Mach Learn (2010) 79: 201–226
DOI 10.1007/s10994-009-5155-1

Infinite factorization of multiple non-parametric views

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

Received: 27 February 2009 / Revised: 9 October 2009 / Accepted: 12 October 2009 /
Published online: 13 November 2009

Abstract Combined analysis of multiple data sources has increasing application interest, in particular for distinguishing shared and source-specific aspects. We extend this rationale to the 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 sources. The commonalities between the sources are modeled by an infinite component 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. We discover complex relationships between the marginals (that are multimodal in both marginals) that would remain undetected by simpler models. Cluster analysis of co-expression is a standard method of screening for co-regulation, and the two-view analysis extends the approach to distinguishing between pre- and post-translational regulation.

Editors: Nicolo Cesa-Bianchi, David R. Hardoon, and Gayle Leen.

S. Rogers · M. Girolami
Department of Computing Science, University of Glasgow, Glasgow, UK

S. Rogers
e-mail: rogers@dcs.gla.ac.uk

M. Girolami
e-mail: girolami@dcs.gla.ac.uk

A. Klami (✉) · J. Sinkkonen · S. Kaski
Department of Information and Computer Science, Helsinki University of Technology, Espoo, Finland
e-mail: arto.klami@tkk.fi

J. Sinkkonen
e-mail: janne.sinkkonen@tkk.fi

S. Kaski
e-mail: samuel.kaski@tkk.fi

Keywords Hierarchical Dirichlet process · Multi-view learning · Contingency table · Protein 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), analysing webpage content and link information (Cohn and Hoffman 2001) and combining mRNA and protein profiles to explore the complex regulatory behavior that underpins a large amount 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 similar aims—to investigate not just the details of the individual sources, but also their commonalities.

Unsupervised analysis of coupled sources is often based on a simple latent variable structure: The data are modeled as having a single latent source shared by the views, and the variation within the views is independent given the latent source. Probabilistic canonical correlation analysis (CCA; Bach and Jordan 2005) implements the idea for linear projections, assuming multivariate normal noise for the marginals, and Klami and Kaski (2008) present a clustering model with the same assumption on conditional marginals. Barnard et al. (2003) and Blei and Jordan (2003) present examples of more complex models following the same principle, using latent Dirichlet allocation (LDA; Blei et al. 2003) to model captions of associated images conditional on a joint latent clustering of the images and captions.

We are particularly interested in exploring the dependencies between views with complex interconnections, and propose a clustering model that relaxes the frequent assumption of unimodal conditional marginal distributions (Bach and Jordan 2005; Barnard et al. 2003; Klami and Kaski 2007; Klami and Kaski 2008). Instead, each cluster is allowed to have multimodal variation within each view, independent of the other view. In brief, the learning problem is to cluster two data sources $\boldsymbol{x}$ and $\boldsymbol{y}$ jointly, in a way that we reveal both the marginal cluster structure in the sources as well as connections between the marginal clusters.

Clustering two data spaces induces a *contingency table* of assignments to the clusters on the two marginals. This table can be mined for dependencies between the marginal clusters. For example, samples in a single cluster on one marginal can be evenly distributed into two separate clusters on the other marginal. While this kind of data could be modeled also with a simple mixture on the concatenation of the two sources, resulting in two separate clusters, the generative process is better captured by modeling it as a single component that has a bimodal distribution in the latter source. This solution is additionally more scalable.

This is a particularly appealing formulation for the analysis of gene regulation data that we will present in Sect. 5. Genetic regulation is the process that controls the production of proteins from genes encoded by DNA. The process can be controlled at any one (or more) of its several stages. For example, the first stage, *transcription*, involves the production of an mRNA molecule and is initiated by the binding of a special protein—a transcription factor—to the DNA near the gene. Hence, this stage is controlled by the presence or absence of the transcription factor protein. In this manner, a single transcription factor protein can control

several genes and a common use of cluster analysis in bioinformatics is to find groups of genes whose mRNA levels are similar (over time or in different conditions) in the hope that they share transcription factors. The next stage involves the translation of mRNA molecules into proteins. Gene-specific control can be exhibited at this stage too—mRNA molecules could, for example, be systematically destroyed to prevent protein production. If we consider mRNA and protein levels as two views of the same set of genes, clustering the genes in the two spaces could help discover which genes (and hence which biological processes) are controlled at the different levels—the principal goal of research into genetic regulation. For example, a group of genes that have a unimodal distribution (i.e. fall into one cluster) in the mRNA space but reside in, say, 2 clusters in the protein space may all be transcriptionally controlled by the same transcription factors *but* be post-transcriptionally controlled by two different, more specific, processes.

In Rogers et al. (2008), a coupled mixture model was presented that implicitly decomposed the contingency table into *rows*—i.e., components were forced to be uni-modal in one source. Here, we generalize the idea further to multimodal distributions on both sides: We propose a model that describes the generative process of two (or in general more) coupled views through such components. The model is implemented as a two-level hierarchy of mixtures. The top-level mixture represents the common variation, 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. From a biological perspective, our principal aim in developing this model is to investigate whether complex (i.e. multi-modal in at least one marginal) top-level components exist and, if they do, whether they have biological significance.

The remainder of the paper is organised as follows. In Sect. 2 we present the mixture model and describe the Gibbs sampling scheme for inference. In Sect. 3 we describe related work and clarify the differences between the proposed model and that recently presented by Rogers et al. (2008), and in Sect. 4 we illustrate the model on two synthetic datasets. In Sect. 5, we present an analysis of the mRNA and protein data and provide a discussion, and finally draw conclusions in Sect. 6.

2 Mixture model

In the following description we will restrict ourselves to data with two marginals, represented by $\boldsymbol{x}$ and $\boldsymbol{y}$, although generalization to > 2 marginals is straightforward. Our aim is to find latent patterns in the joint distribution $p(\boldsymbol{x}, \boldsymbol{y})$, and in our particular application the items are genes with $\boldsymbol{x}$ and $\boldsymbol{y}$ being numerical vectors describing mRNA and protein profiles (over time) for those genes.

We assume that the two sources arise from marginal models having cluster structure, and index these clusters (or mixture components) by j and k respectively. Furthermore, each data point belongs to a top-level component consisting of a number of marginal clusters in both sources. The task is to find a hierarchical representation such that the marginal clusters accurately model the data whilst the top-level components capture processes shared by the sources. We will first describe the fundamental modeling assumptions through a finite variant of the model, and then proceed to the details of the full HDP-based model.

The clustering model corresponds to the generative process

$$p(\boldsymbol{x}, \boldsymbol{y}) = \sum_k \sum_j p(j,k)\, p(\boldsymbol{x}|k)\, p(\boldsymbol{y}|j), \tag{1}$$

where $p(j,k)$ is the joint prior over the marginal clusters and $p(\boldsymbol{x}|k)$ and $p(\boldsymbol{y}|j)$ denote the cluster-conditioned marginal distributions, with any suitable parametric form. Various modeling assumptions can be encoded in the prior $p(j,k)$, also interpretable as a contingency table over the cross-cluster assignments.

Recently, Rogers et al. (2008) proposed a model where the prior was decomposed as $p(j,k) = p(k)p(j|k)$, effectively breaking the table up into components consisting of only one row (or column, depending on which way around it is drawn), as depicted in Fig. 2(a). Another alternative would be a completely free parameterization of the whole matrix. The model of (Rogers et al. 2008) is described more thoroughly in Sect. 3.4.

In the proposed model, we assume that this distribution can be decomposed into components i, each of which is the outer product of component-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 marginal products $\pi_i\, p(j|i)\, p(k|i)$ over top-level components i with mixture weights π_i (depicted in Fig. 2(b)). 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 marginals are not fixed but are part of the latent structure, (2) the number of components is not limited for either i, j, or k: in the final model all matrices are of potentially infinite size.

The advantage of the hierarchical representation is that it allows us to discover complex relationships between the cluster structures in $\boldsymbol{x}$ and $\boldsymbol{y}$. For instance, a single top-level cluster i can associate an arbitrary number of k-clusters to an arbitrary number of j-clusters, and the number of i-components active *a posteriori* is inferred. The hierarchical model also leads to inherent sparsity as typically only a subset of j,k pairs will have non-zero probability, which is an obvious advantage over completely free parameterization of $p(j,k)$. Despite the complex structure, 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.

2.1 Dirichlet process formulation

The above description of the model assumes finite collections of both top-level and marginal clusters. The need to choose three different cardinalities means that using traditional model complexity selection methods, such as cross-validation to estimate the predictive performance, is likely to be very difficult. To overcome this problem we re-formulate the model as a hierarchical Dirichlet Process (HDP; Teh et al. 2006), which enables learning the cluster cardinalities from data.

To introduce the HDP-model, we draw here a parallel between each component in (1) and the infinite model. First, the marginal clustering models are modeled as Dirichlet Process (DP) mixtures. A DP is a distribution over probability distributions, and hence $G \sim \mathrm{DP}(\gamma, H)$, where γ is a concentration parameter (controlling the expected number of

clusters) and H is a base measure, is itself a distribution. A DP mixture is obtained by coupling the DP prior with observation likelihood $f(\boldsymbol{x}|\theta)$ (analogous to $p(\boldsymbol{x}|k)$), such that $\theta_n \sim G$ are treated as the parameters for the particular sample $\boldsymbol{x}_n$. The draws from a DP are discrete, which results in finite probability for any two samples to have exactly identical θ. The collection of samples sharing the same parameters can then be interpreted as a cluster.

The top-level cluster probabilities $\boldsymbol{\pi}$, in turn, are modeled with the GEM distribution (Johnson et al. 1997),[1] which acts as the infinite analogue of Dirichlet distribution. GEM can be described through the stick-breaking construction

$$\beta_k \sim \text{Beta}(1, \alpha),$$
$$\pi_k \sim \beta_k \prod_{l=1}^{k-1} (1 - \beta_l),$$

where $\pi \sim GEM(\alpha)$.

Finally, the full model requires two nested levels of DP priors in order to tie the identities of cluster components of each top-level cluster to each other, as explained by Teh et al. (2006). In brief, the base measure for the marginal clusters within one top-level cluster is itself a draw from a DP that is common for all clusters. The full specification of the model is then

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

Data samples are indexed by n, and the superscripts x and y in general denote marginals. Concentration parameters γ and β are marginal-specific, with γ defining the diversity, or "effective number" of marginal clusters and β defining the size of the top-level components. The concentration parameter α defines the diversity of the top-level clusters over i. Cluster components are parameterised by θ^x and θ^y, originating from the base measures (priors) H^x and H^y. Both marginals have a hierarchy of DPs (Teh et al. 2006), with higher-level Dirichlet processes G_0^x and G_0^y, and lower-level processes G_i^x and G_i^y that are specific to the ith component. In other words, there is a (potentially infinite) set of clusters/components in each marginal (G_0^x and G_0^y) of which a selection are chosen by each G_i^x (and G_i^y). The selections made by each top-level component (G_i^x) do not need to be disjoint and hence components can be *shared* by many different top-level components.

The latent variables z are top-level component identities for the data samples—i.e. z_n holds the assignment for (x_n, y_n). Finally, f^x and f^y are likelihoods of data, specific to each marginal cluster j and k, and are parameterised by the parameters sampled from the base measures and then selected by the top-level components. A plates diagram depicting the model is in Fig. 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 marginals. The two marginals are

[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.

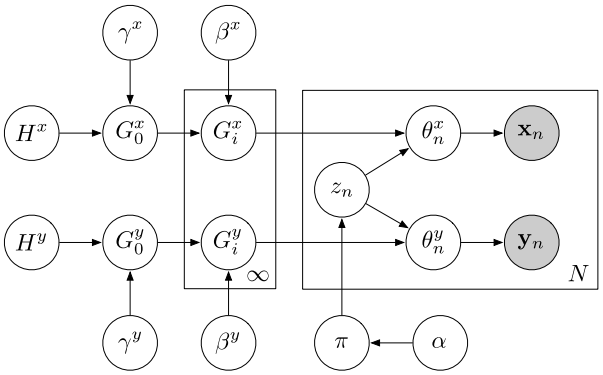
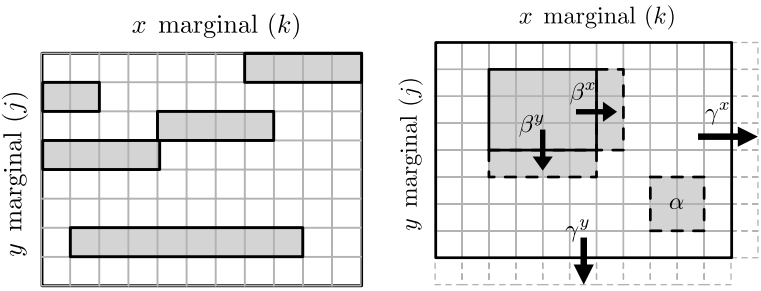


Fig. 1 Plates diagram of the mixture model



(a) Model of Rogers et al (2008) (b) Proposed model

Fig. 2 Schematic representation of the factorisation over contingency tables for the model of Rogers et al. (2008) and the proposed model. The *arrows* and *dashed lines* illustrate the effect of the concentration parameters α, β, and γ. For example, a higher β^x will *a priori* favour top-level clusters that are bigger in the x-direction

again separate. But within a marginal, 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 marginal, 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. The G_0^x and G_0^y, on the other hand, provide discrete base measures and in a sense give identities to the marginal clusters.

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 marginals would not be necessary to bind component identities, which are determined by indices. The infinite version can be obtained as a limit only with hierarchical Dirichlet priors as explained by Teh et al. (2006).

Figure 2(b) 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 marginals controlled (*a priori*) by the concentration parameters γ^x, γ^y. Within the table, we see a decomposition into block-like components.[2]

[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.

New blocks are produced or removed according to the DP over top-level components, controlled by α. The blocks may also grow 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 component structure. The definition of a top-level component is a product of (conditionally, on i) independent distributions over the sets of marginal components. In this sense, one could legitimately split the larger component in Fig. 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 components whilst small α and high β will prefer a small number of large components. Such control is useful when we possess prior knowledge regarding the type of components that are of interest. In the absence of such knowledge, one can place hyper-priors on these parameters and sample them within the Gibbs scheme, as is done in all the experiments in this article. 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).

The hyper-priors can also be used to encode hard constraints. In a recent study (Rogers et al. 2008) a different decomposition was proposed, namely $p(j,k) = p(k)p(j|k)$. This can be seen as a special case of our more general model where each of the K different top level components would link one particular k with some of the js. This is obtained by setting β^x to zero.

Another special case is where each top-level component is only allowed to include one component from each marginal ($\beta^x = \beta^y = 0$). The top-level components recovered here would be similar to clusters found with a standard mixture model applied to the concatenation of the data vectors. Such a model would be appropriate if the partitioning of objects in the two marginals were the same or very similar.

2.2 Inference

In this section we give an overview of the inference procedure for the proposed model. Our proposed sampling scheme uses collapsed Gibbs sampling (the processes are marginalized and not dealt with explicitly) and builds on a large body of published work. For example, Blackwell and MacQueen (1973) solved the marginalization task for standard DPs, and Teh et al. (2006) explain how an hierarchical DP can be marginalized out in a similar fashion. Our sampling scheme follows the Chinese Restaurant (CR) analogy, in particular the Chinese Restaurant Franchise extension for the HDP model (e.g., Teh et al. 2006). The CR metaphor describes the process of assigning individual customers (data points) into tables (clusters) in a restaurant, given the knowledge of table assignments of all other customers. This is directly analogous to how a Gibbs sampler works.

For HDP, the process describes a franchise rather than an individual restaurant; the top-level components are individual restaurants in the franchise of the whole model. The chef in each restaurant constructs a menu by choosing items from some global menu supplied by head office. As with the classical CR analogy for a simple infinite Gaussian mixture, customers sitting at a table all choose the same dish and on entering a restaurant, a diner sits at a table with probability proportional to the number of people sitting at it already, or sits alone at a new table with probability proportional to a DP concentration parameter.

The current work proposes two extensions to this framework. Firstly, we are dealing with multiple data sources. To fit this in to the CR analogy, one can think of eating several courses

(rather than one in the plain hierarchical DP). Each restaurant selects courses (restricting ourselves to two—main and dessert—for brevity) from course-specific menus supplied by head office and diners may eat combinations of these. Unlike the standard scheme where a diner might sit at the same table for their entire meal, in this franchise different courses are served at different tables—after finishing their main, diners then move to the appropriate dessert table. The second innovation is that the diners are no longer assigned to the restaurants *a priori*. Simply speaking, diners enter restaurant r with prior probability proportional to the number of people in there (diners are wary of empty-looking restaurants) and the likelihood is computed by taking into account how much the diner likes the food being served.

As seen above, the CR metaphor has to be stretched uncomfortably to fit the new model. In the remainder of the paper, we will drop the metaphor and instead describe the top-level components as *blocks*, within which are *instances* of marginal components. Every object is assigned to a block and, within that block, to an instance of a marginal component for each data source.

We assume base measures conjugate to the mixture component likelihoods, and hence can integrate out the cluster parameters θ associated to the marginal clusters. We demonstrate implementations with Gaussian and multinomial likelihoods in the experimental section and provide the appropriate conditional distributions below. The resulting sampler 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 (in the following, we often simplify this notation to $p(\boldsymbol{x}|j, \Delta)$, where j is shorthand for $\boldsymbol{X}_j$, the set of data instances currently assigned to component j). 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.

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 marginal component assignments (j, k). The latter are done in a nested scheme of data-instances and marginal-instance assignments. Conditional probabilities for sampling the top-level components can be obtained by marginalizing over the potential marginal assignments within that component (admitting the possibility of creating a new marginal instance). To avoid confusion with the cluster-specific conditional distributions $p(\boldsymbol{x}|j, \Delta)$, we use the notation $p_m(\boldsymbol{x}|i, \Delta)$ for such mixture distributions. We use z_n to denote the top-level assignment of object n. Two sets of variables are required to describe the marginal 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}^y tell which marginal component is assigned to each instance t of top-level component i. The marginal cluster identities of sample n are then obtained by double indexing: $w^x_{z_n v_n^x}$ and $w^y_{z_n v_n^y}$.

Because the marginal-instance assignments are common to many data objects, they constitute a separate sampling step that we run after going 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.

We now describe each of these steps in more detail. 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 the set $(\boldsymbol{X}^{-n}, \boldsymbol{Y}^{-n}, z^{-n}, (v^x)^{-n}, (w^x)^{-n}, (v^y)^{-n}, (w^y)^{-n}, \Delta^x, \Delta^y)$. The counters appearing in the formulas are functions of the latent assignments z, v, and w.

Sampling top-level component-data assignments z The top-level component for a left-out sample is obtained by marginalizing over the potential marginal cluster assignments for the

sample within each component i:

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

We have denoted the number of samples in the component i by C_i, with the superscript $-n$ here and elsewhere denoting the absence of the left-out sample n. Instances on the marginal 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 by marginalizing over marginal component assignments, giving

$$p_m(\boldsymbol{x}_n | t^*, \Delta^x) = \frac{\gamma p(\boldsymbol{x}_n | \Delta^x) + \sum_j d_j^{-n} p(\boldsymbol{x}_n | j, \Delta^x)}{\gamma + \sum_j d_j^{-n}}, \tag{2}$$

where the d_j count the numbers of samples on the x-marginal associated to components j. For existing components we also need the marginal-specific probabilities for top-level component assignments—for instance for the x-marginal,

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

where instance assignments have now been marginalized out, c_{it} counts samples associated to instance t of top-level component i, T_i is the number of instances for top-level component i, and w_{it}^x is used as shorthand for this set of samples.

The formulas for $p_m(\boldsymbol{y}_n | u^*, \Delta^y)$ and $p_m(\boldsymbol{y}_n | i, \Delta^y)$ 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 marginal, we only treat the x-marginal here, without using the marginal superscripts.

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

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

where $p_m(\boldsymbol{x}_n | t^*, \Delta)$ is the probability of setting up a new instance for the sample (2). If a new instance was created, a marginal cluster needs to be associated to the instance, by drawing from the urn associated to the base DP,

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

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

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

All data previously associated to the instance are denoted by $\boldsymbol{X}_{it}$, and $\boldsymbol{X}_j^{-(it)}$ denotes all data in component 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.

Gaussian marginals A common choice of marginal distribution for real valued data is Gaussian with a conjugate Normal-Inverse-χ^2 prior. As mentioned above, the conjugacy means we can marginalize over the component parameters and work directly with distributions of the form $p(\boldsymbol{x}_n|\boldsymbol{X}_j, \Delta^x)$ where $\boldsymbol{X}_j$ describes some set of other data instances. In our examples, we will assume independence over the dimensions of $\boldsymbol{x}$, i.e. the mixture components that we will marginalize over have the following prior

$$p(\boldsymbol{\mu}_j, \boldsymbol{\sigma}_j^2|\Delta^x) \propto \prod_{d=1}^{D} \sigma_{jd}^{-1} (\sigma_{jd}^2)^{-(v_0/2+1)} \exp\left(-\frac{1}{2\sigma_{jd}^2}[v_0\sigma_0^2 + \kappa_0(\mu_{0d} - \bar{x}_{jd})^2]\right).$$

with hyper-parameters $\Delta^x = \{v_0, \kappa_0, \sigma_0, \mu_0 1, \ldots, \mu_{0D}\}$. The distribution of interest is

$$p(\boldsymbol{x}|\boldsymbol{X}_j, \Delta^x) = \int p(\boldsymbol{x}|\boldsymbol{\mu}_j, \boldsymbol{\sigma}_j^2) p(\boldsymbol{\mu}_j, \boldsymbol{\sigma}_j^2|\boldsymbol{X}_j, \Delta^x) d\boldsymbol{\mu}_j d\boldsymbol{\sigma}_j^2.$$

The posterior distribution, $p(\boldsymbol{\mu}_j, \boldsymbol{\sigma}_j^2|\boldsymbol{X}_j, \Delta^x)$ is a product of D Normal-Inverse-χ^2 distributions with parameters (now additionally indexed by dimension d)

$$\begin{aligned}
\mu_{dj} &= \frac{1}{\kappa_0 + N_j}(\mu_0\kappa_0 + N_j\bar{x}_{jd}),\\
\kappa_j &= \kappa_0 + N_j,\\
v_j &= v_0 + N_j,\\
v_j\sigma_{jd}^2 &= v_0\sigma_0^2 + (N_j - 1)s_{jd}^2 + \frac{\kappa_0 N_j}{\kappa_0 + N_j}(\bar{x}_{jd} - \mu_{0d})^2,\\
s_{jd}^2 &= \frac{1}{N_j - 1}\sum_{n=1}^{N_j}(x_{nd} - \bar{x}_{jd})^2,
\end{aligned}$$

where $\bar{\boldsymbol{x}}_j$ is the mean of $\boldsymbol{X}_j$ with individual components $\bar{x}_{jd}$ and N_j is the number of instances in $\boldsymbol{X}_j$. The predictive distribution required by the various sampling steps follows Student-t,

$$p(\boldsymbol{x}|v, \boldsymbol{\mu}, \boldsymbol{\sigma}^2) = \prod_{d=1}^{m} \frac{\Gamma((v+1)/2)}{\Gamma(v/2) + \sqrt{v\pi\sigma_d^2}}\left(1 + \frac{1}{v\sigma_d^2}(x_d - \mu_d)^2\right)^{-(v+1)/2}$$

with parameters

$$\begin{aligned}
v &= v_j + 1,\\
\boldsymbol{\mu} &= \boldsymbol{\mu}_j = [\mu_{j1}, \ldots, \mu_{jD}]^T,\\
\sigma_d^2 &= \frac{v_j\sigma_{jd}^2(\kappa_j + 1)}{\kappa_j(v_j + 1)}.
\end{aligned}$$

Multinomial marginals Another popular choice is multinomial components with a Dirichlet prior on the multinomial parameters. Assuming a Dirichlet prior with a single concentration parameter (extension to dimension-specific concentration parameters is straightforward) δ, a multinomial likelihood for component j with D-dimensional parameter $\mathbf{q}_j = [q_{j1}, \ldots, q_{jd}, \ldots, q_{jD}]^T$ and a set of N_j data-instances, $\boldsymbol{X}_j$ currently assigned to the component of interest, our conditional predictive distribution is given as:

$$p(\boldsymbol{x}|\boldsymbol{X}_j, \delta) = \int p(\boldsymbol{x}|\mathbf{q}_j) p(\mathbf{q}_j|\boldsymbol{X}_j, \delta) d\mathbf{q}_j.$$

The posterior distribution $p(\mathbf{q}_j|\boldsymbol{X}_j, \delta)$ is easily calculated as

$$p(\mathbf{q}_j|\boldsymbol{X}_j, \delta) \propto \prod_{d=1}^{D} q_{jd}^{\delta - 1 + \sum_{n=1}^{N_j} x_{nd}},$$

and the required predictive distribution follows by performing the marginalization described above to give:

$$p(\boldsymbol{x}_*|\boldsymbol{X}_j, \delta) = \left(\frac{\sum_d x_{*d}}{\prod_d x_{*d}} \right) \times \frac{\Gamma(\sum_d \delta + \sum_{n=1}^{N_j} x_{nd})}{\prod_d \Gamma(\delta + \sum_{n=1}^{N_j} x_{nd})} \times \frac{\prod_d \Gamma(\delta + x_{*d} + \sum_{n=1}^{N_j} x_{nd})}{\Gamma(\sum_d \delta + x_{*d} + \sum_{n=1}^{N_j} x_{nd})},$$

where we have used the $*$ subscript to denote the $\boldsymbol{x}$ for which we are computing likelihood to avoid confusion.

Hyperparameter estimation The model specification includes five DP concentration parameters α, β^x, β^y, γ^x, and γ^y, the values of which will determine the readiness with which the model will generate new top-level components, instances and marginal components. The observed number of components at each level is 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 and z contains all the top-level assignments. 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 component and not the total number of instances. Hence, we can think of the I top-level components as I independent realisations of the process controlled by β and obtain a likelihood of the form

$$p(v|\beta) \propto \beta^{\sum_i T_i} \prod_i \frac{\Gamma(\beta)}{\Gamma(C_i + \beta)},$$

where v again contains the instance assignments of all samples.

Alternatively, one could maintain separate β for each top-level component which may be useful if it was expected that components 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. In practice already a simple Metropolis-Hastings is efficient enough, only taking a small proportion of the total computation time.

3 Related work

3.1 Generative modeling of coupled sources

Skipping specific details of the model structure of Fig. 1, it shares basic elements with other generative approaches for modeling dependencies between co-occurring data sources (Archambeau and Bach 2009; Bach and Jordan 2005; Barnard et al. 2003; Blei and Jordan 2003; Klami and Kaski 2007, 2008). 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.

The main difference to the earlier models is that we assume a much weaker link between the structures in the two spaces. For example Klami and Kaski (2008) and the multi-modal hierarchical aspect model of Barnard et al. (2003) both assume each component is unimodal in both spaces, and their models can intuitively be seen as special cases of (a finite variant of) ours. If we would allow each top level cluster to only use one marginal component of each view we would get a similar model. As illustrated later in Sect. 4.1, modeling data having complex interactions with unimodal marginals results in larger number of components as each cross-cluster needs to be modeled as a separate component.

Blei and Jordan (2003) tackle the problem of jointly modeling images and their captions. They propose a generative model (`corr-LDA`) and compare it to a model with unimodal conditional marginals (`GM-Mixture`) and one with LDA-style marginals (`GM-LDA`). Their results show that `corr-LDA` performs better at the task of predicting one view from the other and both of the methods with complex marginals (`GM-LDA` and `corr-LDA`) model the data more accurately than `GM-Mixture`. The choice between allowing simple or complex marginal structure is, ultimately, a question of the application task. If the task is predicting the other view, then strong link between the sources is to be preferred, and it may even be beneficial to explicitly maximize the consensus between the marginal clusterings instead of merely maximizing the joint likelihood (Bickel and Scheffer 2004). Our model does not even attempt accurate prediction, since every component can be multimodal in both directions, and hence the main application is in exploring the dependencies between the views. Previously methods designed for this task have primarily been projection methods (Bach and Jordan 2005; Klami and Kaski 2007) with unimodal Gaussian variation given the latent source, and our model extends their motivation to more complex marginal models.

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 defined for a 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 view-specific branch of our model, but would not be applicable for solving the task of coupled modeling of several views.

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, and presented an example of the use of this process for relational data. The process is based on multidimensional stick-breaking, and in general can be used to induce an axis-aligned partition on any product space. Hence, similarly to our hierarchical structure, the process could be used to define a prior over contingency tables.

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.

The proposed HDP prior process factorizes a matrix into 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 each top-level cluster 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 components and only requiring that the marginals of each component are independent.

3.4 Analysis of coupled mRNA and protein data

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—one of the first such datasets produced. As our model can be seen as an extension of that model, we here describe both the application and the differences between the models in more detail.

As high-throughput protein measurement becomes more common, it is likely that more coupled protein/mRNA data 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 or translational (protein) levels. Simply put, gene expression is a multi-stage process consisting of transcription (the production of mRNA caused by the *switching on* of a gene) and translation (the production of a protein based on the design encoded in the mRNA). Regulation (and hence control) of this process can occur at different stages in this process. For example, the presence/absence of transcription factors (proteins that *switch* genes on/off) will control mRNA production. Similarly, post-transcriptional effects can control whether or not a protein is produced from the mRNA at a particular time. Analysis of mRNA data (measured on microarrays) has led to increased knowledge of transcriptional regulation (e.g. Friedman et al. 2000) but tells us nothing about post-transcriptional effects—such effects are further downstream in the expression process. Similarly, analysis of protein levels, whilst undoubtedly interesting, cannot provide much detail regarding regulation: by only measuring the end product, we cannot tell at which point control is present in the process. It is obvious that to truly begin understanding the whole process, we must analyse both data types concurrently.

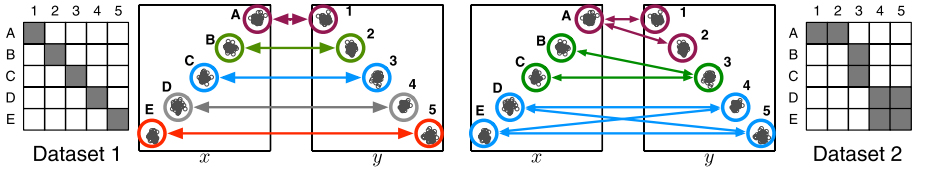
The model proposed by Rogers et al. (2008) attempted to perform such analysis by fitting a mixture model to each view, coupled by a joint prior $p(j,k)$ over the assignments to protein and mRNA clusters. One of the principal differences between the model of Rogers et al. (2008) and the model proposed here is how this distribution is decomposed. In Rogers et al. (2008), the table is split into rows: the probability of assigning a gene to a particular protein cluster is conditioned on which mRNA cluster that gene was assigned to, which translates to factorization $p(j,k) = p(k)p(j|k)$. Biologically, this corresponds to the assumption that protein evolution over time is in some sense conditioned on the mRNA profile over time. When analysing the results of this model, the first thing that stands out is the high level of interconnectivity between clusters from each data source. These connections do not just take the form of each mRNA cluster connecting to more than one protein cluster (as the decomposition implies) but also consists of links connecting some protein clusters to several mRNA clusters. These links cast some doubt on the suitability of the decomposition chosen and inspired the creation of a more appropriate model with more flexible prior.

For example, consider a group of 3 mRNA clusters whose members are all split between 2 protein clusters. In the decomposition of Rogers et al. (2008), we would lose the clear relationship between the mRNA clusters and just discover three separate relationships, each between one mRNA cluster and 2 protein clusters. There may be biological significance in this relationship that we would obviously wish to preserve. The model proposed here overcomes this limitation by decomposing the joint distribution into more flexible components and our principal aim in Sect. 5 is to investigate whether such inter-connected components (with multiple modes in each marginal) exist within this dataset and if so, do they have biological significance.

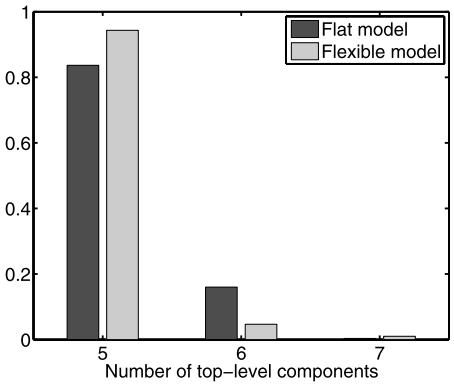
4 Synthetic examples

4.1 Gaussian marginals

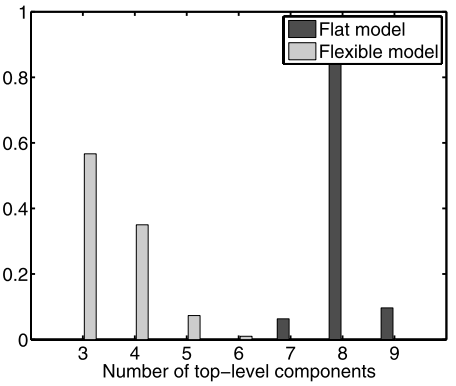
To aid understanding of the model and show its ability to automatically extract top-level components, we present a simple synthetic experiment. We restrict ourselves to a coupled dataset with two dimensions in each marginal to aid visualisation (Fig. 3(a)). Each object consists of two data points—one in data space x and one in y. For example, each object might be a gene with x corresponding to its mRNA expression data and y its protein expression. In each marginal there are 5 distinct data clusters (A, B, C, D, E and 1, 2, 3, 4, 5 in x



(a) Two synthetic coupled datasets. In each dataset, each object is represented by a point in $\boldsymbol{x}$-space and a point in $\boldsymbol{y}$-space. In Dataset 1 (left), clusters in the two spaces have a one to one relationship (relationships shown by arrows) and would be amenable to data concatenation. In Dataset 2 (right), the relationship is more complex and would lead to an over-complex model if the data was concatenated



(b) Posterior distribution over the number of top-level components for Dataset 1 for the proposed flexible model and a restricted variant (flat model). Both models correctly identify the five components (pairs of $\boldsymbol{x}$- and $\boldsymbol{y}$-clusters)



(c) Posterior distribution over the number of top-level components for Dataset 2 for the proposed flexible model and a restricted variant (flat model). The proposed model correctly finds the three components (c.f. Fig. 3) whilst the restricted model is forced to separately model each unique $\boldsymbol{x}$ and $\boldsymbol{y}$ cluster pair

Fig. 3 Demonstration on synthetic data with Gaussian marginals

and y respectively). From this basic setup we generate two datasets. In the first, there is a one-to-one mapping between the clusters in the two datasets. For example, if the data point in the x dataset for an object lies in cluster A, its data point in the y dataset will reside in cluster 1. Similarly, B corresponds to 2, etc. These relationships are shown by the arrows in Fig. 3(a), accompanied with the contingency table representation. It is clear that this dataset has 5 top-level components and could be analysed satisfactorily by concatenating the x and y datasets. In the second dataset, we impose a more complex relationship between the clusters of the x and y data. For example (see arrows in Fig. 3(a) (right)), x-data in either cluster D or E will lie in either of y-clusters 4 and 5. Inspecting this contingency table (Fig. 3(a)), one can see three distinct top-level components (*blocks*) that we would like to recover, as each potentially corresponds to interesting cluster-cluster relationships. Three is the smallest number of top-level blocks that can decompose the table based on our assumption of each block corresponding to the product of independent marginal-specific distributions. Note that a standard clustering model on concatenated data model would require 8 clusters to describe the relationship between the two data sets (there are 8 unique x-y cluster combinations) and in doing so would be unable to highlight the interesting relationship between, for example, D, E and 4, 5.

We present this synthetic data to our proposed, flexible model and also to a constrained version that performs a 'flat' clustering. This constrained version has $\beta_x = \beta_y = 0$, restricting each top-level component to be associated with exactly one component from each of the x and y. This is similar to concatenating the data and applying a standard mixture model, although slightly more efficient as it allows marginal components to be present in more than one top-level component (concatenation would effectively require 8 components in each marginal, some of which would be almost identical). A product of independent Normal-inverse-χ^2 priors was 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 reasonably uninformative Gamma ($\mathcal{G}(1, 1)$) prior distributions.[3]

In Fig. 3(b) we can see the posterior distribution over the number of top-level components for the two models for dataset 1. We notice that both models place high posterior weight onto 5 top-level components as desired. When moving to dataset 2 (Fig. 3(c)) we see a large disparity between the two models. The fully flexible model successfully extracts the 3 interesting components whereas the restricted model has to create 8 components to explain the relationship. From these results, it is reasonable to conclude that the model can successfully discover complex cluster-cluster relationships through the top-level components if such relationships exist.

As an aside it is worth mentioning that the results from the restricted variant of the model used for comparison here, could be analysed to find these top-level components in a *post-hoc* manner, although this is only possible due to the aforementioned 'sharing' of marginal components and would not be possible in a standard mixture model on the concatenation. However, the ability to extract them automatically has clear benefits.

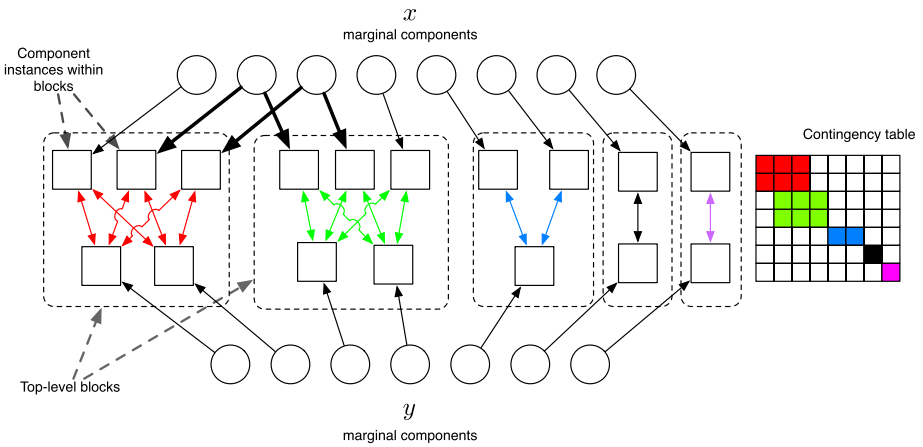
4.2 Multinomial marginals

The method characterizes dependencies between marginal clusterings by extracting component structure in the cross-cluster table. As illustrated in Fig. 2(b) and described in Sect. 2, the level of detail can be tuned with the prior parameters $(\alpha, \beta^x, \beta^y, \gamma^x, \gamma^y)$. Here we demonstrate the effect of these parameters in practice on synthetic data with clear block-like component structure that still overlaps on the marginals, shown in Fig. 4(a), using multinomial data to show that the general model structure is not tied to the marginal likelihoods.

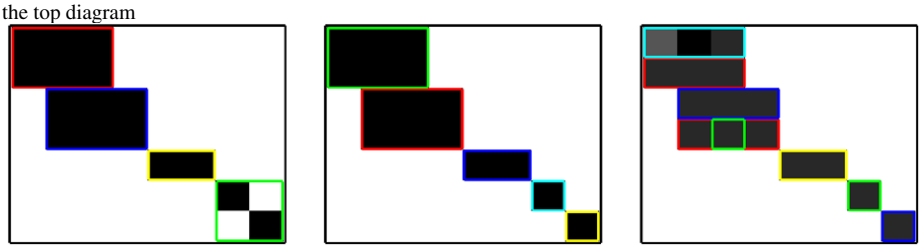
Data in each cluster is drawn from a cluster-specific multinomial. We create 5 independent data points for each link in Fig. 4(a), which results in marginal clusters of varying size and a total of 80 data objects. The dimensionality of the data is 80 for $\boldsymbol{x}$ and 70 for $\boldsymbol{y}$, so that each cluster has a non-zero probability for 20 dimensions and each dimension is used by two different clusters, and each data point consists of 20 trials. In brief, the data collection could be interpreted as small bilingual corpus consisting of 80 documents of 20 words, chosen from a vocabulary size of 80 or 70 depending on the data space. The model is trained with multinomial marginal likelihoods and Dirichlet priors with a count of one for each element.

To illustrate the effect of the top-level concentration parameter α, we run a set of experiments with different fixed values for it. The other hyper-parameters $(\gamma^x, \gamma^y, \beta^x, \beta^y)$ are all given the $\mathcal{G}(1, 1)$ prior. As seen in Fig. 4(b), for very small values of α the model does not find the true component structure, but instead groups several components together. For larger values the model finds component-structures of increasing complexity, starting

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



(a) Illustration of the data generation. The marginals constitute 8 and 7 clusters respectively, and the diagram on top shows the instances of the marginal components within the top-level blocks according to the contingency table shown on the right. Note the two $\boldsymbol{x}$ components that have instances in two different top-level blocks. The contingency table on right is an alternative illustration of the generative process, showing the block structure emerging from the connections between the 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 components 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 components of maximal size, whereas $\alpha = 3$ (right) splits the components into smaller parts. 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 the top-level concentration parameter α

with the right decomposition but for still larger values splitting some of the components into multiple ones. Eventually, for $\alpha > 10$, the model reduces to essentially the flat model, modeling each cross-cluster with a separate component. This is understandable considering we only have 80 data points and $\alpha > 10$ already represents a considerable bias towards a large number of components. Even then the generative process is correct, despite clear over-parameterization. Finally, setting up the prior $\alpha \sim \mathcal{G}(1, 1)$ results in posterior mean at $\alpha = 1.33$ and 60% of posterior mass at the correct result of 5 components, showing that despite the weak link between α and the observed data it is still possible to infer the complexity at least for clear enough component structure. For completeness, the posterior mode of the constrained flat model with $\beta^x = \beta^y = 0$ is at 16–17 components, correctly capturing each cross-cluster as a separate component.

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) and has previously been 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 of cells at 8 unevenly spaced time-points between $t = 0$ and $t = 24$ hours. For clarity, each gene is a data example and for a particular gene, $\boldsymbol{x}$ and $\boldsymbol{y}$ are vectors corresponding to the mRNA and protein time-series for this gene. 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 did not pass the necessary quality controls. Therefore, we were left with 6 mRNA and 7 protein time-points (i.e., $\boldsymbol{x}$ is 6 dimensional and $\boldsymbol{y}$ is 7 dimensional). 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, hyper-priors and sampling details

Following Rogers et al. (2008), we assume independence over time for our marginal base measures and use Gaussian mixture components for the two marginals and a product of univariate Normal-inverse-χ^2 priors (with hyper-parameters $\upsilon_0 = 1, \kappa_0 = 1, \mu_0 = 0, \sigma_0^2 = 1$; see e.g., Gelman et al. 2004). Clearly there is a time dependence in the data (mRNA at time t must have some effect on mRNA at $t + 1$), however, the strong cluster structure found in (Rogers et al. 2008) suggests that assuming independence is not likely to be detrimental to the analysis. In addition, the amount of data, only 542 data-points, is unlikely to be enough to infer full covariance matrices. Reasonably strong priors favoring diagonality would need to be employed—a factorizing distribution is then a neater choice and is more flexible than the frequent choice of isotropic covariance.

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). In previous sections, we discussed how priors over the hyper-parameters can be set to reflect prior knowledge on the number and size of blocks present. For this dataset, we have no such knowledge—we are interested in investigating the size and shape of the blocks present—and so use reasonably uninformative $\mathcal{G}(1, 1)$ priors for all concentration parameters in all models.

The sampler was run for 10000 samples, with the first 5000 discarded as a burn-in phase. The retained samples were then thinned to remove correlations at a ratio of 50-1, resulting in 100 independent posterior samples for further analysis. In our analysis, we show distributions over these posterior samples except where we visualise one prototypical sample—Figs. 6 and 7.

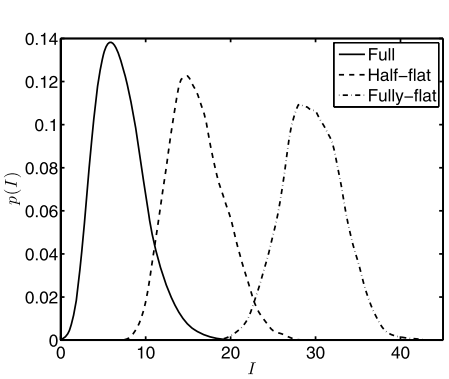
5.2 Comparison with previous models

Our principal aim (as mentioned in Sect. 1.1) was to discover if complex components (multimodal in both marginals) exist in this data, and if they do, do they have any biological significance. To this end, we will first compare the number of top-level components produced by the proposed model with those produced by two restricted variants. The first, which we

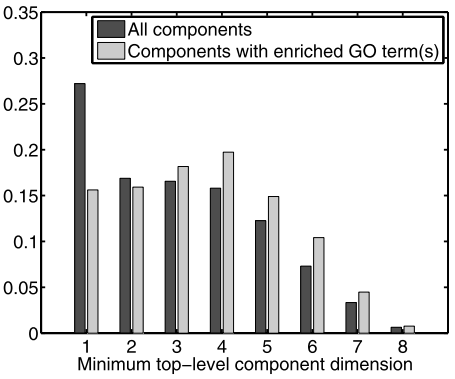
call 'fully flat', was introduced in the synthetic Gaussian experiment and it restricts each top-level cluster to have only one component from each marginal. This model is similar to concatenating the two sources (more details in Sect. 4.1). The second variant is similar to the model proposed in Rogers et al. (2008), in that top-level components may only have one component from the mRNA marginal but as many as necessary from the protein marginal. We could of course also restrict the model in the opposite manner (top level blocks may only have one protein component) but this would not be directly comparable to the previous work. Otherwise the model follows the proposed general model. We compare with this variant rather than the actual model of Rogers et al. (2008) as this way we can remove the effect of the different inference techniques (Gibbs sampling rather than ML), priors and model complexity. In exploratory analysis such as this, it is difficult to directly numerically compare the models as the *true* partitioning is unknown. However, we can examine the results of the proposed model to see if it suggests that interesting biological structure is present that could not be found by the alternative approaches. In this instance, the important question is whether or not top-level components with more than one marginal component on each side exist and, if so, do they have biological significance. Such components could not be discovered by either alternative model.

In Fig. 5(a) we see the posterior distribution over the number of top-level components (with > 5 members to remove transient components—that is, short-lived components generated in the HDP sampling process) for the three models. The significantly smaller number of components for the full model suggests that there are indeed complicated interconnected components present. For further evidence, we look at the size of the components (in terms of number of marginal components present). In Fig. 5(b), we can see the distribution over the smallest component dimension taken across all posterior samples (e.g. if a component has 1 mRNA component and 3 protein components, its smallest dimension is 1). More than half of the components have at least 2 components in both marginals. In addition, we compute whether or not there are enriched Gene Ontology (GO) terms in these components (GO term enrichment is discussed more thoroughly in Sect. 5.4). The lighter bars in Fig. 5(b) show the distribution over minimum size of components that are associated with enriched GO terms. We can see that there are a considerable number of GO terms enriched in components with > 1 marginal components on both sides. This represents excellent evidence that not only do larger blocks exist, they have biological significance. Recall that in either a concatenated model, or the model of Rogers et al. (2008), these components would not be found and important biological information may hence be missed. We examine some of these components in the following sections.

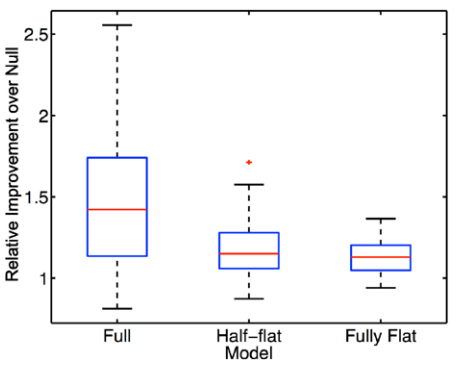
Because the number of hypothesis tests we will be performing for each model in each sample varies, it is hard to directly compare the total number of enriched GO terms in each model. However, we can investigate the relative improvement in GO term enrichment of each model compared to the same model with a permutation of the genes. We denote by g_s the number of enriched terms (at $p \leq 0.05$) in posterior sample s, and by $\hat{g}_s$ the number of terms enriched if we randomly permute the order of the genes. This latter term could be thought of as an approximation to the number of false positives we might expect given the number of components and their sizes. Hence, taking the ratio, $g_s/\hat{g}_s$ provides a measure of the performance of the model (in terms of GO term enrichment) normalised to take into account the number of false positives we might expect given the number of components, their sizes and hence the number of tests performed. We can compute this ratio for all posterior samples, and the three resulting distributions are shown in Fig. 5(c). All three models give (on average) a slight improvement over the null model with the full model showing the largest improvement over the model-specific baseline. One should always take comparisons



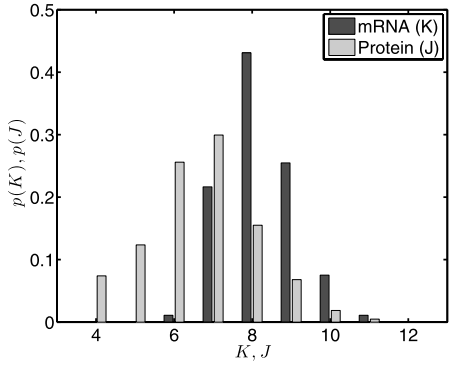
(a) Posterior distribution over the number of top-level components for the full model and the two restricted models corresponding to the model of Rogers et al. (2008) (Half-flat) and data concatenation (Fully-flat)



(b) Size of the minimum component dimension (bars show proportion of components over all posterior samples)



(c) Relative improvement of number of enriched GO terms for each model over appropriate null distribution. In each box, the central line corresponds to the median, the box to the 25th–75th percentile and the dashed line to the range of the data (outliers are plotted separately)



(d) Posterior distribution over the number of components in each marginal

Fig. 5 Posterior distributions for biological data

based on the gene ontology with a pinch of salt (the ontology is incomplete and likely to be noisy), but the results presented here suggest that the proposed model is finding more biological structure than the alternatives.

Finally, in Fig. 5(d) we can see the posterior distribution over the number of marginal components (again, only components with > 5 members). We see fewer components than used by Rogers et al. (2008), however, inspection of Bayesian Information Criteria (BIC) plotted 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

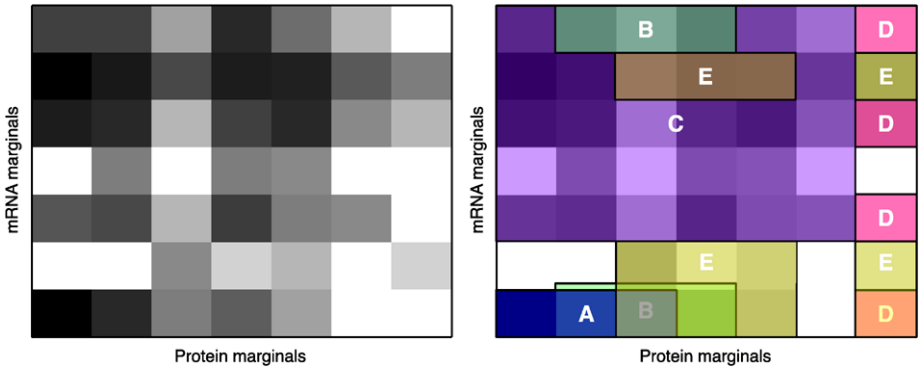


Fig. 6 Example contingency table—cell counts in grayscale (*left*—darker is higher) and top-level component structure (*right*)

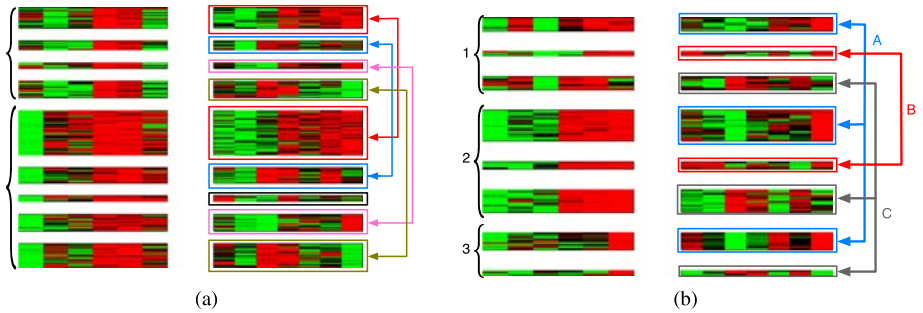


Fig. 7 Examples of two stable top-level components from sampler. In both cases, mRNA marginals are given on the left with curly brackets denoting the different marginal components. Protein marginals are given on the right and marginal components are denoted by *boxes* and *lines*. The genes are presented in the same order for both mRNA and protein marginals. The heatmaps follow the standard biological convention of *red* for high expression, *green* for low expression

for each dimension (time-point)—our parameterisation is more flexible. In light of these observations, the difference is not surprising. We find similar distributions over the number of marginal components in both restricted models.

5.3 Visualising the contingency table

As the sampler explores the posterior, the size of the contingency table and the number of top-level components 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 Fig. 6. The left plot shows the table with entries coloured by the number of genes that fall into that particular marginal combination. In the right hand plot, we have overlaid this plot with the five biggest top level components (labeled A–E). 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 components having independent marginals. 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, for example, the 4 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 components 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 using, for example, the hyper-geometric distribution (see, for example Rivals et al. 2007) to assess how significant it is to observe n_t genes in a particular cluster labeled with a particular GO term t, out of a total of N genes, N_t of which share this label. 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 top-level components. Particularly, if we take components 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 top-level component. In Fig. 7(a) we see one such example. The mRNA profiles in the left plot come from 2 marginal components (denoted by the curly brackets). The protein profiles come from four marginal components (color-coded and connected by arrows). These are the most regularly occurring genes/marginal components in this top-level component, which persisted for $\sim$1500 of the un-thinned posterior samples (30 of the thinned samples). Examining the figure, we see that the two mRNA marginals (left) both interact with 3 of the protein components creating a 2×3 block in the contingency table. Additionally, we can mine these genes for enriched GO terms. We find 12 terms with $p < 0.1$. Analysing the location within the group of genes tagged with these terms, we find that all terms but one have representatives in more than one component on *both* sides. This strongly supports the claim that the top-level components can represent meaningful biological structure.

A second example is given in Fig. 7(b). In this case, we have 3 marginal mRNA components and 3 marginal protein components. This configuration corresponds to a 3×3 block in the contingency table. Again, we can mine these genes for enriched GO terms and find 13 terms with $p < 0.1$. Of these terms, all are present in at least 2 components on one side or the other and 9 of these have members in more than one marginal component on *each* side. Two interesting examples are GO:0003735 tagged to 8 genes in the component in all three mRNA and one protein marginal, and GO:0006412 tagged to 12 genes and present in all three mRNA marginals and all 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 makes 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 1, 2 and 3 (see Fig. 7(b)) and protein component A whereas in general, translation genes appear in all marginals on

both sides. 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 components and the many others produced by the model is an area of ongoing research. The discovery of such components 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 components potentially with representatives from more than one component in both marginals.

The second enrichment analysis method we use averages the enrichments over all posterior samples: for each gene-term pair, we compute the probability of enrichment in the top level 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. 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 Fig. 8 we show 3 examples of terms significant in the top-level components and in Fig. 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 top-level components, we see GO:0003735 (the ribosome, discussed previously) and GO:0000502, the proteasome. 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 Fig. 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 proteasome 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 Fig. 8, we see genes with predominantly homogeneous mRNA and protein profiles. This suggests tight co-regulation at both the transcriptional and post-transcriptional stages. Conversely, in Fig. 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 component 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

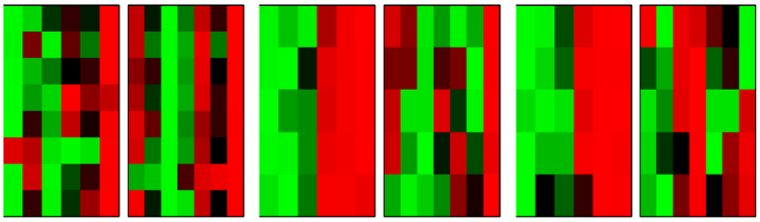


Fig. 8 3 examples of gene ontology terms significantly enriched in top level components. In all cases, *left* heat map is mRNA data, *right* is protein data

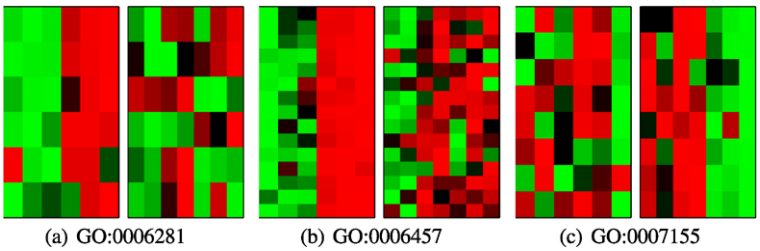


Fig. 9 3 examples of gene ontology terms significantly enriched in one marginal component (mRNA for (**a**) and (**b**), protein for (**c**)) and not the other. (*Left*—mRNA, *right*—protein)

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 four important respects. First, it explicitly attempts to find structure in the joint distribution/contingency table of the marginal components. Secondly, the decomposition of the contingency table allows top-level components to be multimodal in each marginal. Third, model complexity is automatically inferred from the data and, finally, a Gibbs 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.

The development of the model was inspired by a new biological dataset consisting of time-series mRNA and protein profiles for ∼500 human genes, previously analysed by Rogers et al. (2008). In particular, we were interested in discovering whether or not complex relationships exist between clusters in each marginal that have more than one marginal component on each side. Such components could not be found directly with either the model proposed in Rogers et al. (2008) or with a standard mixture model on concatenated data. When applied to this data, the proposed model found significantly fewer components than variants of the model corresponding to the model in Rogers et al. (2008) and concatenated clustering. This suggested that complicated cluster-cluster relationships do indeed exist. Analysis of the size of the resulting components and whether or not they were biologically significant (via testing for enriched gene ontology terms) further substantiated this claim. We have provided a preliminary analysis of a few interesting-looking components in this work and further investigation into the biological implications of this analysis is an area of ongoing research.

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 marginal 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.

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

Archambeau, C., & Bach, F. R. (2009). Sparse probabilistic projections. In D. Koller, D. Schuurmans, Y. Bengio & L. Bottou (Eds.), *Advances in neural information processing systems* (Vol. 21, pp. 73–80). Cambridge: MIT Press.

Bach, F. R., & Jordan, M. I. (2005). A probabilistic interpretation of canonical correlation analysis (Tech. Rep. 688). Department of Statistics, University of California, Berkeley.

Barnard, K., Duygulu, P., Forsyth, D., de Freitas, N., Blei, D. M., & Jordan, M. I. (2003). Matching words and pictures. *Journal of Machine Learning Research*, *3*, 1107–1135.

Becker, S., & Hinton, G. E. (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* (pp. 19–26). IEEE.

Blackwell, D., & MacQueen, J. B. (1973). Ferguson distributions via Polya urn schemes. *The Annals of Statistics*, *1*(2), 353–355.

Blei, D. M., & Jordan, M. I. (2003). Modeling annotated data. In *Proceedings of the 26th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 127–134). New York: ACM Press.

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

Cohn, D., & Hoffman, T. (2001). The missing link—a probabilistic model of document content and hypertext connectivity. In T. Leen, T. Dietterich & V. Tresp (Eds.), *Advances in neural information processing systems* (Vol. 13). Cambridge: MIT Press.

Dhillon, I. S., Mallela, S., & Modha, D. S. (2003). Information-theoretic co-clustering. In *Proceedings of KDD'03, the ninth ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 89–98). New York: ACM Press.

Englebienne, G., Cootes, T., & Rattray, M. (2008). A probabilistic model for generating realistic lip movements from speech. In J. Platt, D. Koller, Y. Singer & S. Roweis (Eds.), *Advances in neural information processing systems* (Vol. 20, pp. 401–408). Cambridge: MIT Press.

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* (pp. 127–135). New York: ACM. doi:10.1145/332306.332355.

Gelman, A., Carlin, J., Stern, H., & Rubin, D. (2004). *Bayesian data analysis* (2nd ed.). London: Chapman and Hall.

Hofmann, T. (1999). Probabilistic latent semantic analysis. In *Proceedings of the 15th conference on uncertainty in artificial intelligence* (pp. 289–296). San Francisco: Morgan Kaufmann.

Johnson, N. L., Kotz, S., & Balakrishnan, N. (1997). *Discrete multivariate distributions*. New York: Wiley.

Klami, A., & Kaski, S. (2007). Local dependent components. In Z. Ghahramani (Ed.) *Proceedings of ICML 2007, the 24th international conference on machine learning* (pp. 425–432). Madison: Omnipress.

Klami, A., & Kaski, S. (2008). Probabilistic approach to detecting dependencies between data sets. *Neurocomputing*, *72*, 39–46. 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, R. M. (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 S. A. Solla, T. K. Leen & K. R. Muller (Eds.), *Advances in neural information processing Systems* (Vol. 12, pp. 554–560). Cambridge: MIT Press.

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

Rodriguez, A., Dunson, D. B., & Gelfand, A. E. (2008). The nested Dirichlet process. *Journal of the American Statistical Association*, *103*(483), 1131–1154.

Rogers, S., Girolami, M., Kolch, W., Waters, K. M., Liu, T., Thrall, B., & Wiley, H. S. (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, D. M., & Teh, Y. W. (2009). The Mondrian process. In D. Koller, D. Schuurmans, Y. Bengio & L. Bottou, (Eds.), *Advances in neural information processing systems* (Vol. 21, pp. 1377–1384). Cambridge: MIT Press.

Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, *101*, 1566–1581.

Vinokourov, A., Hardoon, D. R., & Shawe-Taylor, J. (2003a). Learning the semantics of multimedia content with application to web image retrieval and classification. 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. T. Becker & K. Obermayer (Eds.), *Advances in neural information processing systems* (Vol. 15, pp. 1473–1480). Cambridge: MIT Press.

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 J. Platt, D. Koller, Y. Singer & S. Roweis (Eds.), *Advances in neural information processing systems* (Vol. 20, pp. 1601–1608). Cambridge: MIT Press.

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