# Modelling Away Day Talks

## Programme

| | | |
|---|---|---|
| 09.00 | | Arrival Coffee |
| 09.30 | | Opening Remarks |
| 09.40 | Daniel Preotiuc | Where's @wally? A Graph Based Method for Geolocating Users in Social Networks |
| 09.55 | Nicolò Fusi | Automated learning of hidden expression determinants and their interactions with the genetic state |
| 10.10 | Peter Glaus | Bayesian Inference of Differentially Expressed Transcripts from Sequencing Data |
| 10.25 | Jie Hao | A hybrid method of application of independent component analysis to *invivo* $^1$H MR spectra of childhood brain tumours |
| 10.40 | Ciira Maina | Modeling RNA Polymerase II Dynamics |
| 10.55 | Jens Nielsen | TBA |
| 11.10 | Trevor Cohn | TBA |
| 11.30 | Magnus Rattray | Message passing and the cavity method in biophysics models |
| 11.45 | | **Discussion** |
| 12.30 | | Lunch |
| 14.00 | A. Tomkins | How Degrading Networks Can Increase Certain Cognitive Functions. |
| 14.20 | Eleni Vasilaki | Do synaptic dynamics and STDP govern connectivity motifs? |
| 14.45 | Alfredo Kalaitzis | Residual Component Analysis (RCA) |
| 15.05 | James Hensman | Collapsed Variational Bayes |
| 15.25 | Neil Lawrence | TBA |
| 15.45 | | **Discussion** |
| 16.30 | | Closing Remarks |

## Abstracts

### James Hensman: Collapsed Variational Bayes

Variational approximations are an established method for Bayesian learning. Approximate inference is performed by optimising minimising the KL divergence between the approximate posterior and the true posterior, making inference an optimisation problem.

Collapsing the variational problem is an enticing prospect since it can lead to a lower dimensional optimisation problem. A variety of collapsed methodologies have been proposed, with implementations presented for specific models. Here we consider a general method for collapsed implementations, which unifies previous approaches under a new bound.

In this talk we shall show that the coordinate-ascent approach associated with the VB method (known as VBEM) is in fact performing steepest gradient ascent on our proposed bound. We combine our unified collapsed method with an information-geometric optimisation procedure. I'll illustrate the ideas through discussion of the Gaussian mixture model, and demonstrate empirically that it is possible to achieve significantly faster optimisation.

### Alfredo Kalaitzis: Residual Component Analysis (RCA)

Probabilistic principal component analysis (PPCA) seeks a low dimensional representation of a data set in the presence of independent spherical Gaussian noise, $\mathbf{\Sigma} = \sigma^2\mathbf{I}$. The maximum likelihood solution for the model is an eigenvalue problem on the sample covariance matrix. In this paper we consider the situation where the data variance is already partially explained by other factors, e.g. conditional dependencies between the covariates, or temporal correlations leaving some residual variance. We decompose the residual variance into its components through a generalised eigenvalue problem, which we call residual component analysis (RCA).

This new data analysis technique is combined with GLASSO in an EM framework to estimate the sparse inverse and low rank components of a covariance matrix model.

Full covariance matrix models of data are often problematic as their parameterization scales with $D^2$. Two separate approaches to a reduced parameterization of these matrices are to base them on low rank matrices (as in probabilistic PCA) or on a sparse inverse structure (as in GLASSO). These two approaches have very different characteristics: one involves specifying sparse conditional independencies in the data, the other assumes that a reduced set of latent variables is governing the data. Clearly, in any given data set, both of these characteristics may be present. Our sparse plus low rank approach is the first approach to deal with both these cases in the same model. It is demonstrated to good effect in a motion capture and protein network example.

**Nicolò Fusi**: **Automated learning of hidden expression determinants and their interactions with the genetic state**

Genomic studies have revealed substantial genetic control of the transcriptional state of the cell. To fully understand the mechanisms that underly gene expression variability, it is important to investigate the impact of genotype in the context of changing external conditions. In model systems, explicit control of the environment has been considered for this purpose, allowing for direct analysis of genotype-environment interactions. The clear limitation of this experimental approach is the need to profile identical or similar genotypes in a large number of environments, an undertaken that is clearly infeasible when it comes to human studies. Here, we propose a model-based approach to learn environmental factors from the measured gene expression data when the exact environmental state remains unknown. Our method explicitly accounts for the possibility of multiplicative genotype-environment interactions, which allows for improved reconstruction of the environmental unknown environmental state. In experiments on synthetic data and yeast, we show that our method is able to pinpoint gene-environment interactions with greater accuracy than alternative models. Finally, we show that modeling interactions within genetic analyses can improve the power to detect other genetic factors that alter the transcriptional state.

**Peter Glaus**: **Bayesian Inference of Differentially Expressed Transcripts from Sequencing Data**

High-throughput sequencing enables expression analysis at the level of individual transcripts. The analysis of transcriptome expression levels and differential expression estimation requires a probabilistic approach to properly account for the ambiguity caused by shared exons and finite read sampling as well as the intrinsic biological variance of transcript expression. Another important factor are the biological sources of variance, which, as we show in our analysis, can be substantial and may dependent on the transcript expression level. To avoid false positive differential expression calls, one has to anticipate the intrinsic variance of the transcript expression levels using empirical prior knowledge and information from replicates where they exist.

We use probabilistic generative model of read generation to infer transcript expression levels from high-throughput sequencing experiments. Inferred relative expression is in the form of a probability distribution represented by samples of the distribution obtained by Markov chain Monte Carlo algorithm. We us both regular Gibbs sampling algorithm as well as Collapsed Gibbs sampling in which some of the parameters are marginalised in order to obtain faster convergence.

For differential expression analysis of multiple conditions, we use Log-Normal model to include replicates and account for biological variation. This model is applied to pseudo vectors of single MCMC samples from the replicates in order to propagate uncertainty from the sample-level model. We demonstrate the merits of our approach in comparison with other methods by analysing simulated data with known ground truth.

**Ciira Maina**: **Modeling RNA Polymerase II Dynamics**

Gene transcription by RNA polymerase II (pol-II) is a key step in gene expression. Transcription consists of a number of dynamic events such as recruitment of pol-II to the promoter, elongation and termination. Furthermore, each of these steps may be rate limiting and therefore affect the level of gene expression. This motivates the study of transcription dynamics and in particular the effect of these dynamic events on gene expression.

In this work we present a mathematical model that directly models the movement of pol-II down the gene body. This model allows us to compute the transcription speed for each gene and also determine the genes responding differentially to stimuli using high thoughput sequencing data.

**Daniel Preotiuc**: **Where's @wally? A Graph Based Method for Geolocating Users in Social Networks**

I will present current work on a machine learning approach to geolocating users of online networks based on their social connections. Users regularly interact with those closer to themselves and, in most cases, a person's social network is sufficient to disclose their location.

I demonstrate a high-precision method to identify locations given in the profiles of certain users. Based on these, I will propose a graph based method to assign a probability distribution across locations for each user which doesn't provide an explicit location.

**Adam Tomkins**: **How Degrading Networks Can Increase Certain Cognitive Functions.**

Huntingtons is a genetic, progressive neuro-degenerative dis- ease, causing massive network degradation effecting the Medium Spiny Neurons of the striatum. Despite substantial striatal cell atrophy, some cognitive functions have been shown to improve in manifest Hunting- tons disease patients over healthy and pre-symptomatic Huntingtons disease patients. Using a detailed model of the striatal microciruit, we show that combining current ideas about the underlying causes of the disease could lead to the counter-intuitive result of improved competitive network dynamics for signal selection.

**E.Vasilaki (joint work with M. Giugliano): Do synaptic dynamics and STDP govern connectivity motifs?**

Recent evidences in rodent prefrontal cortex (Wang et al, 2006) and olfactory bulb (Pignatelli, Markram, and Carleton, unpub. data) suggest that synaptic short-term facilitation and depression may be correlated to specific connectivity motifs. In particular, it was observed that two excitatory neurons with facilitating synapses form predominantly reciprocal connections, while two excitatory neurons with depressing synapses form unidirectional connections. However, the causes for these structural differences are unknown.

We propose that connectivity motifs could emerge by the interaction of short-term synaptic dynamics (STD) and long-term spike-timing dependent plasticity (STDP). While the influence of STDP on STP was shown experimentally in vitro (Buonomano 1999), how STP and STDP mutually interact in active recurrent networks is largely unexplored. Our approach combines the Tsodyks-Markram (1997) STD phenomenological model with the STDP triplet model (Pfister et al 2006, Clopath et al, 2010), which captures dependencies on both time and frequency of long-term plasticity. As proof of concept, we implement the STD-STDP on networks with random initial topology, composed by adaptive exponential integrate and fire model neurons (Brette & Gerstner, 2005). Synaptic connections in the networks are either all facilitating or all depressing. Upon identical external stimulation patterns, we find that all networks with depressing synapses evolve non-symmetric connectivity motifs, while networks with facilitating synapses evolve reciprocal connectivity motifs, for the largest part of the simulations.

Our model highlights appropriate biophysical conditions under which STP-STDP could explain the correlation between facilitation and reciprocal connectivity motifs as well as between depression and unidirectional connectivity motifs. These specific conditions may lead to the design of experiments for the validation of the proposed mechanism.

Wang Y, Markram H, Goodman P, Berger T, Ma J, Goldman-Rakic, P (2006) Heterogeneity in the pyramidal network of the medial prefrontal cortex, Nature Neurosci. 9(4):534-42.

Pfister J.-P. & Gerstner W. (2006) Triplets of spikes in a model of spike timingdependent plasticity. J. Neurosci. 26, 96739682.

Clopath C, Buesing L, Vasilaki E & Gerstner, W. (2010) Connectivity reflects coding: a model of voltage-based STDP with homeostasis, Nat Neurosci 13, 344352.

Tsodyks MV, Markram H (1997) The neural code between neocortical pyramidal neurons depends on neurotransmitter release probability. Proc Natl Acad Sci U S A 94: 719-723.

Brette R. & Gerstner W (2005) Adaptive exponential integrate-and-fire model as an effective description of neuronal activity. J. Neurophysiol. 94, 36373642.

Buonomano DV (1999) Distinct functional types of associative long-term potentiation in neocortical and hippocampal pyramidal neurons. J. Neurosci. 19(16):6748-54.

**Jie Hao**: **A hybrid method of application of independent component analysis to *invivo* $^{1}$H MR spectra of childhood brain tumours**

Independent component analysis (ICA) has the potential of automatically extract individual metabolite, macromolecular and lipid (MMLip) components from a series of invivo MR spectra. The traditional feature extraction (FE)-based ICA approach is limited, in that a large sample size is required and a combination of metabolite and MMLip components can appear in the same independent component. The alternative ICA approach, based on blind source separation (BSS), is weak when dealing with overlapping peaks. Combining the advantages of both BSS and FE methods may lead to better results. Thus, we propose an ICA approach involving a hybrid of the BSS and FE techniques for the automated decomposition of a series of MR spectra. The hybrid ICA method showed an improvement in the decomposition ability compared with BSS-ICA or FE-ICA, with an increased correlation between the independent components and simulated metabolite and MMLip signals. We were able to automatically extract metabolites from the patient MR spectra dataset that were not in commonly used basis sets (e.g. guanidinoacetate). It has been demonstrated that hybrid ICA provides more realistic individual metabolite and MMLip components than BSS-ICA or FE-ICA. It can aid metabolite identification and assignment, and has the potential for extracting biologically useful features and discovering biomarkers.