goals of neuroscience is to identify neural networks that correlate with important behaviors, environments, or genotypes. This work proposes a strategy for identifying neural networks characterized by time- and frequency-dependent connectivity patterns, using convolutional dictionary learning that links spike-train data to local field potentials (LFPs) across multiple areas of the brain. Analytical contributions are: (i) modeling dynamic relationships between LFPs and spikes; (ii) describing the relationships between spikes and LFPs, by analyzing the ability to predict LFP data from one region based on spiking information from across the brain; and (iii) development of a novel clustering methodology that allows inference of similarities in neurons from multiple regions. Results are based on data sets in which spike and LFP data are recorded simultaneously from up to 16 brain regions in a mouse.

Shrinkage mappings and their induced penalty functions

Rick Chartrand, Los Alamos National Laboratory

Many optimization problems that are designed to have sparse solutions employ the ℓ^1 or ℓ^0 penalty functions. Consequently, several algorithms for compressive sensing or sparse representations make use of soft or hard thresholding, both of which are examples of shrinkage mappings. Their usefulness comes from the fact that they are the proximal mappings of the ℓ^1 and ℓ^0 penalty functions, meaning that they provide the solution to the corresponding penalized least-squares problem.

In this work, we both generalize and reverse this process: we show that one can begin with any of a wide class of shrinkage mappings, and be guaranteed that it will be the proximal mapping of a penalty function with several desirable properties. Such a shrinkage-mapping/penalty-function pair comes ready-made for use in efficient algorithms, in some cases with provable convergence. We give examples of such shrinkage mappings, and use them to give compressive sensing results that are well beyond the state of the art.

Deep networks with adapted Haar scattering

Xiuyuan Cheng, Ecole Normale Superieure

We introduce an unsupervised deep learning algorithm, which applies to structured and unstructured data, with a Haar transform scattering.

A scattering transform progressively reduces the space volume by applying contractive modulus operators, and builds an invariant representation. We show that for Haar wavelets, learning contraction directions can be reduced to a hierarchical pairing problem. We optimize an adapted Haar scattering through unsupervised learning of data pairing, with a polynomial complexity algorithm. Numerical results are shown for the recognition of handwritten digits from scrambled images.

Dictionary designs for compressive sensing and distributed compressive sensing

Wei Chen, University of Cambridge

For compressive sensing (CS) applications, we propose a new method for the joint design of both the projections and the sparsifying dictionary in order to improve signal reconstruction performance. By capitalizing on the optimized projection matrix design in previous work [Chen,2013], which admits a closed-form expression as a function of any overcomplete dictionary, the proposed method does not need to involve directly the projection matrix. The projection matrix of our joint design can be directly derived based on the learned dictionary. Simulation results show that our joint design framework, which is constituted based on a set of training image patches, leads to an improved reconstruction performance in comparison to other recent approaches.

In distributed compressive sensing (DCS), in addition to intra-signal correlation, inter-signal correlation is also exploited in the joint signal reconstruction, which goes beyond the aim of the conventional dictionary learning framework. We propose a new dictionary

learning framework in order to improve signal reconstruction performance in DCS applications. By capitalizing on the sparse common component and innovations (SCCI) model, which captures both intra- and inter-signal correlation, the proposed method iteratively finds a dictionary design that promotes various goals: i) signal representation; ii) intra-signal correlation; and iii) inter-signal correlation. Simulation results show that our dictionary design leads to an improved DCS reconstruction performance in comparison to other designs. Joint work with Ian James Wassell and Miguel Rodrigues.

Unlocking energy neutrality in energy harvesting wireless sensor networks: An approach based on distributed compressed sensing

Wei Chen, University of Cambridge

We present the use of the emerging distributed compressed sensing (DCS) paradigm to deploy energy harvesting (EH) wireless sensor networks (WSN) with practical network lifetime and data gathering rates that are substantially higher than the state-of-the-art. The basis of our work is a centralized EH WSN architecture where the sensors convey data to a fusion center, using stylized models that capture the fact that the signals collected by different nodes can exhibit correlation and that the energy harvested by different nodes can also exhibit some degree of correlation. Via the probability of incorrect data reconstruction, we characterize the performance of a compressive sensing (CS) based data acquisition and reconstruction scheme and the proposed DCS based approach. These performance characterizations, which are performed both analytically and numerically, embody the effect of various system phenomena and parameters such as signal correlation, EH correlation, network size, and energy availability level. As an illustrative example, our results unveil that, for an EH WSN consisting of five SNs with our simple signal correlation and EH model, a target probability of incorrect reconstruction of 10^{-2}, and under the same EH capability as CS, the proposed approach allows for a ten-fold increase in data gathering capability. Joint work with Yannis Andreopoulos, Ian James Wassell and Miguel Rodrigues.

Mathematically grounded methods for analysing time series data on animal movement

Sarah Chisholm, University College London

In this poster, we introduce two new statistical methods to analyse interactions between individuals. The first establishes whether individuals or groups are more or less often in close proximity to each other than expected by chance. The second determines whether the movement patterns of two individuals or groups are cointegrated. Cointegration implies that there is a (stationary) linear relationship between the movement time series, regardless of the fact that they are individually non-stationary because, for example, they are random-walks that lack a constant mean, variance and covariance.

Over the past two decades, the study of free-ranging animal behaviour has been revolutionised by the growing availability of cheap technology capable of recording information continuously, in all weather conditions and without the need for an observer.

Global Positioning System (GPS) units are becoming smaller and more energy efficient as well as increasingly affordable and accurate. The addition of sensors capable of recording data that is useful in determining activity levels, gait, and physiology, coupled with very large amounts of on-collar data storage and wireless upload means that it is now an expectation that a field study will generate very considerable amounts of data, potentially in near real time. However, there is a considerable semantic gap between the low-level data provided by sensors and the much higher-level information in which a behaviourist is interested. This can only be bridged through a non-trivial domain-specific process of filtering, fusion, and aggregation. The result is that the pace of technological progress has by far outstripped that of analytical methods in the area; new analytical methods have the potential to open new avenues for science in this arena.

One key area of behavioural study concerns the interaction between individuals. Avoidance theory, social networking theory, and related areas such as epidemiology rely on an ability to establish whether two individuals (or groups) are more or less likely to associate with each other than the norm. Two new methods to analyse interactions between individuals are presented in this poster.

Orthogonal matching pursuit (OMP) to reconstruct optical coherence tomography (OCT) image

Yue Dong, University of Liverpool

Sparse approximation methods were recently applied in spectral-domain OCT to reconstruct cross-sectional image with a fraction of measured interferogram data points. It breaks the Shannon-Nyquist sampling theorem to improve the axial resolution and imaging depth simultaneously. In this study, we combined the greedy method, namely orthogonal matching pursuit (OMP), with a spectrum split approach to reconstruct cross-sectional image of pharmaceutical coating of tablet. It provides better imaging depth than Fourier Transform that is used in conventional SD-OCT. In contrast with the widely used $\ell 1$-optimization, the proposed method has much faster reconstruction speed and leads to over 3.7dB improvement in signal to noise ratio (SNR) in reconstruction of coated tablet.

Refined analysis of sparse MIMO radar

Dominik Dorsch, RWTH Aachen University

We consider a multiple-input-multiple-output (MIMO) radar model and prove recovery results for a Compressed Sensing (CS) approach. The MIMO model is based on random linear probes emitted by N_T transmitters and then after a linear transformation, depending on s targets in the azimuth-range(-Doppler) domain, are received by N_R receivers. Since each receiver takes N_t samples during one time interval, one obtains $N_t N_R$ measurements overall. We show that the Restricted Isometry Property (RIP) is fulfilled provided that the number of samples N_t basically grows linearly in the number of targets s (up to some additional log factors). This result comes somehow surprisingly, considering that only the fraction N_t of the total number of measurements $N_t N_R$ appears

in the condition. However, using general results from CS we show that our RIP result cannot be improved substantially. Our arguments reveal how the fine structure of the support set comes into play. Indeed, we introduce a deterministic model for the support sets which incorporates a measure on how equally distributed over the azmiuth angles the support set is. This way certain "bad" support sets which concentrate on "similar" angles are avoided. As a result we are able to show nonuniform recovery results revealing linear dependence of the sparsity in the total number of measurements. We compare our results to a recent analysis by Strohmer & Friedlander who assumed random support sets.

Recovery of wavelet expansion from nonuniform Fourier samples via weighted iterative hard thresholding

Jonathan Fell, RWTH Aachen University

A recent approach for function reconstruction from samples attempts to exploit both smoothness of a function as well as sparsity of its representation coefficients in some basis.

This leads to the notion of weighted sparsity and recovery algorithms such as weighted ℓ_1-minimization or weighted iterative hard thresholding (IHT). We analyze an implementable version of the weighted IHT. As an application we consider the recovery of wavelet expansions from nonuniformly distributed samples of its continuous Fourier transform.

Requiring well-approximability by a weighted sparse wavelet expansion is equivalent to the function being contained in a certain Besov space when the weight is chosen appropriately so that this model is well-suited for image processing and image recovery. In fact, taking (nonuniformly distributed) samples of the continuous Fourier transform corresponds to certain sampling schemes in computerized tomography and in magnetic resonance imaging.

Sparsistent additive modeling in multi-task learning

Madalina Fiterau, Carnegie Mellon University
Mladen Kolar, University of Chicago Booth School of Business

Learning related functions jointly in the multi-task linear regression setting has been shown to outperform learning each individual function separately. Although multi-task learning performance guarantees under various conditions are of increasing interest, existing work on this topic does not typically consider non-linear dependencies between covariates and output. We analyze the performance of nonparametric additive models under the assumption of component sparsity, when the objective is to retrieve the non-zero components given a limited number of samples. We establish results for the component selection consistency of a two stage penalized least squares estimator, comparing them to existing results on component selection in the single-task problem. We also establish convergence results for the non-zero components and the conditional mean functions. Our simulation studies and real data experiments confirm our theoretical findings. This is joint work with Mladen Kolar.

Low-complexity compressive sensing detection for spatial modulation in large-scale multiple access channels

Adrian Garcia-Rodriguez, University College London

In this work we propose a detector, based on the compressive sensing (CS) principles, for multiple access spatial modulation (SM) channels with a large-scale antenna base station (BS). Particularly, we exploit the use of a large number of antennas at the BSs as well as the structure and sparsity of the SM transmitted signals to improve the performance of conventional detection algorithms. The proposed CS-based detector allows the reduction of the signal processing load at the BSs particularly pronounced for SM in large-scale multiple-input multiple-output (MIMO) systems. We further carry out analytical performance and complexity studies of the proposed scheme to evaluate its usefulness. The theoretical and simulation results presented in this work show that the proposed strategy constitutes a low-complexity alternative to significantly improve the systems energy efficiency against conventional MIMO detection in the multiple access channel.

A multiscale approach to discrete optimal transport

Sam Gerber, Duke University

Moving resources at minimal cost, permeates numerous fields, from obvious applications in logistics to economics, geophysical models, image registration and machine learning. Despite these widespread applications, the efficient computation of optimal transport plans remains challenging. This paper proposes a multiscale framework that extends the range and efficiency for solving discrete optimal transport problems.

The discrete optimal transport problem can be solved by a specialized linear program, the minimum network flow problem. The minimum network flow problem has been extensively studied in the operations research community and several fast algorithms exist. However, these algorithms do not scale beyond a few thousand source and target points. This paper describes a framework to extend the applications of these algorithms to problems with millions of points. The framework exploits a multiscale representation of the source and target sets to reduce the problem size and quickly find good initial solutions: The optimal transport problem is solved at the coarsest scale and the solution is propagated to the next scale and refined. This process is repeated until the finest scale is reached. This strategy is adaptable to memory limitations and speed versus accuracy trade-offs and, depending on the refinement strategy, is guaranteed to converge to the optimal solution.

Multichannel adaptive filtering in compressive domains

Karim Helwani, Huawei European Research Center

In this presentation, we give a study on reducing the coefficients to be estimated in an adaptive sparse multichannel system identification problem. We present an approach to perform the adaptation in a compressed representation of the sparse system without requiring prior knowledge about the dimensions in which the system has significant

components. The presented technique exploits the ability of sparse systems to be compressed offering a reduction of the adaptive filter coefficients in addition to high convergence rates.

Modulator design for binary classification of Poisson measurements

Jiaji Huang, Duke University
Robert Calderbank, Duke University

The classification of Poisson measurements is of major interest in many optical sensing systems. Instead of recovering the input then determining its label, classification without recovery is more effective and efficient. Motivated by this idea, we propose to insert a designed modulator (coded aperture) that can increase the diversity between different classes and improve classification accuracy. In particular, this work studies a binary classification setup where each of the two classes is an exactly known signal. The optimal modulator is shown to be a binary mask, which is feasible for real-world implementation. Efficient algorithms are proposed to design the modulator and encouraging results are achieved on real data.

Analyzing the structure of multidimensional compressed sensing problems through local coherence

Alex Jones, University of Cambridge

In work by both Adcock, Anders, Poon, Roman and Krahmer, Ward it was shown that compressed sensing can be applied to cases where there is large coherence. Instead the more precise notion of local coherence was used to allow effective subsampling. A study of local coherence is provided in general and for specific cases, such as Fourier sampling and wavelet sparsity. Theoretical limits for how fast the local coherence can decay in general are provided. It is also shown that by using separable wavelets, the typical problems associated with the corresponding tensor problems can be avoided, giving the same decay as in one-dimension. The geometric structure of local coherences are also studied, demonstrating that in two-dimensions separable wavelets and Fourier sampling are compatible with subsampling strategies used in practice (such as spirals or radial lines) however there are important but avoidable pitfalls with trying to extend these results to three or more dimensions.

Robust uniform recovery of low-rank matrices from Gaussian measurements

Maryia Kabanava, RWTH Aachen University

We consider the problem of robust uniform recovery of low-rank matrices from noisy linear measurements using nuclear norm minimization. We rely on the robust null space property to show that with high probability $r(\sqrt{n_1}+\sqrt{n_2})^2$ Gaussian measurements are sufficient for the recovery of all $n_1 \times n_2$-matrices up to rank r.

Matrix completion on graphs

Vassilis Kalofolias, EPFL

The problem of finding the missing values of a matrix given a few of its entries, called matrix completion, has gathered a lot of attention in the recent years. Although the problem is NP-hard, Candes and Recht showed that it can be exactly relaxed if the matrix is low-rank and the number of observed entries is sufficiently large. In this work, we introduce a novel matrix completion model that makes use of proximity information about rows and columns by assuming they form communities. This assumption makes sense in several real-world problems like in recommender systems, where there are communities of people sharing preferences, while products form clusters that receive similar ratings. Our main goal is thus to find a low-rank solution that is structured by the proximities of rows and columns encoded by graphs. We borrow ideas from manifold learning to constrain our solution to be smooth on these graphs, in order to implicitly force row and column proximities. Our matrix recovery model is formulated as a convex non-smooth optimization problem, for which a well-posed iterative scheme is provided. We study and evaluate the proposed matrix completion on synthetic and real data, showing that the proposed structured low-rank recovery model outperforms the standard matrix completion model in many situations.

Tensor low-rank and sparse light field photography

Mahdad Hosseini Kamal, EPFL

High-quality light field photography has been one of the most difficult challenges in computational photography. Conventional methods either sacrifice resolution, use multiple devices, or require multiple images to be captured. Combining coded image acquisition and compressive reconstruction is one of the most promising directions to overcome limitations of conventional light field cameras. We present a new approach to compressive light field photography that exploits a joint low-tensor-rank and sparse prior on natural light fields with motion. This prior facilitates robust high-resolution light field video capture. We also propose a new camera design that captures light fields with a wider baseline than previously possible and demonstrate that it is well-suited for the proposed compressive reconstruction schemes.

Coherence and sufficient sampling densities for reconstruction in compressed sensing

Franz Kiraly, University College London

We give a new, very general, formulation of the compressed sensing problem in terms of coordinate projections of an analytic variety, and derive sufficient sampling rates for signal reconstruction. Our bounds are linear in the coherence of the signal space, a geometric parameter independent of the specific signal and measurement, and logarithmic in the ambient dimension where the signal is presented. We exemplify our approach by deriving sufficient sampling densities for low-rank matrix completion and distance matrix completion which are independent of the true matrix. (http://arxiv.org/abs/1302.2767)

Learning with cross-kernels and ideal PCA

Franz Kiraly, University College London

We describe how cross-kernel matrices, that is, kernel matrices between the data and a custom chosen set of 'feature spanning points' can be used for learning. The main potential of cross-kernels lies in the fact that (a) only one side of the matrix scales with the number of data points, and (b) cross-kernels, as opposed to the usual kernel matrices, can be used to certify for the data manifold. Our theoretical framework, which is based on a duality involving the feature space and vanishing ideals, indicates that cross-kernels have the potential to be used for any kind of kernel learning. We present a novel algorithm, Ideal PCA (IPCA), which cross-kernelizes PCA. We demonstrate on real and synthetic data that IPCA allows to (a) obtain PCA-like features faster and (b) to extract novel and empirically validated features certifying for the data manifold. (http://arxiv.org/abs/1406.2646)

Modeling correlated arrival events with latent semi-Markov processes

Wenzhao Lian, Duke University

The analysis of correlated point processes data has wide applications ranging from biomedical research to network analysis. We model such data as generated by a latent collection of continuous-time binary semi-Markov processes, corresponding to external events appearing and disappearing. A continuous-time modelling framework is more appropriate for multichannel point process data than a binning approach requiring discretizing, and we show connections between our model and many recent ideas from the discrete-time domain. We describe an efficient MCMC algorithm for posterior inference, and apply our ideas to both synthetic data and a real-world biometrics application.

MUSIC for single-snapshot spectral estimation: Stability and super-resolution

Wenjing Liao, Duke University
Albert Fannjiang, University of California, Davis

We study the problem of line spectral estimation in the continuum of a bounded interval with one snapshot of array measurement. The single-snapshot measurement data is turned into a Hankel data matrix which admits the Vandermonde decomposition and is suitable for the MUSIC algorithm. In the noise-free case exact reconstruction is guaranteed for any arbitrary set of frequencies as long as the number of measurement data is at least twice the number of distinct frequencies to be recovered. In the presence of noise we provide a stability analysis of the MUSIC algorithm while the frequencies are separated by at least twice the Rayleigh length by means of novel discrete Ingham inequalities. Moreover our numerical simulation shows that MUSIC has super-resolution effect - the capability of resolving closely spaced frequencies. In order to understand its super-resolution effect, we systematically study the performance of MUSIC for the reconstruction of certain number of equally spaced frequencies whose separation is below one Rayleigh length and provide quantitative explanations for the results.

Terahertz imaging via block based compressive sensing

Lin Liu, University of Liverpool

The terahertz (THz) region of the electromagnetic spectrum spans the frequency range between the millimetre/microwave and the mid-infrared (300 GHz 30 THz). THz technology has advantages of being non-ionizing, non-destructive, and able to image at depth. A number of imaging applications have been reported in areas such as medical diagnosis of human tissue, detection and chemical mapping of illicit drugs and explosives, analysis of polymer samples and pharmaceutical tablets, and integrated circuit package inspection [1]. However, most existing THz imaging systems remain too slow for time-critic applications due to their pixel-by-pixel raster scans.

Recently Chan et al. [2] reported a fast terahertz imaging system that uses a single pixel detector in combination with a series of random masks to enable high-speed image acquisition. Using this compressed sensing concept [3] we have also demonstrated that both the spatial distribution and the spectral characteristics of a sample could be obtained by compressed terahertz pulsed imaging. Compared with conventional terahertz pulsed imaging, no raster scanning of the object is required, and ten times fewer THz spectra need be taken, making it attractive for real time terahertz imaging applications [4]. Despite these promising results, a single detector could only produce a low-resolution image (up to 96×96 in our experiment). If the resolution is high, i.e. 256×256 pixels, the process of reconstruction can be complex and slow, as the reconstruction process requires a good computation capability.

Here we report an experimental implementation of block based compressed sensing method [5] aiming to achieve high-resolution terahertz imaging at high-speed. We will present results obtained by using a spinning disk as the spatial light modulator [6], and a 2×2 sensor array as the sensing element. Each detector works independently and in parallel (each detector senses a block of the original image) thus both the image acquisition and image reconstruction can be speeded up significantly.

Sparse recovery conditions and realistic forward modeling in EEG/MEG source reconstruction

Felix Lucka, University of Munster

Measuring the induced electromagnetic fields at the head surface to estimate the underlying, activity-related ion currents in the brain is a challenging, severely ill-posed inverse problem. Especially the recovery of brain networks involving deep-lying sources by means of EEG/MEG recordings is still a challenging task for any inverse method. Using different types of spatial sparsity constraints has become increasingly popular to address challenges like correct source localization and separation. We are especially interested in the interplay of realistic forward and sparse inverse modeling. Our focus is on how the intrinsic recovery properties of the system matrix evolve with modeling complexity, not on modeling errors. In this preliminary work, we investigate which mathematical concepts are suitable for such examinations; in particular we examine different recovery conditions from compressed sensing.

Fast and robust multiscale dictionary learning

Mauro Maggioni, Duke University

The dictionary learning problem is that of constructing, given a training set of signals, a set of vectors (dictionary) such that the signals admit a sparse representation in terms of the dictionary vectors. This has a variety of applications in statistical signal processing and modeling of high-dimensional data, as well as connections to compressed sensing, where a dictionary sparsifying the signals of interest is assumed given. We discuss a multiscale geometric construction for learning such dictionaries, its computational cost and online versions, and finite sample guarantees on its quality. It performs extremely well when the data/signals of interests is concentrated near a low-dimensional manifold in high-dimensions. We present applications to constructing probabilistic models for the data and perform anomaly detection.

Distributed compressed sensing algorithms: Completing the Puzzle

João Mota, University College London

Reconstructing signals in compressed sensing involves solving an optimization problem. An example is basis pursuit, which works only in noise-free environments. In noisy environments, it is common to solve basis pursuit denoising, noise-aware basis pursuit, and lasso. We solve all these problems in a distributed environment. The setting is a network with an arbitrary topology, where each node has access only to a portion of the sensing matrix. We address two scenarios: column partition, where the sensing matrix is partitioned by columns; and row partition, where the sensing matrix is partitioned by rows. Prior to our work, solving basis pursuit denoising with a column partition or noise-aware basis pursuit with a row partition was an open problem. Our approach consists of manipulating each of these problems so that a recent general-purpose algorithm for distributed optimization can be applied.

A unified algorithmic approach to distributed optimization

João Mota, University College London

We solve generic optimization problems formulated on networks. Each node in the network has exclusive access to a function, and the goal is to find a vector that minimizes the sum of all the functions. We assume that each function, rather than depending on all the components of the variable, may depend on a subset of them. This creates an additional structure in the problem that is captured by the classification scheme we propose. This scheme enables us both to design an algorithm that solves very general distributed optimization problems, but also to categorize prior algorithms and applications. Despite the generality of our algorithm, it shows an excellent performance, requiring less communications to converge than prior algorithms, even algorithms that are application-specific.

Learning from negative examples for machine translation

Tsuyoshi Okita, Dublin City University

It is common to take advantage of negative examples in a classification problem. This is not the case for complex problems such as Statistical Machine Translation (which is a sequential prediction with reordering). It is advantageous if the Machine Learning model for SMT has capability to incorporate the effect of negative examples (or error patterns), but this is difficult: (1) the cause of negative mistake scatters around different layers in a high dimensional space (hence, it is not straight forward to correct errors without locating the cause of errors), and (2) we have considerably few but important negative examples whose dimensions are insufficient and biased. We start with transformation-based learning [Brill, 95], cross entropy-based language models for domain adaptation [Moore and Lewis, 10], and a deep learning approach [Hinton et al., 06]. This is the joint work with Qun Liu.

Finite dimensional FRI for reconstruction of sparse signals

Jon Onativia, Imperial College London
Pier Luigi Dragotti, Imperial College London

Traditional Finite Rate of Innovation (FRI) theory has considered the problem of sampling continuous-time signals. This framework can be naturally extended to the case where the input is a discrete-time signal. Here we present a novel approach which uses both the traditional FRI sampling scheme, based on the annihilating filter method, and the fact that in this new setup the null space of the problem to be solved is finite dimensional. This method can be applied to the reconstruction of sparse signals for which we have partial knowledge of their Fourier transform and can also be extended to more general cases where the acquisition matrix is structured.

In the noiseless scenario, we show that this new approach is able to perfectly recover the original signal at the critical sampling rate. We also present simulation results in the noisy scenario where this new approach improves performances in terms of the mean squared error (MSE) of the reconstructed signal when compared to the canonical FRI algorithms and compressed sensing (CS).

Supervised learning on an unsupervised atlas

Nikolaos Pitelis, University College London

In many machine learning problems, high-dimensional datasets often lie on or near manifolds of locally low-rank. This knowledge can be exploited to avoid the "curse of dimensionality" when learning a classifier. Explicit manifold learning formulations such as LLE are rarely used for this purpose, and instead classifiers may make use of methods such as local co-ordinate coding or auto-encoders to implicitly characterise the manifold. We propose novel manifold-based kernels for semi-supervised and supervised learning. We show how smooth classifiers can be learnt from existing descriptions of manifolds that characterise the manifold as a set of piecewise affine charts, or an atlas. We experimentally validate the importance of this smoothness vs. the more natural piecewise smooth classifiers, and we show a significant improvement over competing

methods on standard datasets. In the semi-supervised learning setting our experiments show how using unlabelled data to learn the detailed shape of the underlying manifold substantially improves the accuracy of a classifier trained on limited labelled data.

Compressive classification of a mixture of Gaussians: Analysis, designs and applications

Hugo Reboredo, University of Porto–Instituto de Telecomunicações

This work derives fundamental limits on the performance of compressive classification when the source is a mixture of Gaussians, providing an asymptotic analysis of a Bhattacharya based upper bound on the misclassification probability for the optimal Maximum-A-Posteriori (MAP) classifier, which depends on quantities that are dual to the concepts of diversity-order and coding gain in multi-antenna communications. The diversity-order of the measurement system determines the rate at which the probability of misclassification decays with signal-to-noise ratio (SNR) in the low-noise regime. Instead, the measurement gain determines the power offset of the probability of misclassification in the low-noise regime. These two quantities make it possible to quantify differences in misclassification probability between random measurement and (diversity-order) optimized measurement. The behavior of misclassification probability is revealed to be intimately related to certain fundamental geometric quantities determined by the measurement system, the source and their interplay, allowing us to derive projection designs which effectively extract the most discriminating features from the sources. Numerical results demonstrate alignment of the actual misclassification probability with the Bhattacharya based upper bound. Our model is then applied to a face recognition application, in which we aim at classifying face images of various individuals taken under different illumination conditions. In this way, we showcase how the mathematical analysis can provide insights into the mechanisms that regulate the real data experiments.

Reconstruction of high-dimensional GMM data from low-dimensional features

Francesco Renna, University of Porto-Instituto de Telecomunicações

We determine fundamental limits on the reconstruction error incurred in recovering high-dimensional data from low-dimensional features. In particular, we consider the case in which high-dimensional data can be effectively described by a mixture of Gaussian distributions and the feature extraction process is impaired by additive Gaussian noise. The analysis provides a sharp characterization of the minimum number of features needed for reliable reconstruction, considering both random and designed linear feature extraction, which is based on the geometrical description of the Gaussian mixture. Our framework can be applied to the case of image acquisition and reconstruction, where small patches extracted from natural images are shown to be described accurately by the mixture of a restricted number of Gaussians, thus allowing low-complexity, reliable reconstruction. Real data experiments show how high-quality reconstruction can thus be obtained by optimally tuning the number of linear features extracted.

Classification of high-dimensional data from low-dimensional features in the presence of side information

Francesco Renna, University of Porto-Instituto de Telecomunicações

We consider the problem of classifying high-dimensional data from low-dimensional linear features in the presence of side information. In particular, we consider the case in which high-dimensional data can be described by mixtures of Gaussian mixture model (GMM) distributions. Side information data is also described by a mixture of GMM distributions that are correlated to those corresponding to the input data.

We characterize the Chernoff upper bound to the error probability associated to the maximum a-posteriori (MAP) classifier in terms of both phase transition and diversity order, as a function of the number of linear features extracted from the high-dimensional data, the side information, and the geometrical description of their joint distribution. We are then able to quantify the effect of random linear features extracted from side information data in lowering the misclassification probability, thus leading to guidelines on how to optimally distribute feature extraction between the input and side information data.

Order statistics of exponential random variables with imperfect measurement and unknown Gaussian disturbance for resource allocation compression models

Ramiro Samano Robles, Instituto de Telecomunicações/Research Centre of Real Time and Embedded Computer Systems

This paper deals with a scenario where a set of exponentially-distributed random variables are sorted based on measured information that is inaccurate. This is a typical scenario found in channel-gain scheduling problems in wireless networks with imperfect channel state information at the transmitter side (CSIT), where channel strength is usually modeled as an exponentially distributed random variable. Imperfect CSIT is the result of channel estimation errors and feedback delay. Imperfect CSIT results in potential errors in resource allocation, which is typically based on assigning modulation formats according to thresholds of the measured channel strength. Once radio resources have been allocated, transmissions are subject to additional errors due to potentially unknown co-channel interference, which is modeled as a Gaussian process. This setting corresponds to a future wireless network with cognitive radio, where a set of potentially unknown number of secondary opportunistic transmissions (unlicensed) can potentially interfere with a primary transmission due to sensing errors. The main objective here is to derive the order statistics of the sorted random variables based on imperfect measurements, derive the statistics of potential resource allocation errors under inaccurate measurements and Gaussian interference, and ultimately define modified decision thresholds that can reduce the probability of errors. This leads to a statistical compression model of the transmission link that provides a useful abstraction/summary for the optimization of higher layer processes. This compression model allows us to reduce the complexity in the simulation of large wireless networks, while maintaining a statistically accurate evaluation

of the lower level processes: scheduling, transmission, interference and reception.

On asymptotic sparsity in compressed sensing

Bogdan Roman, University of Cambridge

Compressed sensing has been one of the highlights of the last decade in applied mathematics, engineering and compuiter science as it allows one to overcome the traditional Nyquist barrier of classical sampling theory. Assuming the signal is sparse in some basis, and that the sensing and sparsifying bases are incoherent, compressed sensing allows robust recovery from undersampled data using random sensing matrices for sampling, which provide universality.

While there are examples where these principles apply, in many practical applications one or more are lacking. This includes medical imaging – Magnetic Resonance Imaging (MRI), various forms of Tomography e.g. Computerized, Thermoacoustic, Photoacoustic or Electrical Impedance –, Electron Microscopy, Seismic Tomography, Fluorescence Microscopy, Hadamard Spectroscopy, Radio Interferometry etc. In many of these, it is the principle of incoherence that is lacking, making the standard theory inapplicable. Despite this, compressed sensing has been used successfully in many of these areas. Yet, it is typically implemented with subsampling strategies that differ greatly from the uniform subsampling ones suggested by the theory. In fact, as the poster will show, in many cases uniform random subsampling yields highly suboptimal results.

A critical aspect of many real-world problems such as those above is that they do not offer the freedom to design or choose the sensing operator, but instead impose it (e.g. Fourier in MRI). Consequently, much of the existing compressed sensing work which relies on random or custom designed sensing matrices – typically with the purpose of providing universality – is not applicable.

This poster will show that in many such applications the imposed sensing operators are both highly non-universal and coherent with popular sparsifying bases, but that they are however asymptotically incoherent. The so called flip-test that we introduced will also show that the empirical success of compressed sensing witnessed until now in applications such as above is not due to sparsity alone or the Restricted Isometry Property. Most real-world signals in such applications are not sparse, but asymptotically sparse. Based on the new theory we recently introduced, applicable to many practical problems, the poster will show how the asymptotic incoherence phenomenon coupled with multilevel random sampling can exploit asymptotic incoherence, and that it provides several advantages over universal sampling operators, even when the latter are applicable, such as compressive imaging. A consequence of the asymptotic behaviour is that the success of compressed sensing is resolution dependent – as sparsity and incoherence grow asymptotically with resolution – and that substantial gains can be obtained from the same amount of measurements taken using multilevel subsampling.

Taking sparsity structure into account in the sampling procedure, via multilevel sampling of non-universal matrices, has several advantages over approaches which take sparsity structure into account in modified recovery algorithms for universal sampling operators (e.g. message passing, model-based, Bayesian), including higher quality recovery

and ample freedom to choose the sparsifying system (e.g. TV, curvelets, shearlets, contourlets etc).

Variational Bayesian inference for sparse matrix factorization

Evangelos Roussos, University of Oxford

In the exploratory analysis of multivariate data, it is often desirable to decompose the data vectors as a superposition of different prototypical, and possibly overlapping, "patterns". For example, in the linear model of brain activation, the observed four-dimensional spatio-temporal data (indexed by voxels in a brain volume and scan time-points) are modelled as a linear superposition of different activity patterns that are pairs of regressors and corresponding coefficients, respectively capturing the time-courses and spatial brain maps of the patterns of activation. Our goal there is the decomposition of the data set into such spatio-temporal components, representing, respectively, the dynamics and the spatial variation of the data in a parsimonious manner. Unlike model-based approaches, both time-courses and spatial maps are unknown here. This can be stated as a matrix factorization problem. Modern neuroimaging datasets pose unique challenges for many existing approaches to this problem, however, as they do not include ways for handling noise and uncertainty, inherent in these large-scale datasets, or ways for seamlessly fusing information from other parts of the processing pipeline, such as spatial smoothing. We approach the problem from a Bayesian perspective and propose a fully Bayesian hierarchical model for bilinear decompositions. In particular, we take the view of matrix factorization as a generative model, and the problem of estimating the components as a statistical inference problem. Bayesian modelling also allows us to easily incorporate recent theoretical and experimental findings in imaging neuroscience revealing that activations inferred from functional MRI data have sparse structure. To further reinforce sparsity, we derive a hybrid wavelet-sparse matrix factorization model, transforming the signals into a domain where the sparsity of the coefficients with respect to an appropriately chosen dictionary is natural. We then follow a graphical modelling formalism and use sparsity-inducing priors, allowing joint reconstruction and denoising. While the advantages of Bayesian modelling for data analysis are well known, they often come at a computational cost and, often, analytical intractability. Since exact inference and learning in such a model is computationally demanding, we follow a variational Bayesian approach, for efficient unsupervised learning in multidimensional settings. Importantly, employing Bayesian modelling allows us to assess uncertainty, perform sensor noise level estimation, and infer the complexity of the representation, in order to perform model selection. Using fMRI datasets as examples of complex, high-dimensional, spatiotemporal data, we show that our method can better recover the latent generators of the data compared to other state-of-the-art model-free tools, while potentially allowing for more interpretable activation patterns due to sparsity and automatic denoising.

Sparse estimation with generalized Beta mixture and the Horseshoe prior

Zahra Sabetsarvestani, Amirkabir University of Technology

Developing a sparsity-inducing prior for modeling the sparse signal efficiently is a main research topic in Bayesian Compressive Sensing (BCS). We propose the use of Generalized Beta Mixture (GBM) and Horseshoe distributions as priors in the BCS framework. These priors feature appealing behavior both at the tails and in the neighborhood of zero, compared to the other distributions already used in the literature. GBM and Horseshoe are much heavier-tailed, modeling large signals efficiently while shrinking noise-like, small signals toward zero. The distributions are considered in a two-layer hierarchical model, making the corresponding inference problem amenable to Expectation Maximization (EM). We present an explicit, algebraic EM-update rule for the models, yielding two fast and experimentally validated algorithms for signal recovery. Experimental results show that our algorithms outperform state-of-the-art methods on a wide range of sparsity levels and amplitudes in terms of reconstruction accuracy, convergence rate and sparsity. The largest improvement can be observed for sparse signals with high amplitudes.

Portfolio optimization via manifold learning

Alireza Samani, Duke University

In this project a novel method of stock portfolio optimization[1] based on Diffusion Geometry[2] and Factor Analysis[3] is proposed.Diffusion Geometry provides a low dimensional representation of stock data and then Linear Regression Prediction[4] is performed in a select vicinity of the most recent data point using the local covariance matrix obtained by Factor Analysis.I then apply the Markowitz Portfolio Theorem[5] to obtain the optimal capital allocation strategy for the stocks.The proposed method is applied to NASDAQ stock data and is shown to produce generally better predictions compared to similar approaches such as portfolio optimization based on Multifaceted Factor Analysis[6] or low dimensional prediction using ISOMAP[7].

Adaptive MCMC with kernel embeddings

Dino Sejdinovic, Gatsby Unit, University College London

A Kernel Adaptive Metropolis-Hastings algorithm is introduced, for the purpose of sampling from a target distribution with strongly nonlinear support. The algorithm embeds the trajectory of the Markov chain into a reproducing kernel Hilbert space (RKHS), such that the feature space covariance of the samples informs the choice of proposal. The procedure is computationally efficient and straightforward to implement, since the RKHS moves can be integrated out analytically: our proposal distribution in the original space is a normal distribution whose mean and covariance depend on where the current sample lies in the support of the target distribution, and adapts to its local covariance structure. Furthermore, the procedure requires neither gradients nor any other higher order information about the target, making it particularly attractive for contexts such as Pseudo-Marginal MCMC. Kernel Adaptive Metropolis-Hastings outperforms competing fixed and adaptive samplers on multivariate, highly nonlinear

target distributions, arising in both real-world and synthetic examples. Joint work with H. Strathmann, M. Lomeli Garcia, C. Andrieu and A. Gretton.

Learning features for classification

Jure Sokolic, University College London

We present a linear dimensionality reduction method aimed specifically at classification of signals that lie in a union of approximately low-dimensional subspaces. Examples of such signals are face images under varying illumination or sequences of feature points in a video stream associated with the same motion.

We focus on two class case and assume that the class conditioned signals have a Gaussian distribution. This allows us to express the upper bound of the misclassification probability of the maximum-a-posteriori (MAP) classifier in closed form. The problem of dimensionality reduction projection design is then formalised as the minimization of the the upper bound of misclassification probability over a set of projections with desired dimension. The proposed optimization problem is non-convex; however, we obtain a closed form solution via Generalized Singular Value Decomposition (GSVD) of the two data matrices. The GSVD has fast and stable implementation, therefore it can be applied to very high-dimensional data.

Further, we define the incremental discrimination gain which quantifies how much each additional projection contributes to the classification performance. This allows us to exactly determine the number of projections needed to achieve certain performance threshold. The incremental discrimination gain can also be used as the indirect measure of how discriminative the given data representation is. Incremental discrimination gain is related to the decay of generalised singular values and we show how it is related to the eigenvalues and geometry of class conditioned covariance matrices. Additional property of the proposed dimensionality reduction approach is that covariance matrices of the projected signals are diagonal and allow for very efficient MAP classifier implementation.

We also present the extension of the proposed method via one-vs.-all and binary classification tree approaches. Application to face recognition gives competitive results to the state of the art.

Classification of signals with mismatched MAP classifier

Jure Sokolic, University College London

We present a framework for analysis of classification performance when the classifiers parameters are mismatched with respect to the true signal model. Class conditioned signals are assumed to have a Gaussian distribution with a low-rank covariance matrix. The optimal classifier, which minimizes the misclassification probability, is maximum-a-posteriori (MAP) classifier. Central to our approach is the assumption that classifier parameters are mismatched, i.e. mismatched MAP classifier is used for classification. The notion of mismatched MAP classifier is important from two perspectives. First, in practice model parameters have to be estimated from the data, and therefore do not match the signal model exactly. Second, it can be often advantageous from computational efficiency

perspective to use a simpler classifier which is more efficient, however reasonably accurate. Design of such classifiers requires the understanding of the mismatch on classification performance. We approach the analysis of mismatched classification in principled way, by upper bounding the misclassification probability of the mismatched MAP classifier. Specifically, we focus on performance in the low noise-regime and describe the behaviour of misclassification probability in terms of diversity order and measurement gain. Diversity order measures how fast misclassification probability approaches zero in the low-noise regime, and measurement gain measures the offset of the misclassification probability in the low-noise regime. We express diversity order and measurement gain as functions of geometrical properties of true signal model parameters and the mismatched parameters. This allows us to determine the conditions on the mismatched parameters that ensure perfect classification in the low-noise regime, and further, measure how close is the performance of the mismatched MAP classifier to the optimal MAP classifier. We also present the experiments that confirm the validity of our approach.

Achieving compressed sensing physical system via random demodulation

Pingfan Song, Harbin Institute of Technology

In many applications, sampling at the Nyquist rate is inefcient because the signals of interest contain only a small number of signicant frequencies relative to the band limit. For this type of sparse signal with the locations of the frequencies unknown as a priori, this paper describes an efficient physical sensing system based on Random Demodulation, a novel technology inspired by Compressed Sensing. Unlike conventional sensing systems that conform to the Shannon/Nyquist theorem, this sensing system can perform sub-Nyquist sampling for multi-tone analog signal at a much lower rate than Nyquist rate, as well as perform spectrum sensing and signal recovery with high Signal to Noise Ratio (SNR). Several factors such as the accuracy of system impulse response, the cut-off frequency and order of low-pass filters are explored to clarify how much they can impact the performance of this sensing system. Experiments demonstrate that after optimizing these parameters of this sensing system, the sampling rate for successfully recovering multi-tone analog signal and performing spectrum sensing can be as low as only 4 % of Nyquist rate.

Low-rank tensor recovery via Theta bodies

Zeljka Stojanac, University of Bonn

We present the problem of recovery of low-rank tensors from a small number of measurements. We consider one of the generalizations of the matrix singular value decomposition to tensors, called canonical decomposition (CP-decomposition) and the corresponding notions of rank and norm. The CP-decomposition and therefore its norm are in general NP-hard to compute. To overcome this difficulty we suggest convex relaxations of the norm called Theta bodies, which were originally introduced by Lovász for finding a maximal stable set in a graph. The computation of Theta bodies relies heavily on finding the Groebner basis of the appropriate defined ideal. The Theta bodies define

new norms which can be computed using semidefinite programming. We use one of them to recover low-rank tensors from incomplete information via norm-minimization. In addition, numerical results will be presented.

Simple consistent distribution regression on compact metric domains

Zoltan Szabo, Gatsby Unit, University College London

In a standard regression model, one assumes that both the inputs and outputs are finite dimensional vectors. We address a variant of the regression problem, the distribution regression task, where the inputs are infinite dimensional entities, probability measures. Many important machine learning tasks fit naturally into this framework including multi-instance learning, point estimation problems of statistics without closed form analytical solutions, or tasks where simulation-based results are computationally expensive. Learning problems formulated on distributions have an inherent two-stage sampled challenge: only samples from sampled distributions are available for observation and one has to construct estimates based on these sets of samples. We propose an algorithmically simple and parallelizable ridge regression based technique to solve the distribution regression problem: we embed the distributions to a reproducing kernel Hilbert space and learn the regressor from the embeddings to the outputs. We show that under mild conditions (on compact metric domains with characteristic kernels) this solution scheme is consistent in the two-stage sampled setup. Specially, we establish the consistency of set kernels in regression (a 15-year-old open question) and offer an efficient alternative to existing distribution regression methods, which focus on compact domains of Euclidean spaces and apply density estimation (which suffers from slow convergence issues in high dimensions). Joint work with Arthur Gretton, Barnabas Poczos, Bharath Sriperumbudur.

Analysis of brain states from multi-region LFP time-series

Kyle Ulrich, Duke University

The local field potential (LFP) is a source of information about the broad patterns of brain activity, and the frequencies present in these time-series measurements are often highly correlated between regions. It is believed that these regions may jointly constitute a "brain state," relating to cognition and behavior. An infinite hidden Markov model (iHMM) is proposed to model the evolution of brain states, based on electrophysiological LFP data measured at multiple brain regions. A brain state influences the spectral content of each region in the measured LFP. A new state-dependent tensor factorization is employed across brain regions, and the spectral properties of the LFPs are characterized in terms of Gaussian processes (GPs). The LFPs are modeled as a mixture of GPs, with state- and region-dependent mixture weights, and with the spectral content of the data encoded in GP spectral mixture covariance kernels. The model is able to infer an estimate of the number of brain states and the number of mixture components in the mixture of GPs. A new variational Bayesian split-merge algorithm is employed for inference. The model infers state changes as a function of external covariates in two novel electrophysiological datasets, using LFP data recorded simultaneously from multiple

brain regions in mice; the results are validated and interpreted by subject-matter experts.

Nonlinear information-theoretic compressive measurement design

Liming Wang, Duke University

We investigate design of general nonlinear functions for mapping high-dimensional data into a lower-dimensional (compressive) space. The nonlinear measurements are assumed contaminated by additive Gaussian noise. Depending on the application, we are either interested in recovering the high-dimensional data from the nonlinear compressive measurements, or performing classification directly based on these measurements. The latter case corresponds to classification based on nonlinearly constituted and noisy features. The nonlinear measurement functions are designed based on constrained mutual-information optimization. New analytic results are developed for the gradient of mutual information in this setting, for arbitrary input-signal statistics. We make connections to kernel-based methods, such as the support vector machine. Encouraging results are presented on multiple datasets, for both signal recovery and classification. The nonlinear approach is shown to be particularly valuable in high-noise scenarios.

Semi-deterministic sensing matrices by partial randomly phase modulated unit-norm tight frames

Peng Zhang, Imperial College London

In this work, we propose a novel compressed sensing (CS) framework that consist of three parts: a unit-norm frame (UTF), a random diagonal matrix and a unitary matrix. We prove that this structure satisfies the restricted isometry property (RIP) with high probability if the unitary matrix is bounded and the number of measurements is linear in the sparsity level and (poly-)logarithmic in the signal dimension.

We demonstrate that random subsampling operators in existing structured sensing matrices, such as randomly subsampled orthogonal transforms, random convolution, random subsampling of bounded orthonormal system and etc, can be replaced by operators that offer arbitrary/deterministic selections of measurement vectors. Besides the simplification towards practical implementations, the resultant measurement models can provide comparable (or better) recovery performances.

Moreover, we show that the unified framework can encompass many of existing CS measurement models that consist arbitrary/deterministic subsampling operations, such as random circulant matrices, compressive multiplexing, random demodulation and etc. The corresponding analysis leads to either improvements on the existing RIP results or extended sensing schemes that support new applications (e.g. multiple images compression) and more options of practical implementations.

Finally, we propose the construction of new sensing matrices by the addition of deterministic phase modulations to two existing CS measurement models: partial random circulant matrices and matrices by random demodulation. Both of their RIP conditions can be easily proved in our framework. In addition, we discuss a natural application of the structure of random demodulation with deterministic phase modulations for

channel estimation of orthogonal frequency division multiplexing (OFDM) systems. This channel estimation scheme can supersede previous CS based methods due to its capability to achieve a low Peak-to-Average Power Ratio (PAPR) and a low sampling rate simultaneously.

Compressed sensing non-uniformly sparse signals: An asymptotically optimal power allocation

Xiaochen Zhao, Imperial College London
Wei Dai, Imperial College London

In compressed sensing, Gaussian random matrices with i.i.d. entries are commonly used. However, these matrices are not optimal when different components of the unknown sparse signals have different non-zero probabilities (i.e., non-uniformly sparse signals are concerned). In this work, we assume a fixed total power budget but allow a power allocation across the columns of the sensing matrix. We are interested in finding the power allocation policy that minimises the mean squared reconstruction errors (MSE). Based on the approximate message passing (AMP) algorithm and the associated analysis, we first quantify the asymptotic MSE performance for a given power allocation policy, and then use this result to derive an asymptotically optimal power allocation. Simulations demonstrate that the asymptotic results are accurate for reasonably large systems and hence can be useful in practice.

Block-structured sparse tensor decomposition for classification of multi-dimensional data

Syed Zubair , University of Surrey
Wenwu Wang, University of Surrey
Jonathon Chambers, Loughborough University

Block sparsity has been employed recently in vector/matrix based sparse representations to improve their performance for data classification. It is known that tensor based representation has potential advantages over vector/matrix based representation in retaining the spatial relations within the data. In this paper, we extend the concept of block sparsity for tensor decomposition, and develop a new algorithm for obtaining sparse tensor decomposition with block structures. The proposed algorithm is then used for classification of multi-dimensional data, such as face recognition. Experiments are provided to demonstrate the performance of the proposed algorithm, as compared with several sparse representation based classification algorithms.

Whiteboard Presentations

Beyond stochastic gradient descent for large-scale machine learning

Francis Bach, INRIA

Many machine learning and signal processing problems are traditionally cast as convex optimization problems. A common difficulty in solving these problems is the size of the data, where there are many observations ("large n") and each of these is large ("large p"). In this setting, online algorithms such as stochastic gradient descent which pass over the data only once, are usually preferred over batch algorithms, which require multiple passes over the data. Given n observations/iterations, the optimal convergence rates of these algorithms are $O(1/\sqrt{n})$ for general convex functions and reaches $O(1/n)$ for strongly-convex functions. In this talk, I will show how the smoothness of loss functions may be used to design novel algorithms with improved behavior, both in theory and practice: in the ideal infinite-data setting, an efficient novel Newton-based stochastic approximation algorithm leads to a convergence rate of $O(1/n)$ without strong convexity assumptions, while in the practical finite-data setting, an appropriate combination of batch and online algorithms leads to unexpected behaviors, such as a linear convergence rate for strongly convex problems, with an iteration cost similar to stochastic gradient descent. Joint work with Nicolas Le Roux, Eric Moulines and Mark Schmidt.

Hard thresholding pursuit algorithms: The greedy way

Jean-Luc Bouchot, Drexel University // RWTH Aachen University

In this talk, we review ideas from the Hard Thresholding Pursuit algorithm for sparse recovery. It is shown that uniform recovery is achieved in a number of iterations that scale linearly with the sparsity of the unknown signal, under a mild Restricted Isometry Condition. These results hold also true for a greedy variant of the algorithm (Graded HTP - GHTP) that does not require the prior knowledge of the sparsity.

We also show that the GHTP algorithm recovers certain vector in a number of iterations that exactly matches the sparsity of the input signal. Our theoretical findings are illustrated by numerical examples.

Designer Bayes factorizations: Applications to tensors & networks (invited)

David Dunson, Duke University

It is increasingly common to collect high-dimensional data that have a tensor or network structure in a wide variety of applied domains. Examples include multivariate categorical data analysis, contingency tables, relational data, and brain activity networks among others. It has proven challenging to accurately characterize such data from limited training examples while providing uncertainty estimates. With this broad goal in mind, we propose a theoretically supported approach for designing priors for high-dimensional low-sample size data. The proposed approaches are provably flexible, allow automatic learning of the rank and other key tuning parameters, can be implemented efficiently including in high-dimensional settings, and can lead to optimal rate performance. We

focus in particular on a sparse PARAFAC factorization for categorical data, and a novel factorization for population distributions of networks.

High-dimensional change-point detection with sparse alternatives (invited)

Farida Enikeeva, University of Poitiers

We consider the problem of detecting a change in mean in a sequence of Gaussian vectors. We assume that the change occurs only in some subset of the vector components of unknown size. We construct a procedure of testing the change in mean adaptively to the number of changing components. Under high-dimensional assumptions on the vector dimension and on the sequence length we obtain the detection boundary for this problem and show the minimax rate-optimality of the proposed test.

Asymptotic independence of highly coupled very high dimensional data

Erol Gelenbe, Imperial College London

Very high dimensional data presents huge challenges of statistical analysis which can be alleviated if we understand the internal dynamic couplings and interactions which drive the systems that generate the data. We consider a class of systems that we will conveniently call G-Networks, E-Networks and A-Networks that arise in different applications. G-Networks arise when an unbounded number of particles influence each others' routing and dissipation in a high dimensional network. E-Networks arise with unbounded numbers of particles which enable each others passages through a large number of coupled channels, and A-Networks arise from the modelling of a large number of simultaneous and coupled economic interactions such as auctions. In all these cases we show how explicit quasi-independent behaviours emerge asymptotically to simplify the statistical analysis of these systems.

A new look at mean embeddings

Steffen Grunewalder, University College London

We take a fresh look at mean embeddings in the context of empirical process theory. We argue that what makes the embedding approach so successful are less the concrete estimators which are deployed but rather the coarse granularity with which objects are measured. We develop the notation of measurement devices to make this measurement process explicit and we explore new ways to apply the embedding techniques, most of which are made possible through powerful techniques of empirical process theory. In particular, we demonstrate how these can be applied to coarse density estimation, multiple hypothesis testing, expectation estimation of non-RKHS functions, moment estimation and even applications such as the bandit problem. We furthermore develop novel fast and simple conditional expectation estimators by restricting their measurement resolution. The statistical guarantees we gain are a considerable improvement on what is known in the literature: for instance, we neither need approximation errors nor strong assumptions contrary to existing results.

Breaking the coherence barrier – A new theory for compressed sensing

Anders Hansen, University of Cambridge

Compressed sensing is based on the three pillars: sparsity, incoherence and uniform random subsampling. In addition, the concepts of uniform recovery and the Restricted Isometry Property (RIP) have had a great impact. Intriguingly, in an overwhelming number of inverse problems where compressed sensing is used or can be used (such as MRI, X-ray tomography, Electron microscopy, Reflection seismology etc.) these pillars are absent. Moreover, easy numerical tests reveal that with the successful sampling strategies used in practice one does not observe uniform recovery nor the RIP. In particular, none of the existing theory can explain the success of compressed sensing in a vast area where it is used. In this talk we will demonstrate how real world problems are not sparse, yet asymptotically sparse, coherent, yet asymptotically incoherent, and moreover, that uniform random subsampling yields highly suboptimal results. In addition, we will present easy arguments explaining why uniform recovery and the RIP is not observed in practice. Finally, we will introduce a new theory that aligns with the actual implementation of compressed sensing that is used in applications. This theory is based on asymptotic sparsity, asymptotic incoherence and random sampling with different densities. This theory supports two intriguing phenomena observed in reality: 1. the success of compressed sensing is resolution dependent, 2. the optimal sampling strategy is signal structure dependent. The last point opens up for a whole new area of research, namely the quest for the optimal sampling strategies.

Bayesian models for social interactions (invited)

Katherine Heller, Duke University

A fundamental part of understanding human behavior is understanding social interactions between people. We would like to be able to make better predictions about social behavior so that we can improve people's social interactions or somehow make them more beneficial. This is very relevant in light of the fact that an increasing number of interactions are happening in online environments which we design, but is also useful for offline interactions such as structuring interactions in the work place, or even being able to advise people about their individual health based on who they've come into contact with.

I will focus on a recent project which uses Bayesian methods to predict group structure in social networks based on the social interactions of individuals over time in the form of actual events (emails, conversations, etc.) instead of declared relationships (e.g. facebook friends). The time series of events is modeled using Hawkes processes, while relational grouping is done via the Infinite Relational Model. We also look at measuring social influence and power within this framework, temporally via conversational turn-taking, and by incorporating a language model from which we can measure linguistic accommodation.

Inference in high-dimensional varying coefficient models

Mladen Kolar, University of Chicago Booth School of Business

Damian Kozbur, ETH Zurich

Varying coefficient models have been successfully applied in a number of scientific areas ranging from economics and finance to biological and medical science. Varying coefficient models allow for flexible, yet interpretable, modeling when traditional parametric models are too rigid to explain heterogeneity of sub-populations collected. Currently, as a result of technological advances, scientists are collecting large amounts of high-dimensional data from complex systems which require new analysis techniques. We focus on the high-dimensional linear varying-coefficient model and develop a novel procedure for estimating the coefficient functions in the model based on penalized local linear smoothing. Our procedure works for regimes which allow the number of explanatory variables to be much larger than the sample size, under arbitrary heteroscedasticity in residuals, and is robust to model misspecification as long as the model can be approximated by a sparse model. We further derive an asymptotic distribution for the normalized maximum deviation of the estimated coefficient function from the true coefficient function. This result can be used to test hypotheses about a particular coefficient function of interest, for example, whether the coefficient function is constant, as well as construct confidence bands for covering the true coefficient function. Construction of the uniform confidence bands relies on a double selection technique that guards against omitted variable bias arising from potential model selection mistakes. We demonstrate how these results can be used to make inference in high-dimensional dynamic graphical models.

Fast and robust multiscale methods for high-dimensional data (invited)

Mauro Magionni, Duke University

We will discuss techniques for the multiscale analysis of data sets in high-dimensional spaces that have low intrinsic dimensionality. These methods may be used to estimate intrinsic dimension, construct sparse representations (dictionary learning) and provide fast invertible transforms mapping high-dimensional data sets to a sparse set of coefficients, yield a version of compressed sensing for manifold-distributed data, and novel regression and classification techniques. We will discuss the overarching scheme behind these construction, and discuss details and guarantees on applications chosen by the attendants.

Compressed sensing with side information

João Mota, University College London

The success of Compressed Sensing (CS) owes much to the fact that its main assumption is satisfied in practice. That assumption is simply that the signal to reconstruct is compressible, that is, it can be represented in a sparse way. Another assumption that is also common in practice, but has rarely been addressed in CS, is that side information is usually available when reconstructing a signal. By side information we mean a signal similar or correlated to the signal we are acquiring. This is common, for example, when reconstructing sequences of signals, such as in video and estimation problems, or when we know a priori what our signals will look like, as for example in medical imaging. In this work, we consider CS in the presence of side information. In particular, we integrate

side information into CS by adding to the objective of basis pursuit the L1-norm (resp. L2-norm) of the difference between the optimization variable and the side information. This yields a problem that we refer to as L1-L1 (resp. L1-L2) minimization. We establish bounds on the number of measurements necessary to reconstruct the acquired signal for both L1-L1 and L1-L2 minimization and compare them with the best known bound for standard CS. We find that, if side information has a sufficiently good quality, L1-L1 minimization requires much fewer measurements than both standard CS and L1-L2 minimization, thereby improving significantly the performance of CS. We provide an explanation in terms of geometry of the underlying problem and present experimental results that confirm our findings.

Visual pattern encoding on the Poincaré sphere

Aleksandra Pizurica, Ghent University

In this talk we present a new theory of encoding visual patterns as constellations of signal points in a space spanned by psychophysical features, such as randomness, granularity, directionality, etc. The idea is to represent elementary visual patterns, like image patches or learned image atoms, as codewords in a space with well-defined and psychophysically intuitive structure. This new pattern-encoding scheme is inspired by the graphical representation of polarization states of a light wave on the Poincaré sphere, commonly used in optics and in digital optical communications. Defining similar representations for visual patterns is challenging. We illustrate some of the possible applications in visualizing the properties of learned dictionaries of image atoms and in patch-based image processing.

Kernel MMD, the median heuristic and distance correlation in high dimensions

Aaditya Ramdas, Carnegie Mellon University

This paper is about two related methods for two sample testing and independence testing which have emerged over the last decade: Maximum Mean Discrepancy (MMD) for the former problem and Distance Correlation (dCor) for the latter. Both these methods have been suggested for high-dimensional problems, and sometimes claimed to be unaffected by increasing dimensionality of the samples. We will show theoretically and practically that the power of both methods (for different reasons) does actually decrease polynomially with dimension. We also analyze the median heuristic, which is a method for choosing tuning parameters of translation invariant kernels. We show that different bandwidth choices could result in the MMD decaying polynomially or even exponentially in dimension.

Stein shrinkage for cross-covariance operators and kernel independence testing

Aaditya Ramdas, Carnegie Mellon University

Cross-covariance operators arise naturally in many applications using Reproducing Kernel Hilbert Spaces (RKHSs) and are typically estimated using an empirical plugin estimator,

which we demonstrate are poor estimators of operator (eigen)spectra at low sample sizes. This paper studies the phenomenon of Stein shrinkage for infinite dimensional cross-covariance operators in RKHSs, as briefly initiated by Muandet et al (2014) who recently suggested two shrinkage estimators. We develop a third family of shrinkage estimators and undertake a study of how shrinkage improves estimation of operator spectra. We demonstrate an important and surprising application, that shrunk test statistics yield higher power for kernel independence tests and we provide insights into why they improve performance.

The distribution of restricted least squares with a Gaussion matrix (invited)

Galen Reeves, Duke University

We consider the problem of recovering an unknown vector from noisy linear measurements. We assume a random Gaussian matrix but place no assumptions on the signal or the noise. We study the performance of the least squares estimate restricted to a low-dimensional subspace and characterize the bivariate distribution on two key quantities: the reconstruction error and the residual error. This distribution depends only on the distance between the true signal and the subspace, the mean-squared measurement error, and the problem dimensions. This result is important since is provides the conditional distribution on the reconstruction error (which is unknown) as a function of the residual error (which is observed from the data).