Noname manuscript No.
(will be inserted by the editor)

Infinite Factorization of Multiple Non-parametric Views

Simon Rogers · Arto Klami · Janne Sinkkonen · Mark Girolami · Samuel Kaski

Received: date / Accepted: date

Abstract Combined analysis of multiple data sources has increasing application interest, in particular for distinguishing shared and source-specific aspects. We extend this rationale of classical canonical correlation analysis into a flexible, generative and non-parametric clustering setting, by introducing a novel non-parametric hierarchical mixture model. The lower level of the model describes each source with a flexible non-parametric mixture, and the top level combines these to describe commonalities of the sources. The lower-level clusters arise from hierarchical Dirichlet Processes, inducing an infinite-dimensional contingency table between the views. The commonalities between the sources are modeled by an infinite block model of the contingency table, interpretable as non-negative factorization of infinite matrices, or as a prior for infinite contingency tables. With Gaussian mixture components plugged in for continuous measurements, the model is applied to two views of genes, mRNA expression and abundance of the produced proteins, to expose groups of genes that are co-regulated in either or both of the views. Cluster analysis of co-expression is a standard simple way of screening for co-regulation, and the two-view analysis extends the approach to distinguishing between pre- and post-translational regulation.

1 INTRODUCTION

In certain unsupervised learning problems, we are interested in discovering the variation shared by several data sources, sets, modalities, channels, or "views". Practical examples include extracting the shared semantics of original and translated documents (Vinokourov et al, 2003b), discovering dependencies between images and associated text (Vinokourov et al, 2003a), linking the face and sound of a speaker (Englebienne et al, 2008), discovering depth in random-dot stereograms (Becker and Hinton, 1992), and combining mRNA and protein profiles to explore the complex regulatory behavior that underpins a large amount

S. Rogers, M. Girolami
Department of Computing Science, University of Glasgow, UK
E-mail: {rogers,girolami}@dcs.gla.ac.uk

A. Klami, J. Sinkkonen, S. Kaski
Department of Information and Computer Science, Helsinki University of Technology, Finland
E-mail: first.last@tkk.fi

of cellular activity. In some of these examples, the two data sources are defined on similar spaces (Becker and Hinton, 1992; Vinokourov et al, 2003b) whilst in others, the spaces are very different (Englebienne et al, 2008; Vinokourov et al, 2003a). However, in all cases we have the same aim — to investigate not just the details of the individual sources, but also their commonalities.

We will next motivate our work with one particularly timely bioinformatics application. The model we present is, however, generally applicable to any multi-view clustering task and we will return to the general setup after this motivation. In systems biology a re-occurring data analysis setup is the joint analysis of genomic data from multiple sources. Recently, Rogers et al (2008) proposed a coupled clustering model for the specific analysis of a new dataset consisting of two views (mRNA and protein expression) of approximately 500 genes. This is one of the first such datasets produced and as such poses unique challenges. As high-throughput protein measurement becomes more common, it is likely that more data of this kind will be produced, motivating investigation into suitable analysis techniques.

From a biological perspective the primary goal in analysing data of this type is to gain a deeper understanding of regulation at the transcriptional (mRNA) and post-transcriptional/ translational (protein) levels. Simply put, transcriptional regulation involves the switching on and off of genes via the binding of transcription factors to their upstream regions. When a gene is switched on, mRNA is produced. Hence, it has been possible to elucidate networks of transcriptional regulation through the analysis of mRNA data alone (e.g. Friedman et al, 2000). Regulation after transcription can take several forms. For example, translation (producing a protein from the mRNA) can be sped-up or delayed. A third level of control (post-translational; e.g. phosphorylation) is also important but cannot be investigated with the current data — extending the model to extra sources to investigate such effects is straightforward once suitable data is available.

When looking at protein data alone, it is impossible to distinguish whether the observed variance is due to transcriptional or post-transcriptional regulation. Hence, it is necessary to analyse the protein data alongside mRNA data. Cluster analysis is a popular tool for analysing transcriptional regulation due to the reasonable assumption that genes regulated by the same transcription factors will display similar expression profiles. Cluster models are also readily interpretable, which is an important feature for a new method to be taken up by the biological community.

Clustering coupled sources is a problem that has received little attention from the machine learning community. Most of the previous multi-view learning work has been for supervised learning, and the more unsupervised approaches have typically been based on projection approaches like canonical correlation analysis (CCA) and its kernelised (Vinokourov et al, 2003a,b) and probabilistic (Bach and Jordan, 2005) variants. As shown by Rogers et al (2008), the strong cluster structure in the data suggests standard projection techniques would be inappropriate for this data. Earlier Klami and Kaski (2008) proposed a multi-view clustering model based on probabilistic CCA, but their model assumes too simple clustering structure in the joint data space. In particular, each joint cluster should be unimodal within each data space. A related clustering model by Bickel and Scheffer (2004) even explicitly assumes each of the views alone to suffice for learning the shared clustering, and only uses the coupling to improve accuracy.

The model proposed by Rogers et al (2008) attempted to extend coupled clustering to more complex data setups by fitting a mixture to each view, coupled by conditioning the prior for the protein mixture on the assignment to the mRNA mixture. This resulted in a set of probabilities linking the mixture components on each side. When analysing the results from the model, two things stand out. First, there appears to be a high level of inter-

connectivity between the clusters on the two sides. This appears to take the form of *blocks* in the table of connection probabilities corresponding to small, highly connected sets of marginal components from each side. Second, some of the very low probability connections in the table could potentially be due to over-fitting in the maximum likelihood optimisation procedure. These observations motivate the development of a new model.

We introduce a model that removes the restrictions of earlier ones by having very flexible non-parametric models for each of the views, coupled by a flexible model of the interactions between them. The model is implemented as a two-level hierarchy of mixtures. The top-level mixture represents the common variation (each component is a *block* of marginal components), while the second level has a separate independent mixture for each view, describing the view-specific variation conditioned on the top-level component. We present a collapsed Gibbs sampler for estimating the model, allowing us to infer the number of mixture components, both within views and on the top level using a novel hierarchical Dirichlet process (HDP) formulation that extends the standard HDP of Teh et al (2006) by relaxing the assumption of known group assignments.

The remainder of the paper is organised as follows. In Section 2 we present the mixture model, more comprehensively explain the differences between this model and that presented in Rogers et al (2008) and describe the Gibbs sampling scheme. In Section 3 we describe other related work and in Section 4 illustrate the model on two synthetic datasets. In Section 5, we present an analysis of the mRNA and protein data and provide a discussion and draw conclusions in Section 6.

2 MIXTURE MODEL

Considering the specific case of two views represented by $\boldsymbol{x}$ and $\boldsymbol{y}$ (generalization to multiple (> 2) views is straightforward), our aim is to find latent patterns in the joint distribution $p(\boldsymbol{x}, \boldsymbol{y})$. In our application the items are genes, $\boldsymbol{x}$ and $\boldsymbol{y}$ being numerical vectors describing mRNA and protein profiles. We assume that the two sources arise from margin components, mixture components interpretable as clusters and indexed by j and k respectively. There is no restriction on the particular parametric form of these marginal mixtures. In order to understand the proposed model, it is intuitive to consider the joint distribution $p(j, k)$ which is closely related to the contingency table of the assignments of each gene to j, k. In the model, we assume that this distribution can be decomposed into *blocks* i, each of which is the outer product of block-specific distributions $p(j|i)$ and $p(k|i)$ over the two sets of marginal components. The complete table is hence parameterized as an additive mixture of margin products $\pi_i \, p(j|i) \, p(k|i)$ over top-level blocks i; here π_i are the mixture weights. This part of the model is a matrix factorization similar to latent Dirichlet allocation (LDA; Blei et al, 2003), probabilistic latent semantic allocation (PLSA; Hofmann, 1999) and non-negative matrix factorization (NMF; Lee and Seung, 1999), with two extensions: (1) the margins are not fixed but are part of the latent structure, (2) the number of components is not limited for either i, j, or k: all matrices are of potentially infinite size.

This assumption results in a non-parametric description between the $\boldsymbol{x}$ and $\boldsymbol{y}$. For instance, a single top-level cluster i can associate an arbitrary number of k-clusters to one j-cluster. Still the representation can be compact on the top level in that only a few i-components are active *a posteriori*. The model still does multi-view learning in the traditional and easily interpretable sense shared by canonical correlation analysis, for instance; the two views are conditionally independent given the top-level cluster i, which captures similarities between the sources. In brief, the top-level clusters capture the dependencies

between the sources, whereas each source may have arbitrarily complicated variation within each top-level cluster.

Rogers et al (2008) investigated a different decomposition of the joint distribution. Particularly, $p(j,k) = p(k)p(j|k)$ (where k indexes the mRNA marginal, j the protein). This is a special case of our more general model where each of the K different top level components would link one particular k with all of the js — in other words, each block corresponds to one complete row of the contingency table. Results presented in (Rogers et al, 2008) suggest that more flexibility is required — most k-marginal components become connected to many j-marginal components and vice-versa. Hence, imposing a one-to-many constraint in either direction seems prohibitive. With flexibility comes also need for more reliable inference process, solved here with fully Bayesian treatment.

We use Dirichlet Process (DP) priors for the margin clusters, and the GEM distribution (Johnson et al, 1997, p. 237) for the prior probabilities $\boldsymbol{\pi}$ of the top-level clusters[1]. The full specification for the model is then

$$
\begin{aligned}
G_0^x &\sim \mathrm{DP}(\gamma^x, H^x)\ , & G_0^y &\sim \mathrm{DP}(\gamma^y, H^y)\ ,\\
G_i^x &\sim \mathrm{DP}(\beta^x, G_0^x)\ , & G_i^y &\sim \mathrm{DP}(\beta^y, G_0^y)\ ,\\
\boldsymbol{\pi} &\sim \mathrm{GEM}(\alpha)\ , & z_n &\sim \boldsymbol{\pi}\ ,\\
\theta_n^x &\sim G_{z_n}^x\ , & \theta_n^y &\sim G_{z_n}^y\ ,\\
\boldsymbol{x}_n &\sim f_x(\boldsymbol{x}|\theta_n^x)\ , & \boldsymbol{y}_n &\sim f_y(\boldsymbol{y}|\theta_n^y)\ .
\end{aligned}
$$

Data samples are indexed by n, and the superscripts x and y in general denote marginals. Concentration parameters γ and β are margin-specific, defining the diversity, or "effective number" of the j/k-clusters, and α is the concentration parameter defining the diversity of the top-level clusters over i. Cluster parameters, originating from the base measures (priors) H^x and H^y, are denoted by θ^x and θ^y. Both margins have a hierarchy of DPs (Teh et al, 2006), with the higher-level processes G_0^x and G_0^y, and lower-level processes G_i^x and G_i^y that are specific to the components i. The latent variables z are top-level component identities for the data samples. Finally, f_x and f_y are likelihoods of data, specific to each margin cluster j and k, but in the DP notation parameterized directly by the parameters sampled from the base measures and circulated through the DP hierarchies. A plates diagram depicting the model is in Figure 1. Infinite mixtures are involved in the model three-fold. The top level with its GEM prior is straightforward and separate from the mixtures at the margins. The two margins are again separate. But within a margin, the clusters are shared by top-level components, and a hierarchy is needed to give the clusters common identities over multiple components i. If H was directly used as the base measure of a margin, independently for each component i, the atoms sampled from H would be different for each i with probability one, because the base measure is continuous. G_0^x and G_0^y, on the other hand, provide discrete base measures and in a sense give identities to the margin clusters.

Figure 2 illustrates schematically the prior over $p(j,k)$. The number of rows and columns in the table correspond to the number of marginal components and will increase or decrease according to the DPs over the margins controlled (*a priori*) by the concentration parameters γ^x, γ^y. Within the table, we see a decomposition into blocks[2]. New blocks are produced or removed according to the DP over top-level blocks, controlled by α. Blocks may also grow

[1] GEM is like the DP in providing stick lengths, but without a base measure which is irrelevant here. GEM is named after Griffiths, Engen, and McCloskey.

[2] Note that the ordering of rows and columns of the table is arbitrary, and hence the blocks will not in general be contiguous as in the illustration.

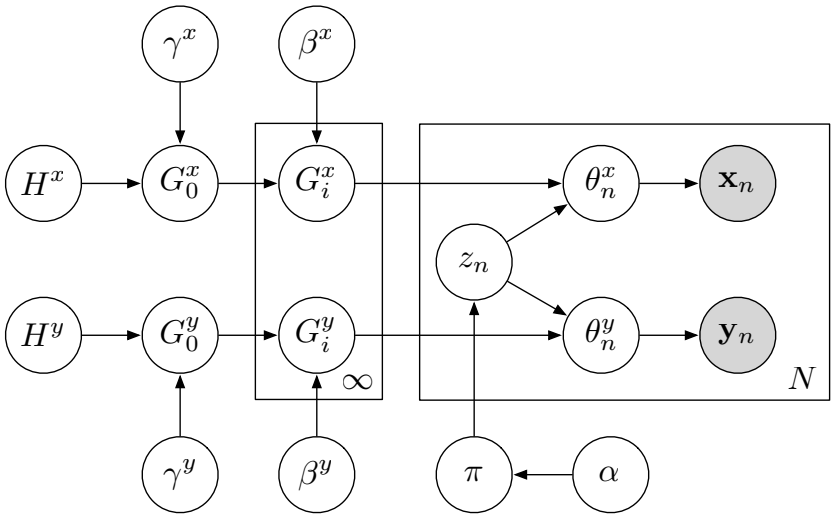


Fig. 1 Plates diagram of the mixture model.

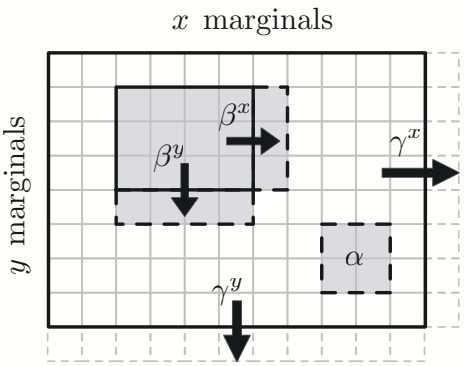


Fig. 2 Schematic representation of the prior over contingency tables. The arrows and dashed lines illustrate the effect of the concentration parameters α, β, and γ.

and shrink as marginal components are added and removed. This process is controlled by β^x, β^y.

It is worth considering the effect of the various concentration parameters on the block structure. The definition of a block is a product of (conditionally, on i) independent distributions over the sets of marginal components. In this sense, one could legitimately split the larger block in Figure 2 into two (or indeed more) components with the same $\boldsymbol{y}$ marginal components and mutually exclusive sets of $\boldsymbol{x}$ marginals (or vice-versa). The decomposition favored by the proposed prior will depend on the concentration parameters α, β and (to a lesser extent) γ. For example, high α and low β encode a preference for a larger number of small blocks whilst small α and high β will prefer a small number of large blocks. Such control is useful when we possess prior knowledge regarding the type of blocks that are of interest. If desired, one can place hyper-priors on these parameters and sample them within the Gibbs scheme. However, it is important to note that the absence of a base measure for α means that the only quantity affecting the posterior sampling is the number of top-level

components (and not how 'good' they are in relation to a base measure, as is the case on the marginals). Hence, the prior specification will have a large influence. We demonstrate these properties in the synthetic examples section.

For concreteness, a finite version of the model would have Dirichlet priors for $p(i)$, $p(j|i)$, and $p(k|i)$. The clusters would be with likelihoods f_x and f_y, and with priors h^x and h^y, the density equivalents of the measures H. In the finite version of the model, the hierarchies for the margins would not be necessary to bind component identities, which are determined by indices. But the infinite version can be obtained as a limit only with hierarchical Dirichlet priors as explained by Teh et al (2006).

2.1 ESTIMATION

Albeit apparent complexity caused by infinite parameterization, DP mixtures can be relatively efficiently estimated by collapsed Gibbs sampling where the processes G have been marginalized out. Blacwell and MacQueen (1973) solved the marginalization task for standard DPs, and Teh et al (2006) explain how the two hierarchical layers of G_0 and G_i can be marginalized out in a similar fashion. Marginalizing out the processes makes it possible to work with conditional distributions that depend directly on the cluster parameters.

We assume base measures conjugate to the mixture component likelihoods, and hence can integrate out the cluster parameters θ associated to the margin clusters. The resulting sampler then operates directly with the cluster likelihoods $p(\boldsymbol{x}|\boldsymbol{X}, \Delta)$, conditioned on the samples $\boldsymbol{X}$ already associated to the clusters, and the hyperparameters Δ. For non-conjugate base measures slightly more advanced sampling techniques would be needed, following for instance the methods presented by Neal (2000) for non-hierarchical DP mixtures.

In general, sampling related to the margin HDPs follows closely the "franchise scheme" (Teh et al, 2006), while conditional probabilities for sampling the top-level components can be obtained by marginalizing over the potential margin assignments. The exact sampling formulas depend on the chosen likelihoods $f_x(\boldsymbol{x}|\theta_n^x)$ and $f_y(\boldsymbol{y}|\theta_n^y)$, which may be different for the two margins. We demonstrate implementations with Gaussian and multinomial likelihoods in the experimental section, but do not write out the conditional distributions $p(\boldsymbol{x}|\boldsymbol{X}, \Delta)$ since they are not specific to our model and are given for instance by Gelman et al (2004) for various distributions.

The model consists of three layers of assignments. First, samples are assigned to top-level components (blocks). Second, within that particular block, groups of one or more samples are assigned to *instances* of marginal components, which are necessary for sharing the marginal components between the top-level clusters. The particular assignment of these instances to marginal components represents the third layer. Note that these instances are the analogue of *tables* in the Chinese restaurant metaphor for hierarchical Dirichlet processes.

The collapsed Gibbs sampler cycles through data n, removing one sample at a time from the "urns" specified below. The assignments are then resampled, first the top-level component i for the sample, then its margin component assignments (j, k). The latter are done in a nested scheme of data-instances and marginal-instance assignments.

In addition to the block assignments z_n of the samples at the top level, the model has two kinds of latent assignments to describe the margin cluster memberships: (1) v_n^x and v_n^y denote how samples are assigned to instances, and (2) w_{it}^x and w_{it}^x tell which marginal component is assigned to each instance t of block i. The margin cluster identities of sample n are then obtained by double indexing: $w_{z_n v_n^x}^x$ and $w_{z_n v_n^y}^y$.

Because the marginal-instance assignments are common to many data objects, they constitute a separate sampling step that we run after going once through all the other assignments. The marginal clusters are drawn from the posterior specified by all of the data points assigned to that particular instance.

All probabilities for the collapsed sampler below are implicitly conditional on data except the left-out sample(s), assignments of samples, and hyper-parameters. That is, if sample n is left out, conditioning is on $(\boldsymbol{X}^{-n}, \boldsymbol{Y}^{-n}, z^{-n}, (v^x)^{-n}, (w^x)^{-n}, (v^y)^{-n}, (w^y)^{-n}, \Delta^x, \Delta^y)$. The counters appearing in the formulas are functions of the latent assignments z, v, and w.

Sampling Block-data Assignments z. The top-level component for a left-out sample is obtained by marginalizing over the potential margin cluster assignments for the sample within each block i:

$$p(z_n = i) \propto \begin{cases} C_i^{-n}\, p(\boldsymbol{x}_n | z_n = i, \Delta^x)\, p(\boldsymbol{y}_n | z_n = i, \Delta^y) & \text{for an existing } i, \\ \alpha\, p(\boldsymbol{x}_n | t^*, \Delta^x)\, p(\boldsymbol{y}_n | u^*, \Delta^y) & \text{for a new } i. \end{cases}$$

We have denoted the number of samples in the block i by C_i, with the superscript $-n$ here and elsewhere denoting the absence of the left-out sample n. Instances on the margin x are in general denoted by t, while y-instances are denoted by u. The likelihood for $\boldsymbol{x}_n$ to be assigned to a new, empty instance t^* is obtained my marginalizing over marginal components, giving

$$p(\boldsymbol{x}_n | t^*, \Delta^x) = \frac{\gamma\, p(\boldsymbol{x}_n | \Delta^x) + \sum_j d_j^{-n} p(\boldsymbol{x}_n | j, \Delta^x)}{\gamma + \sum_j d_j^{-n}}\,, \tag{1}$$

where the d_j count the numbers of samples on the x-margin associated to components j. For $p(z_n = i)$ we also need the margin-specific probabilities for block assignments—for instance for the x-margin,

$$p(\boldsymbol{x}_n | z_n = i, \Delta^x) = \frac{1}{\beta + C_i^{-n}} \beta\, p(\boldsymbol{x}_n | t^*, \Delta^x) + \sum_{t=1}^{T_i} c_{it}^{-n} p(\boldsymbol{x}_n | t_l, \Delta^x)\,,$$

where instances have now been marginalized out, and c_{it} counts samples associated to instance t of block i.

The formulas for $p(\boldsymbol{y}_n | u^*, \Delta^y)$ and $p(\boldsymbol{y}_n | z_n = i, \Delta^y)$ are otherwise identical but with y-specific equivalents of the counters (c, d) and the likelihoods.

Sampling Instance-data Assignments v. As this step and the following are independent and similar for each margin, we only treat the x-margin here, without using the margin superscripts.

On the x-margin, the sample n is assigned to an instance according to

$$p(v_n = t) \propto \begin{cases} c_{it}^{-n}\, p\Big(\boldsymbol{x}_n | \boldsymbol{X}_{it}^{-n}, \Delta\Big) & \text{for an instance } t \text{ in the block } i, \\ \beta\, p(\boldsymbol{x}_n | t^*, \Delta) & \text{for a new instance,} \end{cases}$$

where $p(\boldsymbol{x}_n | t^*, \Delta)$ is the probability of setting up a new instance for the sample, eq. 1. If a new instance was created, a margin cluster needs to be associated to the instance, by drawing from the urn associated to the base DP,

$$p(w_{it^*} = j) \propto \begin{cases} d_j^{-n}\, p\Big(\boldsymbol{x}_n | \boldsymbol{X}_j^{-n}, \Delta\Big) & \text{for an existing component } j, \\ \gamma\, p(\boldsymbol{x}_n | \Delta) & \text{for a new component.} \end{cases}$$

Sampling Instance-marginal Assignments w. All instances are reassigned to components, one by one. For the instance t of block i, the probabilities are

$$p(w_{it} = j) \propto \begin{cases} d_j^{-(it)}\, p\Big(\boldsymbol{X}_{it} | \boldsymbol{X}_j^{-(it)}, \Delta\Big) & \text{for a component } j \text{ in the model,} \\ \gamma\, p(\boldsymbol{X}_{it}|\Delta) & \text{for a new component.} \end{cases}$$

All data previously associated to the instance are denoted by $\boldsymbol{X}_{it}$, and $\boldsymbol{X}_j^{-(it)}$ denotes all data in block i without the data of the instance under reassignment. Note that the conditional probabilities here are exchangeable over permutations of $\boldsymbol{X}_{it}$ and factorize, but the factors are not the probabilities of single samples conditioned on old data. Instead, one needs to sequentially "stack" samples on top of old: For each data point the probability $p(\boldsymbol{x}_n)$ is conditioned on the old data $\boldsymbol{X}_j^{-(it)}$ and all previously assigned samples of this instance.

Hyperparameter Estimation. The model specification includes five DP concentration parameters $\alpha, \beta_x, \beta_y, \gamma_x$, and γ_y, the values of which will determine the readiness with which the model will generate new blocks, instances and marginal components. The observed number of components at each level is rather sensitive to the values of these parameters, especially for the top-level clusters. Bearing this in mind, it is sensible to add an extra level of hierarchy to our model and sample these hyper-parameters along with the various assignments. As Rasmussen (2000), we notice that conditioned on a current set of assignments, conditional distributions for each of these hyper-parameters is only dependent on the number of components and not on the particular distribution of data instances across components. This leads to a likelihood function of the form

$$p(z|\alpha) \propto \frac{\alpha^I \Gamma(\alpha)}{\Gamma(N+\alpha)}$$

where I is the number of top-level components. An identical expression is obtained for γ_x and γ_y with I replaced by K and J respectively. The form for β_x and β_y is slightly different as these parameters tune the number of instances in a particular block and not the total number of instances. Hence, we can think of the I blocks as I independent realisations of the process controlled by β and obtain a likelihood of the form

$$p(v|\beta) \propto \beta^{\sum_i T_i} \left(\frac{\Gamma(\beta_x)}{\Gamma(N+\beta_x)} \right)^I .$$

Alternatively, one could maintain separate β_x for each block which may be useful if it was expected that blocks would be of vastly differing sizes. For the particular combination of Gamma priors and the basic likelihood (that for α and γ), Gibbs sampling is possible through an auxiliary variable method described by West (1992). For other priors, one must resort to a less efficient sampling strategy (for example, Metropolis-Hastings). As noted by Rasmussen (2000), with Inverse-Gamma priors, the posterior is log-concave and adaptive rejection sampling could be used instead.

3 RELATED WORK

3.1 Generative dependency modeling

Skipping all the details of the model structure of Figure 1, it shares basic elements with other generative approaches for modeling dependencies between co-occurring data sources. For

each sample we have a latent variable, here the top-level cluster z_n, that ties the sources together, whereas the rest of the model is conditionally independent given the shared variable. Klami and Kaski (2008) discuss this type of models, stating that they will find dependencies between the views only to the degree permitted by the complexity of the marginal models, conditioned on the shared variable.

Most existing models make rather restrictive assumptions on the marginals. Probabilistic canonical correlation analysis (PCCA) by Bach and Jordan (2005) is based on Gaussian linear marginal models, resulting in a model that seeks pure correlations between the sources. Klami and Kaski (2007, 2008) extend PCCA to clustering-type models, but retain the assumption of unimodal Gaussian marginal within each cluster. The model we propose in this paper, however, uses flexible non-parametric cluster formulation for the marginals within each top-level cluster, making the marginal models capable of capturing in principle any variation specific to each of the sources alone. In this paper we present practical experiments with multinomial and Gaussian DP mixtures as marginal models, but the same principle is expected to apply more generally: Novel dependency-seeking generative models can be developed by plugging in suitable non-parametric marginal models.

In a sense, the model by Klami and Kaski (2008) could be considered as a special case of a finite variant of the proposed model. It would correspond to a variant where each top level cluster only uses one marginal component on both views, resulting in a diagonal contingency table. Which of the methods to prefer is ultimately a question of the application. For data with complex interconnections, like in the data analysed in this paper, diagonal contingency table would not be suitable as illustrated later in Section 5. However, models that assume one-to-one mapping between the marginal clusters are more efficient for data where both views share nearly identical cluster structure. The constraint regularizes the solution, and also makes possible modified learning algorithms that try to enhance the similarity of the marginal clusterings (Bickel and Scheffer, 2004).

3.2 Non-parametric modeling

The proposed model relies heavily on the hierarchical DP of Teh et al (2006), and is consequently related to various extensions and modifications of the model as well. The crucial difference to the alternatives is that they are typically defined for single source only, lacking the multi-view aspect completely. Non-parametric Pachinko allocation by Li et al (2007) presents a similar hierarchy for topic models, also relaxing the assumption of fixed group assignments. Nested DPs (Rodriguez et al, 2008), in turn, are an alternative hierarchical formulation for fixed groupings. Both of these are hence related to one branch of our model, but would not be applicable for solving the task of coupled modeling.

On the other hand, the presented model is a special case of the very general infinite-state Bayes network (ISBN) by Welling et al (2008). ISBN encompasses most HDP-based models in the same fashion as all graphical models based on directed acyclic graphs are special cases of standard Bayes networks. Our paper treats extensively the practical case of coupled clustering of two data sources with complex interconnections between marginal clusters.

Recently Roy and Teh (2009) proposed a multidimensional non-parametric prior process, called Mondrian, for modeling relational data. The process is based on multidimensional stick-breaking, and constructs an infinite factorization of a matrix (in two-dimensional case) into non-overlapping axis-aligned blocks that cover the whole table. The process

shares technical properties with our hierarchical structure that is based on hierarchical stick-breaking, but Mondrian is a prior for tables of fixed size.

3.3 Matrix factorization

The model could be mimicked by first computing marginal clusters and then analyzing the resulting contingency table of sample assignments as a discrete count matrix. In practice, solving the problem in two stages is bound to be suboptimal, but it is worth contrasting our prior process to the alternatives that could be applied in such a two-stage approach. It should still be kept in mind that none of the methods discussed in this subsection would be directly applicable to the kind of data analyzed in this paper. Instead, they merely have close connections with a part of our model.

The proposed HDP prior process factorizes a matrix as a sum of outer products of marginal probability densities. The finite version of the process would be identical to PLSA (Hofmann, 1999), and closely related to more general matrix factorizations such as Latent Dirichlet Allocation (Blei et al, 2003) and NMF (Lee and Seung, 1999), each giving a factorization in terms of non-negative components. Compared to these the main novelty in the proposed model is that neither the number of components in the factorization nor the size of the matrix are fixed.

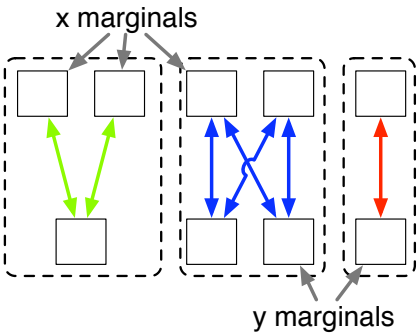
The proposed prior also has inherent sparsity due to top-level clusters only using a subset of marginal clusters, creating a close connection to bi-clustering models, especially to methods like that of Dhillon et al (2003) intended for bi-clustering probability matrices. As a bi-clustering model the prior process is highly flexible, allowing even overlapping blocks and only requiring that the margins of each block are independent.

4 SYNTHETIC EXAMPLES

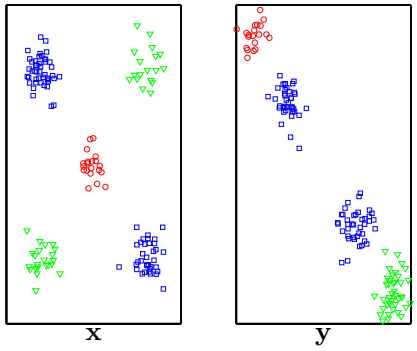
4.1 Gaussian marginals

We begin by demonstrating the performance of the method on a simple synthetic dataset, using Gaussian likelihoods. Figure 3(b) illustrates the data, where the data points are labeled according to their assignment to one of three top-level components, shown in contingency table format in Figure 3(e). A product of independent Normal-inverse-χ^2 priors were used for the marginal base measures (with hyper-parameters $v_0 = 1, \kappa_0 = 1, \mu_0 = [0\ 0]^T, \sigma_0^2 = 1$; see, e.g., Gelman et al., 2004). For this example the concentration parameters $(\alpha, \beta^x, \beta^y, \gamma^x, \gamma^y)$ were given $\mathcal{G}(50, 10)$ prior distributions[3]. Figure 3(c) shows the posterior distribution over the number of top-level blocks and the distribution over the number of top-level blocks that have at least two members. The correct number of components is 3, and we can see that the vast majority of posterior weight over the number of components with 2 or more members is placed at 3 blocks. In addition, we show the autocorrelation for the number of components which suggests the Markov-Chain is mixing reasonably well. Finally, in Figure 3(d), we show the posterior distribution over the number of marginal components for $\mathbf{x}$ and $\mathbf{y}$. We see that for both marginals, the posterior is heavily concentrated on the correct number. Overall, the model reliably extracts the correct generative process.

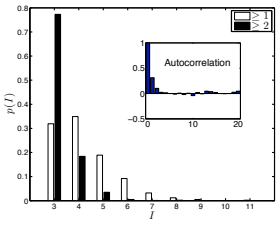
[3] $\mathcal{G}(a, b) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-b/x}$



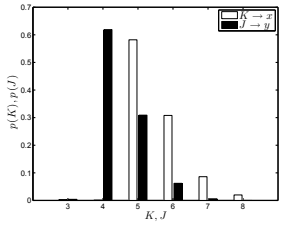
(a) Structure of Gaussian synthetic example. Top boxes represent $\boldsymbol{x}$ marginal components, bottom $\boldsymbol{y}$. Arrows and dashed lines represent the block structure.



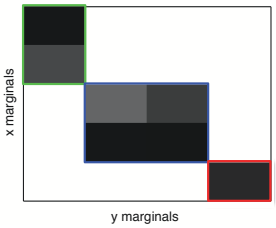
(b) Synthetic dataset for $\mathbf{x}$ (left) and $\mathbf{y}$ (right). Symbols/colors represent top-level clustering.



(c) Marginal posterior distribution over the number of top-level components, I. White bars show the histogram over all clusters, whereas black bars ignore singleton clusters. (Inset: autocorrelation)



(d) Posterior distribution over the number of marginal components. For both margins the mode corresponds with the true solution.



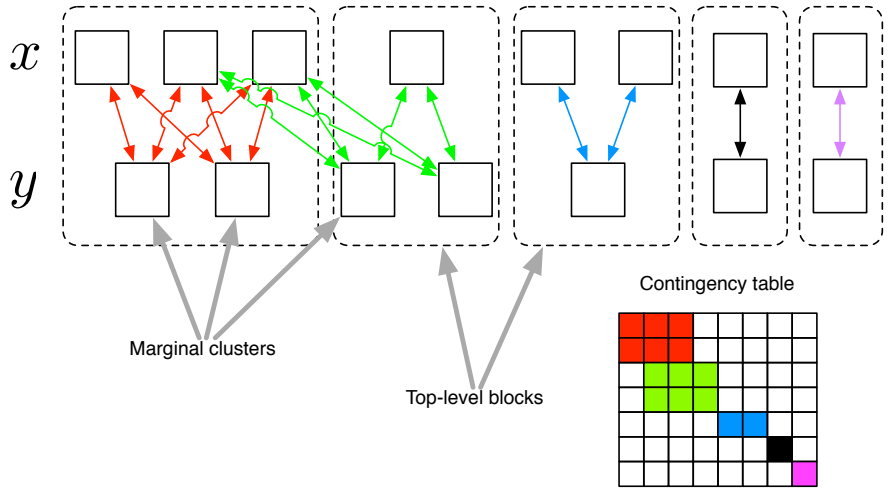
(e) Example contingency table from the sampler. Gray shades denote the count of samples in each cell, and the block structure corresponds exactly to that shown in subfigure (a).

Fig. 3 Model results on synthetic data with Gaussian marginals.

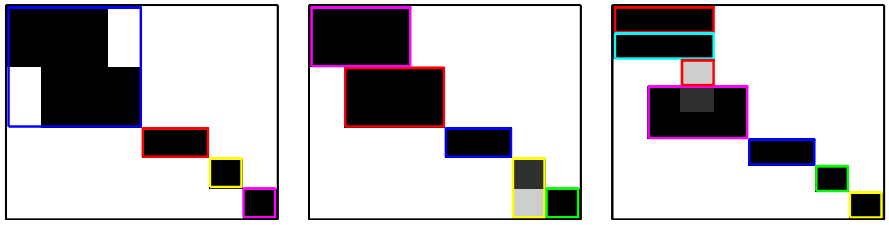
4.2 Multinomial marginals

The method characterizes dependencies between marginal clusterings by extracting block structure in the cross-cluster table. As illustrated in Figure 2 and described in Section 2, the level of detail can be tuned with the prior parameters (α, β, γ). Here we demonstrate the effect of the parameters in practice on synthetic data with clear block structure that still overlaps on the marginals, shown in Fig. 4(a), using multinomial data to show how the general model structure is not tied to the margin likelihoods.

Data in each cluster is drawn from Multinomial(θ,100) with 10-dimensional θ (that is, we get "documents" of 100 "words" taken from a vocabulary of size 10), so that we have 5 independent documents for each link in Figure 4(a). This results in marginal clusters of varying size and a total of 80 data objects. The model is trained with multinomial marginal likelihoods and Dirichlet priors with a count of one for each element. Here we do not infer the concentration parameters here but instead use fixed values to illustrate their effects. In particular, Figure 4(b) shows the effect of tuning α while keeping the other parameters fixed



(a) Illustration of the data generation. The marginals constitute of 8 and 7 clusters, respectively, and the diagram on top shows connections between the marginal clusters. The contingency table on right is an alternative illustration of the generative process, showing the block-structure emerging from the connections between marginal clusters. Each non-empty cell corresponds to 5 data points drawn from multinomial distribution, and the color codes correspond to the links in the top diagram.



(b) Concentration parameter α controls the number of top-level clusters. Very small value (left; $\alpha = 0.01$) results in blocks that are not truly independent, whereas larger values give solutions consistent with the data generation process with different degrees of detail. Here $\alpha = 0.3$ (middle) results in blocks of maximal size, whereas $\alpha = 10$ (right) splits the blocks into smaller parts. The extra y-cluster between the two large blocks arises because relatively large γ_y favors having more marginal clusters. Each contingency table corresponds to one posterior sample, with gray shade denoting the amount of data in a cell and borders (with arbitrary color coding) indicating the top-level clusters.

Fig. 4 Synthetic multinomial data for demonstrating the effect of top-level concentration parameters α.

to values $\gamma_x = \gamma_y = 3$ and $\beta_x = \beta_y = 1$. For very small values of α the model does not find the true block structure, but instead groups several blocks together. For larger values the model finds block-structures of increasing complexity. In practice, we recover the true block structure with wide range of parameter values, and even at extreme parameter values the result is informative of the underlying structure.

It is worth keeping in mind that even though we here illustrate only the subdivision of the contingency table, which could be achieved with various other methods as well, our model operates directly on the marginal multinomial measurements and constructs the contingency table as part of the analysis. Standard matrix factorization or bi-clustering methods used for

analysing the table would not be applicable without separate preprocessing step clustering the marginals independently, losing the advantage of the coupling at that stage.

5 ANALYSIS OF mRNA AND PROTEIN TIME-SERIES

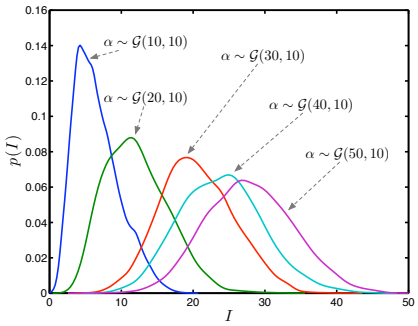
We now turn our attention to the analysis of coupled mRNA and protein time-series datasets. The data was originally described in (Waters et al, 2008), has been previously analysed in (Rogers et al, 2008) and consists of mRNA and protein time-series for a total of 542 genes measured from human breast epithelial cell line. Measurements were taken from the same population cells at 8 unevenly spaced time-points between $t = 0$ and $t = 24$ hours. As in (Rogers et al, 2008), data were normalised by dividing by the value at $t = 0$ (and hence this time-point was discarded) and then normalised so that each representation of each gene had zero mean and unit standard deviation over time. Additionally, one mRNA time point (15 minutes) was rejected from the analysis as it didn't pass the necessary quality controls. Therefore, we were left with 6 mRNA and 7 protein time-points. Genes were tagged with gene ontology (GO) terms to enable us to objectively determine the biological significance of the groupings produced by the model. Terms were removed if they were tagged to fewer than 5 genes.

5.1 Base measures and hyper-priors

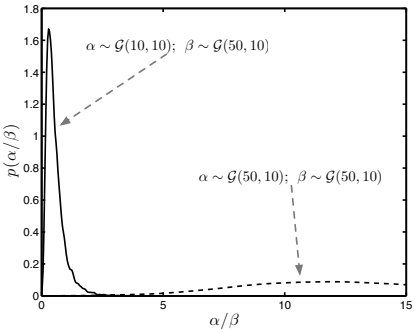
A product of independent, univariate Normal-inverse-χ^2 priors was used for the marginal base measure (with hyper-parameters $v_0 = 1, \kappa_0 = 1, \mu_0 = 0, \sigma_0^2 = 1$; see, e.g., Gelman et al., 2004). The use of conjugate priors allows us to marginalize over the marginal cluster parameters which is known to help mixing and convergence in DP models (Neal, 2000). As discussed in previous sections, the choice of values for the various concentration parameters, particularly α, is rather important. We demonstrate the effect this has by fixing the priors on γ and β and varying the prior on α. For this data, we found that reasonably informative priors were most effective. Therefore, we placed $\mathcal{G}(50, 10)$ priors on β and γ and then placed a $\mathcal{G}(a, 10)$ prior on α and varied the location parameter, a, from 10 to 50. Figure 5(a) shows how the posterior distribution over the number of top-level components changes as a is varied. As we might expect, the number of blocks increases with a. We can also see this effect if we look at the posterior distribution of α/β (in this case, β for the mRNA side) shown in Figure 5(b) for $a = 10$ and $a = 50$. For $a = 10$, we see that $\alpha < \beta$. Hence the model will be *a priori* more likely to grow current blocks than create new ones. In the other extreme, $\alpha > \beta$, the model prefers making new components to expanding existing ones. In the remaining analysis, we will investigate the results created with $a = 50$. There are two reasons for this. First, results presented in (Rogers et al, 2008) suggest that there is high connectivity between the components, suggesting that there might be a reasonably large number of blocks. Second, we can compute the number of enriched GO terms from our sampler (details below) and find that the number ($p < 0.1$) for $a = 50$ (~ 60 terms) is considerably higher than that for $a = 10$ (~ 10 terms).

5.2 Model complexity

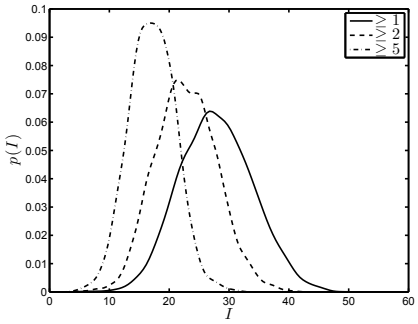
In Figure 5(c) we show the posterior distribution over the number of blocks of varying sizes. The solid line shows all blocks (i.e., those of size ≥ 1). The dashed line shows those with at



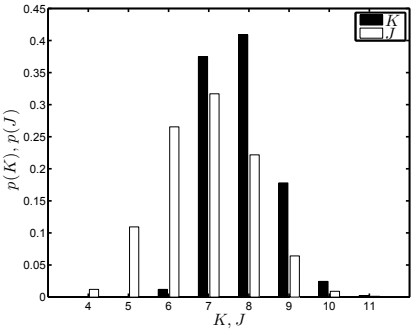
(a) Effect of varying prior over α on posterior over number of top-level blocks.



(b) α/β



(c) Posterior distribution over the number of top-level blocks (solid) with $\alpha \sim \mathcal{G}(50, 10)$; Posterior distribution over the number of top-level blocks with at least 2 members (dashed); At least 5 members (dash-dot).



(d) Posterior distribution over the number of marginal components for the mRNA marginal (K) and the protein marginal (J).

Fig. 5 Posterior distributions over model complexity.

least 2 members and the dash-dot line shows those with at least 5 members. This tells us two things. First, as in all DP models, the posterior is biased a little by singleton components that are unlikely to be stable; they are produced and immediately destroyed. Second, the increase in the number of components as a is increased cannot only be due to the production of singleton components; the curve for ≥ 5 is significantly higher than the complete posterior (i.e. ≥ 1) when $a = 10$ (c.f. Figure 5(c)). Therefore, we can be reasonably confident that the blocks that are being found are of a sensible size and are quite stable. Finally, in Figure 5(d) we see the posterior distribution over the number of marginal components. We see fewer components than used by Rogers et al (2008), however, inspection of Bayesian Information Criteria (BIC) plots in the supplementary material of Rogers et al (2008) suggests that the mode of the posterior here (7/8 components) corresponds to a BIC score very close to the optimal value. In addition, the Gaussians of Rogers et al (2008) were constrained to be spherical whilst in the current model, they are axis-aligned but with different variance parameters for each dimension (time-point) — our parameterisation is more flexible. In light of these observations, the difference is not surprising.

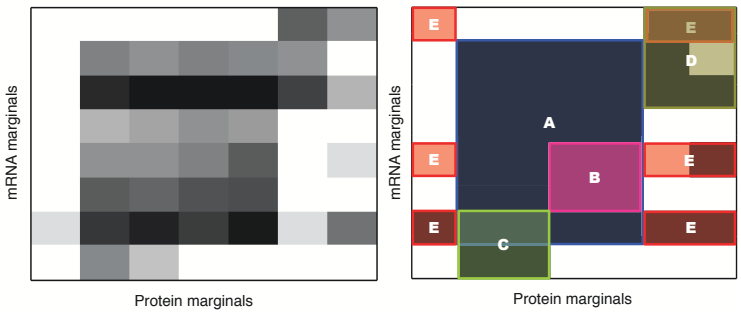


Fig. 6 Example contingency table - cell counts in grayscale (left) and block structure (right).

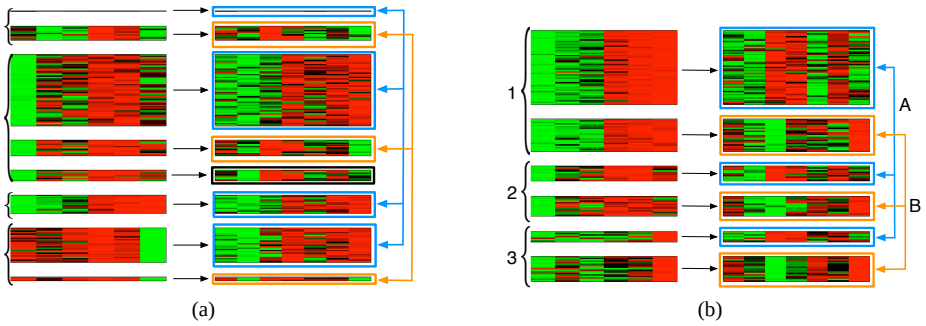


Fig. 7 Examples of two stable blocks from sampler. In both cases, mRNA marginals are given on the left with curly brackets denoting the different marginal components. Protein marginals are given on the right and marginal components are denoted by boxes and lines. The genes are presented in the same order for both mRNA and protein marginals.

5.3 Visualising the contingency table

As the sampler explores the posterior, the size of the contingency table and number of blocks is continually varying. For this reason, visualising the contingency table for this application is rather difficult. However, for completeness, we present the contingency table corresponding to one randomly chosen posterior sample in Figure 6. Most cells have non-zero gene count, showing complex connectivity between marginal clusters as noticed already by Rogers et al (2008) and demonstrating how multi-view clustering methods assuming unimodal marginal variation within clusters would not be applicable here. We also illustrate how the model subdivides the contingency table into blocks having independent margins. Five clusters with the largest membership are labeled from A to E, covering already the main characteristics of the contingency table and being partially overlapping. Note that the six separate parts of cluster E would be merged into a visual block in a different row/column ordering. In general, it is not possible to visualize the structure so that all top-level clusters would form contiguous blocks.

5.4 Gene ontology enrichments

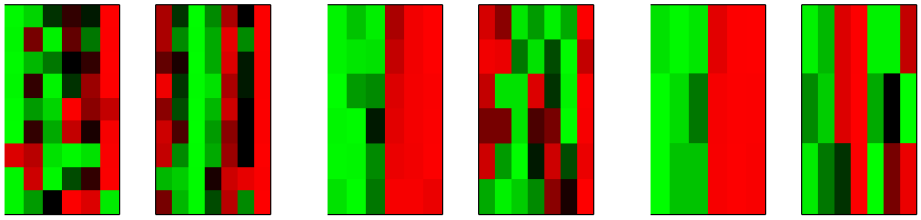
It is standard practice when clustering genomic data to mine the clustering for enriched (and depleted) Gene Ontology (GO) terms. Given a single partitioning of the genes, this is straightforward. However, the Gibbs sampler produces many samples from the posterior distribution over clusterings and it is less obvious how to mine for enrichments. Here we use two different methods. The first method involves exploring the posterior samples for individual, stable blocks. Particularly, if we take blocks that survive for a reasonable number of sampler iterations, we can find the subset of marginal components and genes that are consistently assigned to this block. In Figure 7(a) we see one such example. The mRNA profiles in the left plot come from four marginal components (denoted by the curly brackets). The protein profiles come from three marginal components (color-coded and connected by arrows). These are the most regularly occuring genes/marginal components in this block, which persisted for ~ 700 posterior samples. Examining the figure, we see that three of the mRNA marginals (left) interact with both of the large protein components creating a fully connected 3×2 block in the contingency table. Additionally, we can mine these genes for enriched GO terms. We find 21 terms with $p < 0.1$. Analysing the location within the block of genes tagged with these terms, we find that all terms but one have representatives in more than one component in at least one marginal and 13 of the terms have representatives in more than one component on *both* sides. This strongly supports the claim that blocks can represent meaningful biological structure.

A second example is given in Figure 7(b). In this case, we have three marginal mRNA components and 2 marginal protein components. This configuration corresponds to a 2×3 block in the contingency table. Again, we can mine these genes for enriched GO terms and find 19 terms with $p < 0.1$. Of these terms, 1 only has representatives in one marginal on each side. The remaining 18 are present in at least 2 components on one side or the other and 8 of these have members in more than one marginal component on *both* sides. Two interesting examples are GO:0003735 tagged to 9 genes in the block in two mRNA and one protein marginal, and GO:0006412 tagged to 18 genes and present in all three mRNA marginals and both protein marginals. The reason that these two are of particular interest is that they are related. GO:0006412 corresponds to the process of translation whilst GO:0003735 is tagged to genes whose product make up the ribosome, a large protein complex involved in translation. Hence, all genes tagged with GO:0003735 are involved in translation and are also tagged with GO:0006412 whilst the reverse does not necessarily apply (there are genes involved in translation that do not make up the ribosome). It is extremely encouraging that we see genes tagged with GO:0003735 in a subset of the marginal components of those tagged with GO:0006412. The implication is that through our model, we are able to see variations within particular biological processes (in this case, translation). Specifically, ribosomal genes are restricted to mRNA components 2 and 3 (see Figure 7(b)) and protein component B whereas in general, translation genes appear in all marginals. Inspection of the marginals shows that whilst the mRNA levels follow similar profiles across the 3 marginal clusters (start low, finish high, albeit with some cluster-specific variation), the protein profiles are very different, possibly suggesting overall transcriptional control with specific behavior being controlled at the post-transcriptional level. Further biological investigation into these blocks and the many others produced by the model is an area of ongoing research. The discovery of such blocks is a direct consequence of the new model and the factorisation of the contingency table. The method proposed in Rogers et al (2008) would not be capable of discovering flexible blocks potentially with representatives from more than one component in both marginals.

The second enrichment analysis method averages the enrichments over all posterior samples: for each gene-term pair, we compute the probability of enrichment in the top level blocks and marginal components. These enrichments can then be averaged across samples to give a measure of how significant this term is for this gene in this dataset (conditioned on the model) rather than how significant it is in any particular partitioning. Using this measure, and computing enrichment using the one-sided mid-P-value of the hyper-geometric distribution described in (Rivals et al, 2007) we find (at $p \leq 0.1$) 65, 352 and 430 significant gene-term pairs for the top level blocks, mRNA marginal, and protein marginal, respectively. One drawback of this approach when compared to the previous one is that as we are averaging over all samples from the posterior, it is not possible to break the genes up by their marginal components. In Figure 8 we show 3 examples of terms significant in the top-level blocks and in Figure 9 two examples of terms significant in the mRNA and not in the protein marginals and one example that is significant in the protein and not the mRNA marginal. Each pair of plots has the mRNA data on the left and protein on the right. Rows correspond to genes and columns to time-points. Of the example terms significant in the blocks, we see GO:0003735 (the ribosome, discussed previously) and GO:0000502, the proteosome. As both of these are protein complexes requiring all of their constituent parts to be present, it is not surprising that they appear to be tightly regulated with homogeneous mRNA and protein profiles (not withstanding the observations regarding the ribosome and translation in the previous section). The third, GO:0008380 is related to mRNA processing and may be an interesting group of genes for further analysis. The three terms in Figure 9 correspond to DNA repair (GO:0006281), protein folding (GO:0006457) and cell adhesion (GO:0007155) and in each case we can see considerable diversity in the marginal for which the term is not significant (protein in (a) and (b) and mRNA in (c)). Re-assuringly, both GO:0003735 and GO:0007155 are discovered to be significant and discussed by Rogers et al (2008) as well. The proteosome was not mentioned by Rogers et al (2008) — it is possible that its small size made it hard to extract from the connectivity probabilities although further validation would be required to test this hypothesis. It is clear from these plots how this technique can provide insight into regulatory mechanisms. For example, in Figure 8, we see genes with predominately homogeneous mRNA and protein profiles. This suggests tight co-regulation at both the transcriptional and post-transcriptional stages. Conversely, in Figure 9 we see examples with tight mRNA co-expression and varied protein expression ((a) and (b)) suggesting co-regulation at the transcriptional level but different control at later stages and co-regulation at the protein level (c) but diverse mRNA profiles pointing towards genes that are differently regulated at the transcriptional level but exposed to some post-transcriptional control. Whilst the biological conclusions drawn here are similar to those by Rogers et al (2008), it is important to remember that these small modules are *automatically* exposed through the block decomposition of the contingency table.

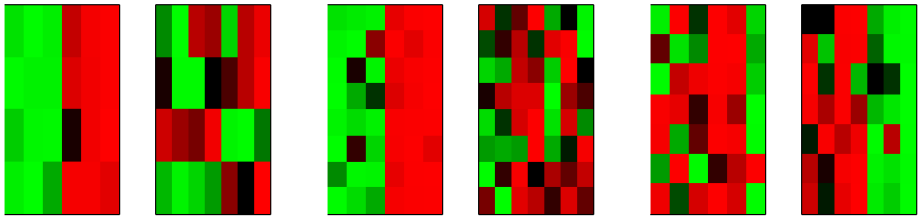
6 DISCUSSION

We have introduced a hierarchical non-parametric model for multi-view data inspired by a new biological dataset. The model couples two, or more, hierarchical DPs, has a potentially infinite number of subgroups for data, and does not assume the group assignments of data to be known *a priori*, as in the original HDP. The motivation behind the model is similar to that by Rogers et al (2008). However, it differs in three important respects. First, it explicitly attempts to find structure in the joint distribution/contingency table of the marginal components. Second, model complexity is automatically inferred from the data and finally a Gibbs



(a) GO:0003735 (b) GO:0008380 (c) GO:0000502

Fig. 8 3 examples of gene ontology terms significantly enriched in top level components. In all cases, left heat map is mRNA data, right is protein data.



(a) GO:0006281 (b) GO:0006457 (c) GO:0007155

Fig. 9 3 examples of gene ontology terms significantly enriched in one marginal component (mRNA for (a) and (b), protein for (c)) and not the other. (Left - mRNA, right - protein)

sampling scheme is presented rather than the maximum likelihood approach previously proposed. In summary, the model explores the similarities and differences between the views whilst permitting complex structure in the individual views. The analysis is performed in the original space, making the results readily interpretable.

In a more general sense, the latent structure is also a factorization of an infinite joint probability matrix, and a prior for contingency tables of potentially infinite dimension. As any margin cluster likelihood can be plugged in and we are not restricted to using the same likelihood for each marginal, the framework is quite general and not just applicable to datasets defined in the real space, such as the Omics datasets of molecular biology. It could just as readily be used in, for example, domains consisting of text, images, strings (e.g., DNA sequences) or combinations thereof.

We have demonstrated the model on a dataset consisting of time-series mRNA and proteomic profiles for ~ 500 human genes, previously analysed by Rogers et al (2008), where it appears that the model is able to extract interesting regulatory effects at both the transcriptional and translational levels. At a more abstract level, the small number of marginal components compared to the large number of top-level components provides some evidence that the relationship between the two representations is a complex one. Further analysis of the biological results of this dataset is an area of ongoing research. The results are broadly comparable with those obtained by Rogers et al (2008), with the added benefits of inferring the number of components at each level, and automatically extracting useful relationships from the contingency table through the top-level components.

6.0.1 Acknowledgements

SR and MG are supported by EPSRC grants EP/C010620/1 and EP/E052029/1 respectively. This work was made possible by funding on the PASCAL2 short visit programme. JS is supported by Academy of Finland (AoF), grant number 119342. SK and AK belong to the Finnish CoE on Adaptive Informatics Research of AoF, and to the Helsinki Institute for Information Technology HIIT, and are partially supported by PASCAL2.

References

Bach FR, Jordan MI (2005) A probabilistic interpretation of canonical correlation analysis. Tech. Rep. 688, Department of Statistics, University of California, Berkeley

Becker S, Hinton GE (1992) Self-organizing neural network that discovers surfaces in random-dot stereograms. Nature 355:161–163

Bickel S, Scheffer T (2004) Multi-view clustering. In: Proceedings of the IEEE International Conference on Data Mining, IEEE, pp 19–26

Blacwell D, MacQueen JB (1973) Ferguson distributions via Polya urn schemes. The Annals of Statistics 1(2):353–355

Blei D, Ng A, Jordan M, Lafferty J (2003) Latent Dirichlet allocation. Journal of Machine Learning Research 3:993–1022

Dhillon IS, Mallela S, Modha DS (2003) Information-theoretic co-clustering. In: Proceedings of KDD'03, the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM Press, New York, NY, pp 89–98

Englebienne G, Cootes T, Rattray M (2008) A probabilistic model for generating realistic lip movements from speech. In: Platt J, Koller D, Singer Y, Roweis S (eds) Advances in Neural Information Processing Systems 20, MIT Press, Cambridge, MA

Friedman N, Linial M, Nachman I, Pe'er D (2000) Using bayesian networks to analyze expression data. In: RECOMB '00: Proceedings of the fourth annual international conference on Computational molecular biology, ACM, New York, NY, pp 127–135, DOI http://doi.acm.org/10.1145/332306.332355

Gelman A, Carlin J, Stern H, Rubin D (2004) Bayesian Data Analysis, 2nd edn. Chapman and Hall

Hofmann T (1999) Probabilistic latent semantic analysis. In: Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann Publishers, San Francisco, CA, pp 289–296

Johnson NL, Kotz S, Balakrishnan N (1997) Discrete multivariate distributions. John Wiley & Sons, New York, NY

Klami A, Kaski S (2007) Local dependent components. In: Ghahramani Z (ed) Proceedings of ICML 2007, the 24th International Conference on Machine Learning, Omnipress, pp 425–432

Klami A, Kaski S (2008) Probabilistic approach to detecting dependencies between data sets. Neurocomputing 72:39–46, DOI doi:10.1016/j.neucom.2007.12.044

Lee D, Seung H (1999) Learning the parts of objects by non-negative matrix factorization. Nature 401:788–791

Li W, Blei D, McCallum A (2007) Nonparametric Bayes Pachinko allocation. In: Proceedings of the 23rd conference on Uncertainty in Artificial Intelligence, AUAI Press

Neal RM (2000) Markov chain sampling methods for Dirichlet process mixture models. Journal of Computational and Graphical statistics 9(2):249–265

Rasmussen C (2000) The infinite Gaussian mixture model. In: Solla SA, Leen TK, Muller KR (eds) Advances in Neural Information Processing Systems 12, MIT Press, Cambridge, MA, pp 554–560

Rivals I, Personnaz L, Taing L, Potier MC (2007) Enrichment or depletion of a GO category within a class of genes: which test? Bioinformatics 23(4):401–407

Rodriguez A, Dunson DB, Gelfand AE (2008) The nested Dirichlet process. Journal of the American Statistical Association 103(483):1131–1154

Rogers S, Girolami M, Kolch W, Waters KM, Liu T, Thrall B, Wiley HS (2008) Investigating the correspondence between transcriptomic and proteomic expression profiles using coupled cluster models. Bioinformatics 24(24):2894–2900, DOI 10.1093/bioinformatics/btn553

Roy DM, Teh YW (2009) The Mondrian process. In: Koller D, Schuurmans D, Bengio Y, Bottou L (eds) Advances in Neural Information Processing Systems 21, MIT Press, Cambridge, MA, pp 1377–1384

Teh YW, Jordan MI, Beal MJ, Blei DM (2006) Hierarchical Dirichlet processes. Journal of the American Statistical Association 101:1566–1581

Vinokourov A, Hardoon DR, Shawe-taylor J (2003a) Learning the semantics of multimedia content with application to web image retrieval and classification. In: In Proceedings of Fourth International Symposium on Independent Component Analysis and Blind Source Separation, pp 697–701

Vinokourov A, Shawe-Taylor J, Cristianini N (2003b) Inferring a semantic representation of text via cross-language correlation analysis. In: S Becker ST, Obermayer K (eds) Advances in Neural Information Processing Systems 15, MIT Press, Cambridge, MA, pp 1473–1480

Waters K, Liu T, Quesnberry R, Qian W, Willse A, Bandyopadhyay S, Kathmann L, Weber T, Smith R, Wiley H, Thrall B (2008) Systems analysis of response of human mammary epithelial cells to egf by integration of gene expression and proteomic data. Under Submission

Welling M, Porteous I, Bart E (2008) Infinite state Bayesian networks. In: Platt J, Koller D, Singer Y, Roweis S (eds) Advances in Neural Information Processing Systems 21, MIT Press, Cambridge, MA

West M (1992) Hyperparameter estimation in Dirichlet process mixtures. Tech. Rep. 92-A03, Duke University, Institute of Statistics and Decision Sciences