Sparse structures: statistical theory and practice, Bristol, June 2010

Abstracts

Invited session 1

Alexandre Tsybakov (Paris VI, France)

Estimation of high-dimensional low rank matrices

Suppose that we observe entries or, more generally, linear combinations of entries of an unknown $m \times T$-matrix A corrupted by noise. We are particularly interested in the high-dimensional setting where the number mT of unknown entries can be much larger than the sample size N. Motivated by several applications, we consider estimation of matrix A under the assumption that it has small rank. This can be viewed as dimension reduction or sparsity assumption. In order to shrink towards a low-rank representation, we investigate penalized least squares estimators with a Schatten-p quasi-norm penalty term, $p \leq 1$. We study these estimators under two possible assumptions – a modified version of the restricted isometry condition and a uniform bound on the ratio "empirical norm induced by the sampling operator/Frobenius norm". The main results are stated as non-asymptotic upper bounds on the prediction risk and on the Schatten-q risk of the estimators, where $q \in [p, 2]$. The rates that we obtain for the prediction risk are of the form rm/N (for $m = T$), up to logarithmic factors, where r is the rank of A. The particular examples of multi-task learning and matrix completion are worked out in detail. The proofs are based on tools from the theory of empirical processes. As a by-product we derive bounds for the kth entropy numbers of the quasi-convex Schatten class embeddings $S_p^M \hookrightarrow S_2^M$, $p < 1$, which are of independent interest. Joint work with A. Rohde.

Ann Lee (Carnegie Mellon University, USA)

Exploiting sparse structure by Spectral Connectivity Analysis

For naturally occurring data, the dimension of the given input space is often very large while the data themselves have a low intrinsic dimensionality. Spectral kernel methods are non-linear techniques for transforming data into a coordinate system that efficiently reveals the geometric structure – in particular, the "connectivity" – of the data. In this talk, I will focus on one particular technique – diffusion maps – but the analysis can be used for other spectral methods as well. I will give examples of various applications of the method in dimensionality reduction, image retrieval and texture discrimination. I will also present recent results on how spectral kernel methods relate to classical kernel smoothing. (Part of this work is joint with L. Wasserman, C. Schafer, S. Lafon, and R.R. Coifman)

Contributed session 1

Jianxin Pan (University of Manchester)

Modelling of Large Covariance Matrices

It is well known that when analysing longitudinal/clustered data, misspecification of covariance structures may lead to very inefficient or even biased estimators of parameters in the mean. Covariance structures, like the mean, can be modelled using linear or nonlinear regression models techniques. Various estimation methods have been recently developed for modelling of mean and covariance structures, simultaneously. In this talk, methods on modelling of mean-covariance structures will be reviewed, including linear/non-linear regression models, variable selection, semiparametric models, etc. Real examples and simulation studies will be presented for illustration.

Barbara Engelhardt (University of Chicago)

Sparse low-dimensional matrix factorization methods applied to biological data with latent structure

In recent work, we applied sparse low-dimensional matrix factorization methods to population genetic data with various types of latent structure. We found that the ability of the diverse methods to elucidate the latent structure depended heavily on the interaction between the type of model the data conformed to (as described, e.g., by the covariance matrix of the individuals) and the methods for inducing sparsity. In current work, we generalize these observations on simulated and real biological data, including gene expression data. In these data, the correlation structure may partly reflect underlying relationships among networks of genes. We compare the performance of a number of different low-rank matrix factorization methods, including L_1 type methods such as sparse PCA, and Bayesian models with sparsity-inducing priors such as automatic relevance determination and spike-slab priors.

Invited session 2

Martin Wainwright (Berkeley, USA)

Graphical model selection in high dimensional settings: Practical methods and fundamental limits

Undirected graphical models or Markov random fields are used to model a variety of spatial phenomena. The problem of graphical model selection is to use data, assumed to be generated by some underlying graphical model, to determine the correct graph structure. This structure itself is of interest in various applied settings, including gene network analysis in biology, and social network analysis.

In this talk, we discuss the use of ℓ_1-regularization and related methods for undirected graphical model selection. We present some theory that provides sufficient conditions for success under high-dimensional scaling, meaning that the graph size p and graph degrees d are allowed to scale with the sample size n. We illustrate the use of these methods in application to social network analysis of politician's voting records. Finally, we discuss the use of information-theoretic methods to show that our method is optimal up to constant factors, meaning that no method can perform with substantially fewer samples.

Based on joint works with John Lafferty (CMU), Pradeep Ravikumar (UT Austin) and Prasad Santhanam (Univ. Hawaii).

Christophe Ambroise (CNRS, Paris, France)

Inferring Sparse Gaussian Graphical Models for Biological Networks

Gaussian Graphical Models provide a convenient framework for representing dependencies between variables. In this framework, a set of variables is represented by an undirected graph, where vertices correspond to variables, and an edge connects two vertices if the corresponding pair of variables are dependent, conditional on the remaining ones. Recently, this tool has received a high interest for the discovery of biological networks by l1-penalization of the model likelihood.

In this presentation, we introduce various ways of inferring sparse co-expression networks based on partial correlation coefficients from either steady-state or time-course transcriptomic data. All proposals search for a latent structure of the network to drive the selection of edges through an adaptive ℓ_1-penalization of the model likelihood. We focus on inference from samples collected in different experimental conditions and therefore not identically distributed.

Contributed session 2

Guillaume Obozinski (INRIA) (joint with Jenatton, R. and Bach, F.)

Structured Sparse Principal Component Analysis

We present an extension of sparse PCA, or sparse dictionary learning, where the sparsity patterns of all dictionary elements are structured and constrained to belong to a prespecified set of shapes. This structured sparse PCA is based on a structured regularization recently introduced by Jenatton et al.(2009). While classical sparse priors only deal with cardinality, the regularization we use encodes higher-order information about the data. We propose an efficient and simple optimization procedure to solve this problem. Experiments with two practical tasks, the denoising of sparse structured signals and face recognition, demonstrate the benefits of the proposed structured approach over unstructured approaches.

Kevin Sharp (University of Manchester)

Dense Message Passing for Sparse Principal Component Analysis

We describe work on inference algorithms for sparse Bayesian Factor Analysis with a zero-norm prior on the model parameters and present a novel message passing algorithm for sparse Bayesian PCA [1]. So called zero-norm priors assign finite probability mass to sparse solutions. They may be preferable to other types of shrinkage prior since they better characterise a prior belief in sparsity and should therefore lead to more meaningful posterior beliefs. However, Bayesian inference is very challenging in probabilistic models of this type. The highly multimodal posterior causes severe difficulties for standard mean-field variational Bayes algorithms, while in high-dimensional settings MCMC procedures are too slow to be practical.

We present a novel Dense Message Passing algorithm (DMP) [1] similar to algorithms developed in the statistical physics community and previously applied to inference problems in coding and sparse classification. For the single factor case with isotropic noise, which corresponds to sparse probabilistic principal component analysis (PCA), a statistical mechanics theory of optimal learning can be derived. In previous work [2] we found the theoretical predictions to be confirmed by MCMC while deterministic approximate Bayesian inference methods were shown to be suboptimal. Here we show how DMP achieves near-optimal performance on synthetic data and outperforms two other published, state-of-the-art algorithms for sparse PCA, SPCA [3] and emPCA [4]. We also compare the performance of DMP on the same two gene expression datasets used in [3] and [4] and find our

method to outperform both of these algorithms under a standard performance measure.

[1] Kevin Sharp and Magnus Rattray, Dense Message Passing for Sparse Principal Component Analysis. To appear in Proc. of the 13th International Conference on Artificial Intelligence and Statistics (AIstats 2010).

[2] M. Rattray, O. Stegle, K. Sharp, and J. Winn. Inference algorithms and learning theory for bayesian sparse factor analysis. Journal of Physics: Conference Series, 197:012002 (10pp), 2009.

[3] H. Zou, T. Hastie, and R. Tibshirani. Sparse principal component analysis. Journal of computational and graphical statistics, 1:265, 2006.

[4] C. Sigg and J. Buhmann. Expectation maximization for sparse and non-negative PCA. In Proceedings of the 25th Annual International Conference on Machine Learning (ICML 2008), p.960-967, 2008.

Invited session 3

Chris Holmes (University of Oxford)

Bayesian nonparametric clustering of sparse signals

Bayesian mixture modelling is a widely used method for cluster analysis as it allows for characterisation and uncertainty surrounding principle dependence structures within data. Clustering is often used as an exploratory method to uncover hidden structure or to recover suspected structure. As data becomes ever cheaper to capture there is an increasing need for methods that can adjust towards many recorded variables being irrelevant to the clustering task. Variable selection has a substantial literature in regression or supervised learning settings but is much less studied in unsupervised clustering problems.

In this talk we will report on the use of variable selection priors for nonparametric mixture models. Such priors tend to induce sparsity in the posterior model space and help characterise the relative information of those explanatory variables useful for clustering. We demonstrate how the use of hierarchical Dirichlet process priors allows for a principled way to construct these models; providing accurate indication of irrelevant variables whilst being able to quantify the relative relevance of informative variables. We pay particular attention to efficient MCMC sampling schemes for inference allowing for an unknown number of mixture components and an unknown number of relevant variables.

Contributed session 3

Haeran Cho and Piotr Fryzlewicz (Department of Statistics, London School of Economics)

Variable selection via tilting

Recent technological advances have led to the explosion of data across many scientific disciplines, with the dimensionality of the data often exceeding the number of observations. Consider the problem of estimating a p-vector of parameters from a linear model with n observations. In this setting (as in many others), identifying the subset of significant variables can improve estimation accuracy as well as model interpretability. Many approaches to variable selection have recently been developed in which the component-wise correlation screening between each variable and the response plays an important role. However, ordinary correlation can mislead the variable selection procedure due to high cross-correlation between the variables, especially in high dimensions.

In this paper, we propose a new measure of correlation between the variables and the response,

which takes into account the between-variables cross-correlation. This is done by "tilting" the data structure as follows. Each variable is projected onto the orthogonal complement of the space spanned by the variables with which it attains high correlation, and the correlation with the response is observed in this space, so that the true contribution of this variable can be measured more accurately. We discuss the conditions under which the tilting leads to successful variable selection and propose an iterative algorithm for practical application.

Silvia Liverani (University of Bristol)

Bayesian Partition Model selection for high-dimensional time series

One important recent domain where partition models are used is in the study of microarray models. Microarray experiments that measure gene expression of tens of thousands of genes are now widespread and their sheer size presents a challenge to any probability distribution guided clustering algorithm. More recently a large number of experiments have been performed that collect short longitudinal time courses - or time profiles - of microarrays. These profiles have been very useful in helping scientists to discover new genes in the various regulatory pathways in the studied organisms. Because of its transparency one particularly successful methodology within this domain has been the use of MAP model selection on partitions of different clusters.

However, this type of exploration is usually performed under the assumption that units lying in different clusters express independently of one another. In the context where the clusters form part of a mutual regulatory system this independence assumption is not tenable since regulation must logically entail dependence between the regulating unit in one cluster affecting the development of a regulated unit in a different cluster. In particular this means that no search can output hypotheses about how clusters might communicate with one another: one of the features of the process of most interest to the scientist.

Search over this dependence space would be infeasible in general. However, it is fortunate in the context of the given problem that the number of relationships between clusters can be safely hypothesised to be sparse for two reasons. First it makes biological sense that only certain sets of co-expressing genes regulate those in another set. Second, even if this were not the case if a model describes too many simultaneous regulators on a single cluster the associated science of the model becomes impenetrable. This makes it possible for us to design intelligent exploration algorithms so that some of the most promising regulatory models are traversed. The MAP score functions of the dependent models, whist being of closed form, are not as simple as those in the original space but it is nevertheless possible to adapt current propagation algorithms to provide exact or fast approximate scores for the dependent partition. We feed back the MAP dependence model to the scientist using a novel graph. Its semantics not only provides a formal description of the underlying explanatory probability model but also closely matches semantics familiar to the scientist.

Contributed session 4

Florian Frommlet (with Malgorzata Bogdan, Arijit Chakrabarti, Jayanta K.Ghosh, Vienna)
Bayes oracle and asymptotic optimality of multiple testing procedures under sparsity

We investigate asymptotic optimality of a large class of multiple testing rules using the framework of Bayesian Decision Theory. We consider the parametric setup, where the observations come from a normal scale mixture model, and assume a loss which is additive with respect to individual tests. Our model can be used for testing point null hypotheses of no signals (zero effects), as well as to distinguish large signals from a multitude of very small effects. Optimality of a rule is proved by showing that the ratio of its Bayes risk and that of the Bayes oracle (a rule which minimizes the Bayes risk) converges to one within our chosen asymptotic framework. Our main interest is in the asymptotic scheme under which the proportion p of "true" alternatives converges to zero. We fully characterize the class of fixed threshold multiple testing rules which are asymptotically optimal and hence derive conditions for the asymptotic optimality of the rules controlling the Bayesian False Discovery Rate (BFDR). We also provide conditions under which the popular Benjamini-Hochberg procedure is asymptotically optimal and show that for a wide class of sparsity levels, its threshold can be approximated very well by a non-random threshold. Our results extend to more general priors as well as to model selection in a linear regression setting under orthogonality.

Andrew Smith & Arne Kovac (University of Bristol)
Penalised Regression on a Graph

Nonparametric regression means taking observations with complicated structure and fitting simpler estimates to them. In many types of regression 'simpler' means 'smoother' but in some types of penalised regression, in particular total variation denoising, 'simpler' means 'sparser'.

We will briefly consider how total variation denoising may be used to find a sparse structure in one-dimensional regression, before generalising it to perform regression on a graph. This is a new type of regression that fits an estimate to observations made at the vertices of a graph. It can be employed when there are no covariate values but there is a graphical structure that suggests which observations are near to each other. With appropriate penalty terms it may be thought of as a generalisation of the nonparametric lasso, and can be used to detect sparse structures in, among other examples, images and UK house price data.

Our generalised version of total variation denoising penalises distance from the data on the vertices, and roughness on the edges of the graph. There are computational challenges in the implementation, so we will see the results of a new, fast algorithm for regression on a graph, and discuss some examples.

Invited session 4

Marten Wegkamp (Florida State University)

Adaptive Rank Penalized Estimators in Multivariate Regression

We introduce a new criterion, the Rank Selection Criterion (RSC), for selecting the optimal reduced rank estimator of the coefficient matrix in multivariate response regression models. The corresponding RSC estimator minimizes the Frobenius norm of the fit plus a regularization term proportional to the number of parameters in the reduced rank model. The rank of the RSC estimator provides a consistent estimator of the rank of the coefficient matrix. The consistency results are valid not only in the classic asymptotic regime, when the number of responses n and predictors p stays bounded, and the number of observations m grows, but also when either, or both, n and p grow, possibly much faster than m. Our finite sample prediction and estimation performance bounds show that the RSC estimator achieves the optimal balance between the approximation error and the penalty term. Furthermore, our procedure has low computational complexity, linear in the number of candidate models, making it particularly appealing for large scale problems. We contrast our estimator with the nuclear norm penalized least squares estimator (NNP). We show that NNP has estimation and prediction properties similar to those of RSC, albeit under stronger conditions. However, it is not as parsimonious as RSC. We offer a simple correction of the NNP estimator which leads to consistent rank estimation.

David Madigan (Columbia University, USA)

High-dimensional modelling for drug safety surveillance

The pharmaceutical industry and regulatory agencies rely on various data sources to ensure the safety of licensed drugs. Recent high profile drug withdrawals have led to increased scrutiny of this activity. Many statistical challenges arise in this context. This talk will describe some of these data sources and the challenges they present, focusing especially on newer large-scale data analyses.

Contributed session 5

S.J. Steel, N. Louw and S. Bierman (University of Stellenbosch)

Variable selection for kernel classification

A variable selection procedure, called surrogate selection, is proposed which can be applied when a kernel classifier such as the support vector machine or kernel Fisher discriminant analysis is used in a binary classification problem. Surrogate selection applies the lasso after substituting the kernel discriminant scores for the binary group labels, and replacing the input variable observations by values calculated from the kernel function. Empirical results are reported illustrating the performance of surrogate selection. The underlying idea is general enough to make its extension to variable selection in other classification and regression contexts feasible.

Colin Campbell (with Yiming Ying and others, University of Bristol)

Multiple kernel learning methods for handling large and complex datasets

Substantial quantities of data are being generated within the biomedical sciences and the successful integration of different types of data remains an important challenge. We begin the talk with a brief overview of unsupervised learning methods for handling multiple types of data [1]. In the principal part of the talk we outline a set of novel approaches to multi-kernel learning for supervised learning [2,3,4] which can handle disparate types of input data. We outline an information-theoretic

approach [2], a simple EM-based approach [3] and probabilistic classifiers related to the Relevance Vector machine [4] which attempt to find a decision function for classification which is sparse in the number of features. These multi-kernel learning algorithms can eliminate data sources from the decision function if the relevant information is contained in another type of data or if the information is very noisy and redundant compared to other, more informative, input data. We show that these methods can give state-of-the-art performance on certain biomedical prediction problems.

[1] Phaedra Agius, Yiming Ying and Colin Campbell. Bayesian Unsupervised Learning with Multiple Data Types. Statistical Applications in Genetics and Molecular Biology: Volume 8, Issue 1, Article 27 (2009).

[2] Yiming Ying, Kaizhu Huang and Colin Campbell. Enhanced Protein Fold Recognition through a Novel Data Integration Approach. BMC Bioinformatics, 2009, 10:267.

[3] Yiming Ying, Colin Campbell, Theodoros Damoulas and Mark Girolami. Class Prediction from Disparate Biological Data Sources using an Iterative Multi-kernel Algorithm. Lecture Notes in Bioinformatics 5780 (2009) pp.427-438.

[4] Theodoros Damoulas, Yiming Ying, Mark Girolami and Colin Campbell. Inferring Sparse Kernel Combinations and Relevance Vectors: An application to subcellular localization of proteins. Proceedings of the Seventh International Conference on Machine Learning and Applications (ICMLA'08), San Diego, California.

Invited session 5

Sara van de Geer (ETH Zurich)

The Lasso with within group structure

We study the group Lasso, where the number of groups is very large, and the sizes of the groups is large as well. We assume there is within group structure, in the sense that the ordering of the variables within groups in some loose sense expresses their relevance. We propose a within group weighting of the variables, and show that with this structure, the group Lasso satisfies a sparsity oracle inequality.

François Caron (INRIA, Bordeaux)

Hierarchical models for dependent sparse linear regression

One of the most common problems in machine learning and statistics consists of estimating the mean response $X\beta$ from a vector of observations y assuming $y = X\beta + \varepsilon$ where X is known, β is a vector of parameters of interest and ε is a vector of stochastic errors. We consider here the case where we have a sequence of dependent regression models $y_t = X_t\beta_t + \varepsilon_t$, $t = 1, \ldots, T$. Typical applications include multi-task learning or time-varying linear regressions. We describe some flexible hierarchical Bayesian models that enforce sparsity over the $\beta_t's$ as well as sharing of this sparsity pattern over different tasks or successive time indices.

Contributed session 6

Keith Knight (Toronto)

An adaptive lasso for correlated predictors

In the case of a group of strongly correlated predictors, the lasso tends to enforce sparsity by selecting a single predictor from the group while setting the coefficients for the remaining predictors to zero. Extensions of the lasso (such as the elastic net (Hastie and Zou, 2004), fused lasso (Tibshirani et al, 2005), OSCAR (Bondell and Reich, 2008) among others) modify the lasso penalty in some way so that the effect of a group of correlated predictors can be shared among the group. In this talk, we will consider augmenting the set of p predictors with sums and differences of predictors and using the lasso (taking $\lambda \to 0$) to select a new set of predictors that can then be used in the lasso for $\lambda > 0$.

Paul Kirk (Imperial College London)

Stability Selection Methods for Biomarker Discovery

Biomarker discovery is a major challenge in computational biology, in which we seek to select a small number of variables (usually genes or proteins) that are useful indicators of biological state or clinical outcome. However, reported biomarkers are often treated with suspicion due to lack of reproducibility. We here address this problem in the context of several disease phenotypes that are all characterised by a strong inflammatory response. We employ a method very closely related to stability selection (Meinshausen and Bhlmann, 2010) in order to identify proteomic disease markers using lasso logistic regression. Our aim is to find a minimal set of stable predictors that permit successful diagnosis of disease outcome. Here, "stable" predictors are ones that have a high probability of appearing in our selected set, as determined by repeated subsampling from the data. We demonstrate that dependencies between variables affect predictor stability, and that the elastic net may be employed to specify how willing we are to accept groups of correlated predictors in our selected set. We further consider ways in which predictive performance and stability may be assessed simultaneously. We argue that the selection probabilities calculated by stability selection methods provide a useful means to target subsequent experimental investigations, potentially allowing us to shift our focus away from selections that are specific to one particular data set.

N. J.-B. Brunel and F. d'Alch-Buc (Université d'Evry)

Sparse autoregressive models for module extraction in biological networks

In the context of biological dynamical systems, we define modules as subgroups of variables with inner interactions but with no interaction between subgroups. We address the identification (extraction) of these modules in order to analyse high-dimensional regulatory networks based on short time series, e.g gene expression profiles. Our approach is based on a (linear) sparse vector autoregressive (AR) modelling of the time series, as it has been already proposed [4, 3, 5, 6, 9]. It consists in two stages: an estimation stage where the parameters are supposed to have a sparse (unknown blockwise) structure, and a re-ordering of the variables based on the estimated autoregressive matrix that extracts automatically the modules. We consider classical estimators of the parameters of autoregressive models and we propose and compare several sparse versions of these estimators based on recently proposed sparse estimators of covariance and concentration matrices [2, 1, 7, 8]. In particular, we consider the case where the variance of the innovation process is unknown and have to be estimated. Once a sparse AR model is identified, the re-ordering of the rows and columns of the AR matrix is cast into a clustering problem of variables. Tests are run on artificial data and benchmark datasets.

References

[1] J. Fan, Y. Feng, and Y. Wu. Network exploration via the adaptive lasso and scad penalties. Annals of Applied Statistics, 2009.

[2] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. Biostatistics, 9(3):432-441, 2008.

[3] Sato J.R. Garay-Malpartida H.M. Sogayar M.C. Ferreira C.E. Miyano S Fujita, A. Modeling nonlinear gene regulatory networks from time series gene expression data. Journal of Bioinformatics and Computational Biology, 5(5):961-979, 2008.

[4] Sato J.R. Garay-Malpartida H.M. Yamaguchi R. Miyano S. Sogayar M.C. Ferreira C.E. Fujita, A. Modeling gene expression regulatory networks with the sparse vector autoregressive model. BMC Systems Biology, 1(39), 2007.

1 Posters

1. **Vanna Albieri** (Danish Cancer Society)

   A comparison of structural learning procedures for biological networks

2. **Luke Bornn** (University of British Columbia) and Francois Caron (INRIA Bordeau Sud-Ouest)

   The Product Graphical Model

3. **Colin Campbell** (with Yiming Ying and Kaizhu Huang, University of Bristol)

   Sparse regularisation methods for metric learning

4. **Sohail Chand**, (with Chris Brignell and Andrew Wood, University of Nottingham)

   Oracle properties of lasso-type methods

5. **Tom Diethe** (Department of Computer Science, University College London)

   Learning in a Nystrm Approximated Space

6. **Doyo Gragn**, Nickolay T. Trendafilov (Open University)

   Sparse principal components based on semi-divisive clustering of genes

7. **Shota Gugushvili** (Vrije Universiteit Amsterdam)

   $\sqrt{n}$-consistent estimation for systems of ordinary differential equations: bypassing numerical integration via smoothing

8. **Zhou Fang** (University of Oxford)

   Group sparsity through concave penalties

9. **Marie Fitch** and Beatrix Jones (Massey University, New Zealand)

   Sparsity vs Computational Convenience for Estimation of a Sparse Inverse Covariance Matrix

10. **Edmund Jones** (University of Bristol)

    Learning sparse graphical model structures from sparse data

11. **Shakir Mohamed** (joint with Zoubin Ghahramani, University of Cambridge)

    Bayesian Learning with Correlated Spike-and-Slab Priors

12. **Dino Sejdinovic** (University of Bristol)

    Message-passing algorithms in coding and information theory

1. **Vanna Albieri** (Danish Cancer Society)

   A comparison of structural learning procedures for biological networks

   Over the past years, microarray technologies have produced a tremendous amount of gene expression data. The availability of these data has motivated researchers to assess genes function and to gain a deeper understanding of the cellular processes, using network theory as tool for the analysis. An elegant framework for modeling and inferring network structures in biological systems is provided by graphical models. They allow the stochastic description of network associations and dependence structures in complex highly structured data. However, typically gene expression data set includes a large number of variables but only few samples making standard graphical model theories inapplicable. The issues presented by genetic data have led to further extend the theory of graphical models to allow their applications in this area. The main aim of this work is the comparison of recent procedures, which estimate sparse concentration matrices and learn the structure of biological networks, through the use of both simulated and real data. The compared procedures are: G-Lasso algorithm (Friedman et al., 2008), Shrinkage estimator with empirical Bayes approach for model selection (Schafer and Strimmer, 2005), and PC-algorithm (Kalisch and Buhlmann, 2007)

2. **Luke Bornn** (University of British Columbia) and Francois Caron (INRIA Bordeau Sud-Ouest)

   The Product Graphical Model

   In this paper, the authors propose a class of prior distributions on decomposable graphs, allowing for improved modeling flexibility. While existing methods solely penalize the number of edges, the proposed work allows practitioners to control clustering, level of separation, and other features of the graph. Emphasis is placed on a particular prior distribution which derives its motivation from the class of product partition models; the properties of this prior relative to existing priors is examined through theory and simulation. The authors then explore the use of graphical models in the field of agriculture, showing how the proposed prior distribution alleviates the inflexibility of previous approaches in properly modeling the interactions between the yield of different crop varieties.

3. **Colin Campbell** (with Yiming Ying and Kaizhu Huang, University of Bristol)

   Sparse regularisation methods for metric learning

   In application to classification the concept behind metric learning is to devise an adaptive distance measure such that datapoints belonging to the same class are drawn closer together in data space whereas datapoints belonging to different classes are pushed further apart. The application of metric learning to real-life classification problems has lead to significant drops in test error. However, the question remains of how best to scale metric learning to large and complex datasets. In this talk we present several novel schemes for metric learning [1,2,3,4]. We introduce a rank-reducing trace norm regulariser and other regularisers which implement a mapping of higher-dimensional data to a lower dimensional space in addition to implementing metric learning. We show that this approach can lead to state-of-the-art performance on real-life classification problems.

[1] Sparse Metric Learning via Smooth Optimization. Yiming Ying, Kaizhu Huang and Colin Campbell. Advances in Neural Information Processing Systems (NIPS), 2009.

[2] GSML: A Unified Framework for Sparse Metric Learning. Kaizhu Huang, Yiming Ying and Colin Campbell. Proceedings IEEE International Conference on Data Mining, ICDM 2009.

[3] Generalized Sparse Metric Learning With Relative Comparisons. Kaizhu Huang, Yiming Ying and Colin Campbell. Journal submission (2009).

[4] Metric Learning with Sparse Constraints. Yiming Ying, Kaizhu Huang, Massimiliano Pontil and Colin Campbell. Submission to AISTATS, 2010.

4. **Sohail Chand**, (with Chris Brignell and Andrew Wood, University of Nottingham)

   Oracle properties of lasso-type methods

   Lasso-type methods in the regression context are popular for their simultaneous estimation and variable selection. Identification of the right subset is one of the desired oracle properties of these methods. Zhao and Yu (2006, J Mach Learn Res, 7:2541-2563) and Zou (2006, J Am Stat Assoc, 101(476):1418-1429) discussed a necessary and almost sufficient condition for the oracle performance of lasso-type methods. In this talk, we present numerical results which investigate how well the oracle properties of the lasso (Tibshirani, 1996, J Roy Stat Soc B Met, 58(1):267-288) and adaptive lasso (Zou, 2006) are achieved in practice. Both of these methods have been implemented using the LARS algorithm (Efron et al., 2004, Ann Stat; 40-451), which suggests normalisation of the predictors. The adaptive lasso uses adaptive weights which makes it an oracle procedure. Our numerical results show how normalisation of the predictors can nullify the advantage of using the adaptive weights and may lead to failure of the necessary and sufficient condition for correct subset selection. Our simulation studies clearly show that there is no advantage, and even a disadvantage in some scenarios, in normalising the predictors. The choice of the regularisation parameter is critical for the oracle performance of these methods. We have compared the performance of cross validation with the Wang and Leng (2009, J Roy Stat Soc B Met; 71(3):671-683) BIC approach in choosing the appropriate value of regularisation parameter. Our results show that the cross validation choice of regularisation parameter may lead to inconsistent variable selection.

5. **Tom Diethe** (Department of Computer Science, University College London)

   Learning in a Nystrm Approximated Space

   Given n observations we define a framework that carries out learning in a $k \leqslant l << n$ dimensional subspace that is constructed using the Nystrm method. We adopt a recently advocated and theoretically justified approach of uniform sub-sampling without replacement to cheaply find a k-dimensional subspace in time complexity $O(1)$. We propose to use any linear learning algorithm in this uniformly sampled k-dimensional Nystrm approximated subspace to help tackle large data sets. Furthermore, we prove a novel upper bound for any general Lipschitz loss function such that we are guaranteed not to loose too much in the Nystrm space, implying successful learning for both classification and regression. Finally, we demonstrate our proposed methodology on several UCI repository data sets for the SVM, Kernel Ridge Regression and Kernel Fisher Discriminant Analysis.

6. **Doyo Gragn**, Nickolay T. Trendafilov (Open University)

   Sparse principal components based on semi-divisive clustering of genes

   A new method for semi-divisive hierarchical clustering of variables (genes) is proposed. The method forms clusters of genes sequentially in two steps. First, the genes are ordered sequentially, either according to the highest sum of squared correlation, or based on the leading singular value of the already sorted genes and one of the unsorted ones. Then, the ordered genes are split in two parts such that the determinant of the correspondingly partitioned correlation matrix is maximized. The first group of genes becomes an output cluster, while the second one – input for another run of the sequential process. After the optimal clusters has been formed, sparse components can be constructed from the leading principal components in each cluster. The method is applied to a real gene expression data and the results compared with other existing approaches.

7. **Shota Gugushvili** (Vrije Universiteit Amsterdam)

   $\sqrt{n}$-consistent estimation for systems of ordinary differential equations: bypassing numerical integration via smoothing

   We consider the problem of parameter estimation for a system of ordinary differential equations from noisy observations on a solution of the system. Most realistic systems are high-dimensional with a high-dimensional parameter space. In case the system is nonlinear, as it typically is in practical applications, an analytic solution to it usually does not exist. Consequently, straightforward estimation methods like the ordinary least squares method depend on repetitive use of numerical integration in order to determine the solution of the system for a sequence of parameter values and to find subsequently the parameter value that minimises the objective function. This induces a huge computational load on such estimation methods. We propose an estimator that is defined as a minimiser of an appropriate distance between a nonparametrically estimated derivative of the solution and the right-hand side of the system applied to a nonparametrically estimated solution. Our estimator bypasses numerical integration altogether and reduces the amount of computational time drastically compared to ordinary least squares. Moreover, we show that under suitable regularity conditions this estimation procedure leads to a $\sqrt{n}$-consistent estimator of the parameter of interest.

   Joint work with Chris Klaassen.

8. **Zhou Fang** (University of Oxford)

   Group sparsity through concave penalties

   In a regression context, sparsity can appear in many forms other than the simple case of covariate selection. For example, many problems can be posed as a situation where we wish to conduct linear regression under a group sparsity constraint - that is to say, where we wish to have model selection where selection of some covariates allow other related covariates to be selected as well, whilst tending to leave unrelated covariates estimated at zero. Several authors have suggested methodology based on the group lasso - consisting of a Lasso-like penalty involving aggregation of L2 norms for each group. However, we suggest an alternative formulation, using concave penalty functions. Such methods can be easily implemented using

LLA or path following algorithms, can be very competitive with the group lasso in many cases, allow multiple levels of sparsity, and can be extended readily to a wide range of problems.

9. **Marie Fitch** and Beatrix Jones (Massey University, New Zealand)

   Sparsity vs Computational Convenience for Estimation of a Sparse Inverse Covariance Matrix

   Graphical models are a popular tool for describing the patterns of conditional independence, where vertices represent variables. In the Gaussian setting, edges in the graph are equivalent to zeros in the inverse covariance matrix. Although a non-decomposable graph (e.g. a large cycle) may be very sparse, and thus have (relatively) few parameters to estimate, it is common in a high-dimensional Bayesian setting to restrict model selection procedures to decomposable models, which are typically less sparse, for computational convenience. We consider estimation of the covariance and inverse covariance matrix where the true model forms a cycle, but estimation is performed supposing that the pattern of zeros is a decomposable graphical model, where the elements restricted to zero are a subset of those in the true matrix. The variance of the maximum likelihood estimator based on the decomposable model is demonstrably larger than for the true non-decomposable model, and which decomposable model is selected affects the variance of particular elements of the matrix. However, the penalty for using the decomposable model is fairly small, even when the difference in sparsity is large and the sample size is fairly small (eg the true model is a cycle of size 50, and the sample size is 51).

10. **Edmund Jones** (University of Bristol)

    Learning sparse graphical model structures from sparse data

    Several methods have been proposed for learning the structure of a Gaussian graphical model from sparse data. In the Bayesian approach, the conjugate prior for the covariance matrix is the generalized hyper inverse Wishart distribution. For the graphs themselves, the only commonly used prior has been the uniform distribution, where each graph is equally likely. This favours medium numbers of edges. But in molecular biology applications, sparse graphs are believed to be more likely than dense ones.

    Considerations about the graph prior are related to two difficulties in calculating the posterior. Firstly, for non-decomposable graphs, there is no analytic formula for the marginal likelihood. Secondly, for large numbers of nodes, the number of possible graphs is enormous, even if only the decomposable ones are counted.

    A further issue is that with sparse data, posteriors are sometimes very sensitive to the prior. This poster will present ideas on how to quantify the sensitivity of the posterior graph distribution to the prior, and results using real and simulated data on small graphs.

11. **Shakir Mohamed** (joint with Zoubin Ghahramani, University of Cambridge)

Bayesian Learning with Correlated Spike-and-Slab Priors

The spike-and-slab prior, which is a mixture of a point mass at zero and any other continuous distribution, has emerged as a useful tool in numerous sparse Bayesian learning problems, including feature selection, regression and unsupervised learning. These priors assume that each of the sparse dimensions are independent, and provide a penalty on the number of non-zero elements, similar to that of the L0-norm.

In certain cases, we may believe that subsets of the these sparse dimensions are correlated and we may wish to explicitly encode this correlation in our prior representation. In this work, we will focus on problems of sparsity in unsupervised learning in latent variable models and provide a Bayesian perspective on learning sparse latent representations with correlation.

A common construction for the spike-and-slab prior is to represent them as a binary vector indicating whether or not a particular latent feature lies in the spike or the slab component, and representing the slab component as a draw from a Gaussian distribution. Learning is achieved in the sampling framework using a pairwise sampling of the latent indicators and the corresponding continuous value. To extending this construction to take into account correlation, we look to construct a correlated binary vector and proceed with a similar pairwise sampling using Metropolis-Hastings.

There are numerous approaches to constructing correlated binary vectors which have been proposed in the literature. Our construction is based on the dichotomisation of a Gaussian distribution. This involves a moment matching procedure to specify the mean and the covariance of a Gaussian distribution, such that after dichotomising a draw from this distribution at zero, we obtain a correlated binary vector with known mean and covariance. We can combine this approach with the existing construction of the spike-and-slab prior to specify a correlated spike-and-slab prior, and demonstrate its applicability in the context latent variable modelling. This approach to learning sparse and correlated representations aims to extend the applicability of the spike-and-slab prior and to highlight the subtleties of Bayesian learning in this setting.

12. **Dino Sejdinovic** (University of Bristol)

Message-passing algorithms in coding and information theory

Sparse graph codes revolutionised the field of error correction coding and are already an important part of modern communication systems. Their novelty lies within the use of iterative message-passing decoding algorithms, inspired by belief propagation for efficient marginalisation of multivariate functions, which allow the error correction performance close to Shannon's capacity with exceptionally low computational complexity. More recently, message-passing has warranted another resurgence of interest as a basis for the iterative thresholding schemes for compressed sensing with an equivalent sparsity-undersampling tradeoff to that of the LASSO-based reconstruction. I will discuss the connections between these two applications of message-passing algorithms in the context of distributed communications.