Statistical Applications in Genetics and Molecular Biology

Volume 8, *Issue* 1 2009 *Article* 27

Bayesian Unsupervised Learning with Multiple Data Types

Phaedra Agius* Yiming Ying†
Colin Campbell‡

*MSKCC, phaedragius@gmail.com
†University of Bristol, mathying@gmail.com
‡University of Bristol, c.campbell@bristol.ac.uk

Bayesian Unsupervised Learning with Multiple Data Types

Phaedra Agius, Yiming Ying, and Colin Campbell

Abstract

We propose Bayesian generative models for unsupervised learning with two types of data and an assumed dependency of one type of data on the other. We consider two algorithmic approaches, based on a correspondence model, where latent variables are shared across datasets. These models indicate the appropriate number of clusters in addition to indicating relevant features in both types of data. We evaluate the model on artificially created data. We then apply the method to a breast cancer dataset consisting of gene expression and microRNA array data derived from the same patients. We assume partial dependence of gene expression on microRNA expression in this study. The method ranks genes within subtypes which have statistically significant abnormal expression and ranks associated abnormally expressing microRNA. We report a genetic signature for the basal-like subtype of breast cancer found across a number of previous gene expression array studies. Using the two algorithmic approaches we find that this signature also arises from clustering on the microRNA expression data and appears derivative from this data.

KEYWORDS: multiple datasets, correspondence model, Bayesian learning, unsupervised learning, clusters, breast cancer, cancer subtypes, genes, microRNA

1 Introduction

Rapid developments in genomics and proteomics have lead to the generation of many different types of data which has in turn stimulated the development of data fusion techniques. Thus, for supervised learning, a number of different kernel-based methods have been proposed which enable class assignment based on the use of disparate types of input data. Successful multiple kernel classification methods have been proposed which use Bayesian methods [16], semi-definite programming [21], semi-infinite linear programming [29] and column generation methods [4], for example. In a bioinformatics context, examples have been presented where the classification test error is demonstrably reduced through the use of multiple types of data, encoded in different kernels, over the best single data type [15].

Though much less investigated, unsupervised learning could be performed using multiple types of data in certain contexts. In particular, the clustering of samples which use various different feature sets or data types for describing each sample, would require an unsupervised approach that jointly handles these multiple datasets. For example, one may wish to cluster a set of pictures using both descriptors of objects in the pictures and the corresponding picture captions: both types of data carry valid information about the pictured objects. This issue becomes intriguing if we can establish the relationships between the multiple datasets that describe the samples. For the clustering of pictures, the picture captions are descriptive sentences with some partial dependency on the types of objects in the pictures. In this paper we propose a Bayesian unsupervised method for the joint modelling of two types of data with such an assumed dependency. The *correspondence model* we propose is inspired by *correspondence LDA* (Latent Dirichlet Allocation [5, 6]), originally developed for the joint modelling of images and their corresponding caption words. An important aspect of the proposed model is that we can establish the appropriate model complexity i.e. the number of clusters in the data. In addition the model generates density estimates for the data belonging to the two component datasets. We propose two algorithmic approaches which we call *corrMAP* (maximum a posteriori) and *corrVB* (variational Bayes). The *corrVB* approach has two advantages over *corrMAP*: it gives a full posterior distribution estimate instead of point estimates, and model complexity can be determined far more easily without the computationally expensive cross validation approach necessary in *corrMAP*. Although we only present results for the more intuitive *corrMAP* model, we used the *corrVB* approach for validation and obtained very similar results.

In the experimental Section 3 we first evaluate performance on artifi-

cially created datasets with known labelling. This enables objective assessment of performance using a Jaccard score. We then pursue a breast cancer study in which microRNA and gene expression array datasets have been derived from the same patients. Biological studies suggest that there is a directed dependence of gene expression, at least in part, on microRNA activity. Therefore, taking into account this data dependence via our correspondence model, the goal is to cluster the patients and derive meaningful cancer subtypes, and to isolate abnormally expressing genes or microRNA within these subtypes. We show that the resultant model is consistent with previous findings and is biologically plausible.

2 Bayesian Models and Inference

Before describing the model we first introduce some notation. Let D be the sample set to be clustered, indexed by d. Each sample has two component datasets, labelled as C and E, with an assumed dependence of E on C. While C comprises H features (indexed by h) and D samples, E comprises G features (indexed by g) and D samples. Thus, for our breast cancer example in Section 3.2, these datasets correspond to microRNA expression and gene expression respectively. After training, the method represents samples as a combinatorial mixture over a finite set of $\mathcal{K}$ soft clusters, with a probabilistic measure given for the assignment of sample d to cluster k.

2.1 A correspondence model for the joint modelling of two datasets

In this section we introduce a correspondence model that captures an underlying functional interaction between two component data sets. In line with previous models, such as correspondence LDA [5], the two data sets are assumed to share a common prior distribution and latent variables. The correspondence model is applicable to the joint modelling of two datasets where there is a directed dependence of one type of data on another. In Section 3.2 we illustrate the model with a dataset for breast cancer, where we assume gene expression data (denoted E) is potentially dependent on microRNA data (denoted C): we will make reference to this example in our following discussion of the method, to illustate the approach. Thus in this example we have pairs of samples (C_d, E_d), i.e. both these readings are taken from the same patient, denoted sample d. We first cluster the microRNA data C_{hd} and then cluster gene expression E_{gd} conditioned on the generated cluster for microRNA sample C_{hd}.

DOI: 10.2202/1544-6115.1441

The correspondence function is realized by a latent variable $y_{dg} \in [1, H]$ modeling the interaction between gene expression and microRNA measurements. This probabilistic graphical model is represented in Figure 1.

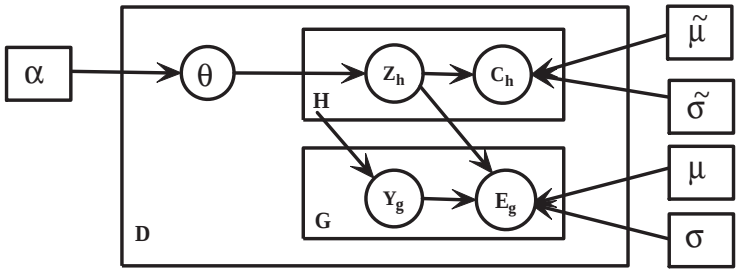

Figure 1: A graphical representation of the generative correspondence model. C_{hd} and E_{gd} are experimental observations and $\{\alpha, \mu, \sigma, \tilde{\mu}, \tilde{\sigma}\}$ are model parameters.

The model is described as follows:

For a given data index d for both E ($G \times D$ matrix) and C ($H \times D$ matrix)

1. Prior distributions: $\theta_d \sim \text{Dir}_K(\alpha)$
2. Choose C_d:
   (a) Choose cluster for C_{hd}: $z_{dh} \sim \text{Multi}(\theta_d)$
   (b) Sample $C_{hd} \sim \mathcal{N}(C_{hd}|\tilde{\mu}_{hz_{dh}}, \tilde{\sigma}_{hz_{dh}})$ where $\mathcal{N}(C_{hd}|\tilde{\mu}, \tilde{\sigma}^2)$ denotes a normal distribution with mean $\tilde{\mu}$ and variance $\tilde{\sigma}^2$. Note that $\tilde{\mu}_{hz_{dh}}$ and $\tilde{\sigma}_{hz_{dh}}$) refer to the mean and standard deviation for samples in cluster z_{dh}.
3. Choose E_d:
   (a) Sample gene correspondence: $y_{dg} \sim \text{Uniform}(1, \ldots, H)$
   (b) Sample $E_{gd} \sim \mathcal{N}(E_{gd}|\mu, \sigma, z, y_{dg}) = \mathcal{N}(E_{gd}|\mu_{gz_{dh}}, \sigma^2_{gz_{dh}}, y_{dg} = h)$

Using the notation $\Theta = \{\alpha, \mu, \sigma, \tilde{\mu}, \tilde{\sigma}\}$, the joint distribution for a given index d is then specified by

$$\begin{aligned} p(C_d, E_d, z_d, y_d, \theta_d|\Theta) &= p(\theta_d|\alpha) \prod_h \big[p(z_{dh}|\theta_d)\mathcal{N}(C_{hd}|\tilde{\mu}_{hz_{dh}}, \tilde{\sigma}_{hz_{dh}})\big] \\ &\quad \times \prod_g p(y_{dg}|H)\mathcal{N}(E_{gd}|\mu_{gz_{dh}}, \sigma^2_{gz_{dh}}, y_{dg} = h) \end{aligned}$$

and the overall joint distribution is given by

$$p(C, E, z, y, \theta|\Theta) = \prod_d p(C_d, E_d, z_d, y_d, \theta_d|\Theta) \tag{1}$$

Extensions to classical Gaussian mixture models (GMM) (e.g. [32]) are possible to handle multiple data sets in a similar fashion to the mixture model presented here. However, for a GMM each datapoint E_d is only related to a latent variable z_d: this restricts the datapoint to an association with one cluster only. In contrast, as with Latent Process Decomposition [26] and several other soft cluster models, in our model each data point C_d is associated with multiple latent variables $\{z_{dh} : h \in [1, \ldots, H]\}$. This means there is no implicit mutual exclusion of clusters assumption and C_d can be associated with multiple clusters. As a correspondence model this also means the data E_d can stochastically share C_d clusters through the correspondence latent variable y_{dg}.

2.2 CorrMAP: maximum a posterior approach

Having introduced the model, we now focus on approximation inference and parameter estimation for the correspondence model. Let the overall set of latent variables be denoted by $\mathcal{H} = \{\theta, z, y\}$ and model parameters by $\Theta = \{\alpha, \mu, \sigma, \widetilde{\mu}, \widetilde{\sigma}\}$. Then the target of model inference is to compute the posterior distribution $p(\mathcal{H}|E, C, \Theta) := p(E, C, \mathcal{H}|\Theta)/p(E, C|\Theta)$ and to learn the model parameters Θ. Unfortunately, this would involve computationally intensive estimation of the integral in the evidence $p(E, C|\Theta)$ and thus we will use *variational inference* instead [18] (we will discuss MCMC methods in the Conclusion).

The goal of variational inference is essentially to minimize the KL-divergence between the variational distribution $q(\theta, z, y)$ and posterior distribution $p(E, C, \theta, z, y|\Theta)$:

$$\arg \min_{\Theta, q \in \mathcal{Q}} \mathrm{KL}\big(q(\theta, z, y) \| p(E, C, \theta, z, y|\Theta)\big) \tag{2}$$

Since this expression is not convex, we employ the mean field approach [18]. The derivations are standard (see Jordan *et al* [18]) and referred to as variational EM-steps.

DOI: 10.2202/1544-6115.1441

We will briefly describe the general methodology. For simplicity, we assume that the latent variables $\mathcal{H}$ can be split into sub-variables $\mathcal{H}_i$ [18]. Then we choose the hypothesis family $\mathcal{Q}$ of *variational distributions* $q(\mathcal{H})$ to be a *fully factorized family*, that is, $q(\mathcal{H}) := \prod_i q(\mathcal{H}_i)$. Consequently, for the variational E-step, we conclude that the variational distribution of latent variables is given by [18, 5]:

$$q(\mathcal{H}_i) \propto \exp\Big(\mathbf{E}_{q^{\setminus i}}\big[p(E, \mathcal{H}|\Theta)\big]\Big) \tag{3}$$

where $q^{\setminus i}$ represents the distribution $\prod_{j\neq i} q(\mathcal{H}_j)$ and $\mathbf{E}_{q^{\setminus i}}$ denotes the expectation with respect to distribution $q^{\setminus i}$. For the M-Step, we take the derivative of the KL-divergence with respective to model parameter Θ and obtain the updates for Θ.

We now apply the above methodology to the correspondence model and obtain the following update equations. Let $\mathcal{H} = \{z, \theta, y\}$ be the set of latent variables and assume that the family of variational distributions $\mathcal{Q}$ takes the form:

$$q(\theta, z, y) = \Big[\prod_d q(\theta_d|\gamma_d)\Big]\Big[\prod_{d,h} q(z_{dh}|R_{dh})\Big]\Big[\prod_{d,g} q(y_{dg}|Q_{dg})\Big],$$

where $q(\theta_d|\gamma_d)$ is a Dirichlet distribution, $q(z_{dh}|R_{dh})$ and $q(y_{dg}|Q_{dg})$ are multinomial distributions. γ, R, Q are often called *variational parameters* and describe sufficient statistics of the variational distributions q. Equation (3) tells us that the optimal q can be found via the updates:

$$q(\theta|\gamma) = \prod_d q(\theta_d|\gamma_d) \propto \mathbf{E}_{z,y}\big[\log p(E, C, \theta, z, y|\Theta)\big], \tag{4}$$

$$q(z|R) = \prod_{d,h} q(z_{dh}|R_{dh}) \propto \mathbf{E}_{\theta,y}\big[\log p(E, C, \theta, z, y|\Theta)\big], \tag{5}$$

$$q(y|Q) = \prod_{d,g} q(y_{dg}|Q_{dg}) \propto \mathbf{E}_{\theta,z}\big[\log p(E, C, \theta, z, y|\Theta)\big]. \tag{6}$$

In summary, the estimation of the log of the joint distribution yields variational EM-type updates for variational and model parameters, as follows:

- **Variational E-step:**

$$\gamma_{dk} = \alpha_k + \sum_h R_{dhk}$$

$$R_{dhk} \propto \mathcal{N}(C_{hd}|\widetilde{\mu}_{hk}, \widetilde{\sigma}_{hk}) \exp\left(\Psi(\gamma_{dk}) - \Psi(\sum_j \gamma_{dj}) + \sum_g Q_{dgh} \log \mathcal{N}(E_{gd}|\mu_{gk}, \sigma^2_{gk})\right)$$

$$Q_{dgh} \propto \exp\Big(\sum_k R_{dhk} \log \mathcal{N}(E_{gd}|\mu_{gk}, \sigma^2_{gk})\Big).$$

where Ψ is the digamma function.

In the variational M-step we update the model parameters Θ. To this end, we just take the derivatives of the KL-divergence. The updates are listed as follows:

- **Variational M-step:**

$$\tilde{\mu}_{hk} = \frac{\sum_d R_{dhk} C_{hd}}{\sum_d R_{dhk}}, \qquad \tilde{\sigma}^2_{hd} = \frac{\sum_d R_{dhk}(C_{hd} - \tilde{\mu}_{hk})^2}{\sum_d R_{dhk}} \tag{7}$$

$$\mu_{gk} = \frac{\sum_{d,h} Q_{dgh} R_{dhk} E_{gd}}{\sum_{d,h} Q_{dgh} R_{dhk}}, \qquad \sigma^2_{gk} = \frac{\sum_{d,h} Q_{dgh} R_{dhk}(E_{gd} - \mu_{gk})^2}{\sum_{d,h} Q_{dgh} R_{dhk}} \tag{8}$$

For the updates for α, we use a Newton-Raphson method (see the Appendix of [6]). The gradient is given by: $\frac{\partial \mathcal{L}}{\partial \alpha_i} = D\big(\Psi(\sum_k \alpha_k) - \Psi(\alpha_i)\big) + \sum_d \big(\Psi(\gamma_{di}) - \Psi(\sum_k \gamma_{dk})\big)$, the Hessian is $H_{ij} = D(\Psi'(\sum_k \alpha_k) - \delta_{ij}\Psi'(\alpha_i))$. Hence, we have an iterative update procedure:

$$\alpha_{\text{new}} = \alpha_{\text{old}} - \big(H(\alpha_{\text{old}})\big)^{-1} \frac{\partial \mathcal{L}(\alpha_{\text{old}})}{\partial \alpha}.$$

We pursue the above iterative procedure until convergence of the KL-divergence (details are given in Appendix A: for discussion of numerical stability issues for the variational E-step update see Rogers *et al* [26] section 5.3). Since the latent variable θ_{dk} is the k-th cluster probability for sample d and its expectation with respect to the posterior distribution $q(\theta_d)$ is γ_{dk}, we could assign data E_d and C_d to cluster k using $k^* = \arg\max_k \gamma_{dk}$, for example. In Section 3, with a knowledge of the means and variances (μ, σ^2) and $(\tilde{\mu}, \tilde{\sigma}^2)$ for each cluster, we can use statistical scores to perform gene-ranking and

DOI: 10.2202/1544-6115.1441

thus find abnormally expressing genes or microRNA that are grouped in the same cluster. Following our earlier practice [26], we can choose the appropriate number of clusters using cross-validation on the predictive likelihood (see Appendix A for details).

We end this subsection with some comments. The above method can also handle cases where some values E_{gd} or C_{hd} are missing by omitting corresponding contributions in the M-step updates and corresponding parameters Q_{dgk} and R_{dgh}. In the original correspondence model of Blei *et al* [5] clustering was performed over samples. Here we are more interested in clustering over samples and trying to find a linkage between features (e.g. microRNA and genes). Unfortunately, whereas a direct linkage is calculable in the original correspondence model, $p(E_{gd}|C_{hd})$ is not meaningfully calculable here. Thus the model proposed here gives a picture of altered features within each cluster but does not individually link these: such a direct linkage would require methods outside the algorithm such as correlation analysis.

2.3 CorrVB: a variational Bayes approach

In the above maximum a posteriori approach, a computationally expensive cross validation study is required to infer the appropriate number of clusters. This involves setting aside a certain percentage of the data and then estimating the parameters on the remaining data. A model accuracy score is then found from the estimated likelihood on left-out data (refer to Figure 3 which shows the log likelihood estimated for the breast cancer data). Also, this variational inference approach only gives point estimates for $\{\mu, \sigma, \widetilde{\mu}, \widetilde{\sigma}\}$. An alternative approach is a variational Bayesian method which allows us to estimate the full posterior distribution in place of point estimates. Another advantage of a variational Bayesian approach over a maximum a posteriori solution is that there is an inbuilt mechanism for model comparison, which can be performed easily without use of cross validation data.

We now turn our attention to the description of variational Bayesian inference for our correspondence model. To this end, we further regard $\Xi = \{\mu, \sigma, \widetilde{\mu}, \widetilde{\sigma}\}$ as latent variables. Specifically, we further assume their prior distributions as follows. Let

$$p(\mu|m_0, v_0) = \prod_{g,k} \mathcal{N}(\mu_{gk}|m_0, v_0), \quad p(\widetilde{\mu}|m_0, v_0) = \prod_{h,k} \mathcal{N}(\widetilde{\mu}_{hk}|m_0, v_0),$$

and

$$p(\beta|a_0, b_0) = \prod_{g,k} \Gamma(\beta_{gk}|a_0, b_0), \quad p(\widetilde{\beta}|a_0, b_0) = \prod_{h,k} \Gamma(\widetilde{\beta}_{hk}|a_0, b_0)$$

where the Gamma distribution is defined by $\Gamma(x|a_0, b_0) = x^{a_0-1}e^{-\frac{x}{b_0}}/\Gamma(a_0)b_0^{a_0}$.

For fixed α, the variational Bayesian (ensemble learning) method (see e.g. [2]) aims to find an approximate posterior distribution $q \in \mathcal{Q}$ to the true posterior distribution $p(\theta, z, y, \Xi|E, \alpha))$, i.e.

$$\min_{q\in\mathcal{Q}} \mathrm{KL}\big(q(\theta, z, y, \Xi)\|p(\theta, z, y, \Xi|C, E, \alpha))\big).$$

Note, for any *variational distribution* $q(\theta, z, y, \Xi)$, that

$$\begin{aligned}\log p(E|\alpha) &= \log \int \sum_Z p(C, E, \theta, z, y, \Xi|\alpha) d\theta dz dy d\Xi \\ &= \mathbb{E}_q\Big[\log \tfrac{p(C,E,\theta,z,y,\Xi|\alpha)}{q(\theta,z,y,\Xi)}\Big] + \mathrm{KL}\big(q(\theta, z, y, \Xi)\|p(\theta, z, y, \Xi|C, E, \alpha))\big).\end{aligned} \tag{9}$$

Since $p(E)$ is a constant, our optimization target is equivalently reduced to maximizing the *free-energy* lower bound defined by

$$\max_q \mathcal{F}_{\mathcal{K}}(q|\alpha) := \max_q \mathbb{E}_q\Big[\log \frac{p(C, E, \theta, z, y, \Xi|\alpha)}{q(\theta, z, y, \Xi)}\Big]. \tag{10}$$

If we have no restriction on variational distributions q, then the maximizer of the free energy bound is trivially the true posterior which is already assumed intractable. Hence, we should introduce the *hypothesis family* $\mathcal{Q}$ where the variational posterior distributions $q(\theta, Z, \Theta)$ exist. For simplicity, we assume that the overall latent variables $\mathcal{H} = \{\theta, z, y, \mu, \sigma, \widetilde{\mu}, \widetilde{\sigma}\}$ can be partitioned into disjoint subsets $\{\mathcal{H}_i\}$. As before, we choose the hypothesis family $\mathcal{Q}$ of *variational distributions* $q(\mathcal{H})$ to be a *fully factorized family*, that is, $q(\mathcal{H}) := \prod_i q(\mathcal{H}_i)$. Consequently, in analogy to the E-step in MAP inference the variational distribution of latent variables is given by [18, 2]:

$$q(\mathcal{H}_i) \propto \exp\Big(\mathbf{E}_{q^{\setminus i}}\big[\log p(C, E, \theta, z, y, \Xi|\alpha)\big]\Big) \tag{11}$$

where $q^{\setminus i}$ represents the distribution $\prod_{j\neq i} q(\mathcal{H}_j)$. Specifically, the variational posterior distribution can be represented by their corresponding *variational parameters* as follows.

$$\begin{aligned}q(\theta, z, y, \Xi) &= \Big[\prod_d q(\theta_d|\gamma_d)\Big]\Big[\prod_{d,h} q(z_{dh}|R_{dh})\Big]\Big[\prod_{d,g} q(y_{dg}|Q_{dg})\Big] \\ &\quad\times\Big[\prod_{h,k} q(\widetilde{\mu}_{hk}|\widetilde{m}_{hk}, \widetilde{v}_{hk}) q(\widetilde{\beta}_{hk}|\widetilde{a}_{hk}, \widetilde{b}_{hk})\Big] \\ &\quad\times\Big[\prod_{g,k} q(\mu_{gk}|m_{gk}, v_{gk}) q(\beta_{gk}|a_{gk}, b_{hk})\Big].\end{aligned}$$

DOI: 10.2202/1544-6115.1441

Since the Gamma distribution is the conjugate prior of the Normal distribution, the variational posterior distribution on the latent variables μ and β are respectively the Normal distribution and the Gamma distribution, likewise for $\widetilde{\mu}$ and $\widetilde{\beta}$. The detailed updates are listed on Appendix B.

The above inference method assumes that the number of clusters is fixed. We now turn our attention to inferring the number of clusters in the variational Bayesian framework (see e.g. [2]). Given a maximum possible number of clusters M, the target is to determine the optimal number of clusters $\mathcal{K} \in [2, M]$. The variational Bayes approach provides an appealing way to infer correct model complexity. Let $p(\mathcal{K})$ be the prior distribution on the number of clusters. Recall that the joint distribution $p(C, E, \theta, z, y, \Xi|\alpha)$ given by equation (1) is dependent on $\mathcal{K}$, the number of clusters, so we could denote it as $p(C, E, \theta, z, y, \Xi|\mathcal{K}, \alpha)$. Thus, regarding $\mathcal{K}$ is a latent variable, we can consider the joint distribution

$$p(C, E, \theta, z, y, \Xi, \mathcal{K}|\alpha) = \prod_{\mathcal{K}=1}^{M} p(C, E, \theta, z, y, \Xi|\mathcal{K}, \alpha)p(\mathcal{K}).$$

In analogy to the above inference procedure, we assume that the variational posterior distribution $q(\theta, z, y, \Xi, \mathcal{K})$ is of the form $q(\theta, z, y, \Xi, \mathcal{K}) = q(\theta, z, y, \Xi)q(\mathcal{K})$ where the updates for $q(\theta, z, y, \Xi)$ (conditioned on a specific $\mathcal{K}$) are already provided by the previous updates. The posterior distribution for $\mathcal{K}$ turns out to be

$$q(\mathcal{K}) \sim \exp(\mathcal{F}_{\mathcal{K}}).$$

Here, the free energy $\mathcal{F}_{\mathcal{K}}$ is defined by equations (10). Usually, the prior distribution for $\mathcal{K}$ is chosen to be a uniform distribution, i.e. $p(\mathcal{K}) = 1/M$. In this case, the optimal number of clusters $\mathcal{K}^*$ is determined by maximum of the free energy bound:

$$\mathcal{K}^* = \arg\max_{\mathcal{K}} \mathcal{F}_{\mathcal{K}}.$$

We can also update the parameter α together with the latent variables $\mathcal{H}$. Specifically, we maximize the lower (free energy) bound of the log likelihood $\log p(E|\alpha)$ with respect to both the latent variables $\mathcal{H}$ and α:

$$\max_{q,\alpha} \mathcal{F}_{\mathcal{K}}(q|\alpha) := \max_{q,\alpha} \mathbb{E}_q\Big[\log \frac{p(C, E, \theta, z, y, \Xi|\alpha)}{q(\theta, z, y, \Xi)}\Big]. \tag{12}$$

Just like the MAP approach, the updates for the latent variables $\mathcal{H}$ fall under the E-step. Keeping a fixed variational posterior distribution q in the M-step,

we can use a Newton-Raphson method to update α by

$$\alpha_{\text{new}} = \arg\max_{\alpha} \mathcal{F}(q|\alpha).$$

where, in a similar fashion to the MAP method, the updates for α can be solved by the Newton-Raphson method $\alpha_{\text{new}} = \alpha_{\text{old}} - \left(H(\alpha_{\text{old}})\right)^{-1} \frac{\partial \mathcal{F}(q|\alpha^{\text{old}})}{\partial \alpha}$.

3 Experiments

In this section we will numerically validate the proposed correspondence model using *corrMAP* to present our results and *corrVB* to confirm them. First we demonstrate that the correspondence model performs as expected on artificially generated data, where the cluster structure and sample labels are known. In addition, in Section 3.1, we consider an expression array dataset for *S. Cerevisiae* which illustrates a biological context in which correspondence models would be relevant. We compare against three other clustering methods. We then consider the breast cancer example referred to earlier where gene expression array data is assumed dependent on microRNA data. The results for *corrVB* validate the results for *corrMAP* and the results are consistent with previous breast cancer studies.

3.1 Comparison with other clustering methods

To validate performance we first generated artificial datasets. Data for C was randomly generated to give three distinct clusters (consisting of 10 samples per cluster with 10 features per sample). Then the data in E was generated per sample in C, so that it had blocks of features positively correlated to features in C. The number of features in E (corresponding to the index G) was varied between $N = 2, \ldots, 10$ times the size of dataset C. Thus each vector in C had from 2 to 10 replicate features in E with each such feature perturbed by a small Gaussian random deviate. Since the sample labels of our artificially generated data were known, we were able to use the Jacard score to compare our clusterings with the correct labels and thereby validate our results. The Jaccard score J is used to compare clusterings, or to compare a clustering with the correct labels. If we let n_{11} denote the number of point pairs correctly placed together in the clustering, n_{01} the number of incorrectly identified pairs and n_{10} the number of missed pairs, then $J = n_{11}/(n_{11} + n_{01} + n_{10})$, where $0 \leq J \leq 1$, with $J = 1$ indicating a perfect clustering. In Figure 2 (left) we present a bar plot of the Jaccard scores obtained. Apart from the proposed correspondence

DOI: 10.2202/1544-6115.1441

model, we also amalgamated C and E and performed spectral clustering and k-means clustering on the amalgamated dataset. To create this amalgamated dataset, both datasets were normalised to zero mean, unit variance and combined into a single column vector per sample. All of the models perform better for a small N, with the correspondence model (*corrMAP*) consistently outperforming the rest. As N increases, the difference in Jaccard scores diminishes considerably as it is hard for any of the models to pick up the correct clustering. We also tried a novel joint mixture model (*JMM*) which assumes the two datasets to be independent and extracts a joint clustering without the correspondence assumption (details outlined in Appendix C). We include *JMM* for comparison for two reasons. First, although it is more similar to the correspondence model than the other two methods that use the amalgamated data, it does not perform as well, and this highlights the importance of modeling the correspondence between the datasets when appropriate. Second, like the correspondence model, *JMM* outputs a normalised γ_{dk} values (normalised over k) which represents the confidence in the assignment of sample d to cluster k. In Figure 2 (right) we show bar plots for these confidence measures for both *corrMAP* and *JMM*. The confidence values for the correspondence model are consistently higher than *JMM*.

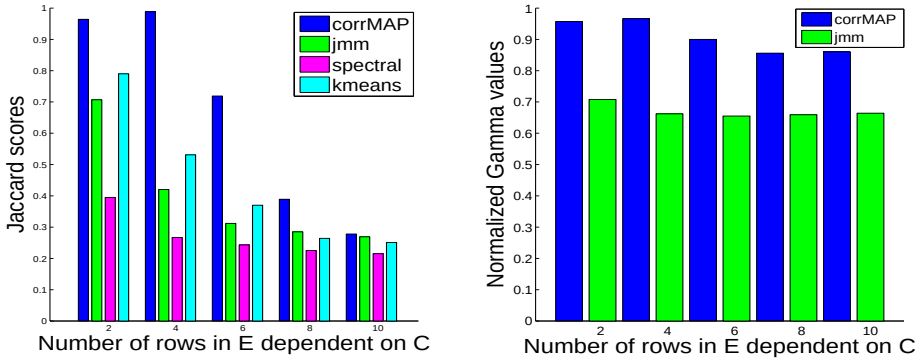

Figure 2: Average Jaccard scores on the 3 cluster artifical dataset (left) and associated confidence measures (right) using the *corrMAP* algorithm and *JMM*, a joint mixture model outlined in Appendix C.

As a second example, we used microarray expression data for *S. cerevisiae* from a series of experiments run by Middendorf *et al* [24]. These authors identified a strong regulating factor, *USV1*, which was believed to influence

up to 305 other genes in the dataset. In this second example, the C data now consists of only one gene, *USV1*, while E comprises the 305 regulated genes. Though an extreme example, since C only has one value per sample, this is a context where the application of a correspondence model makes sense since the significance of C is maintained by this model, but would be lost if we amalgamated the datasets, for example.

The samples were derived from three groups of experiments: heat shock, nitrogen depletion and a set of stationary phase experiments used as a time-zero reference. A PCA plot (not illustrated) suggested these were reasonably well defined groupings. We ran the correspondence and joint mixture models and they correctly classified all three groups ($J = 1$, based on the highest predictive log-likelihood solution after 30 random initialisations). We also used k-means clustering and spectral clustering on the amalgamated dataset. Both k-means clustering and spectral clustering can give different results depending on the start point, hence we investigated performance over 100 restarts. k-means clustering gave $J = 1$ with 62 restarts from the 100 with an overall average Jaccard score of 0.81. Spectral clustering correctly classified ($J = 1$) 81 from 100 restarts with an average Jaccard score of 0.90. These results are not that surprising since C is considerably smaller than E in size so the significance of C is lost when the two datasets are amalgamated together.

3.2 Evaluation on a real-life dataset: breast cancer

For the two examples given above we have argued that there are instances where joint modelling of the data is more appropriate than clustering on an amalgamated dataset. We now extend the discussion to a real-life example in cancer biology to illustrate the extra biological insights provided by the correspondence model. We will show that the results which emerge are consistent with previous findings. In addition, we have not commented so far on model complexity: how many clusters are present in the data. We will show that the estimated log-likelihood on hold-out data provides a principled approach to finding the correct model complexity for *corrMAP*.

We applied our models to a dataset consisting of two types of data derived from the same patients. The first data set, C, consisted of microRNA expression data from 78 primary human breast tumors using a bead-based array to identify 133 microRNA found in normal and breast tumors [7]. The second set of data, E, comprised gene expression data for the same 78 patients. In both cases, the data was normalized to zero mean and unit variance.

The first goal was to determine the optimal number of clusters. To do so, we first performed a cross validation study on the predictive log likelihood

DOI: 10.2202/1544-6115.1441

using *corrMAP* (see Appendix A). We held out 8 datapoints as test data and the remaining 70 datapoints were used to construct the model: performance was averaged over 10 random partitionings of the data into training and test data. The log-likelihood on the hold-out data was calculated using the model obtained from training data. Figure 3 (left) shows the corresponding log-likelihood curve for the correspondence model. A 5 cluster model appears optimal: if more than 5 clusters are used overfitting occurs and the log-likelihood falls. To confirm this result we then used the variational Bayes method of section 2.3. In this approach, we do not need hold-out data to estimate a log-likelihood. Instead, a free energy expression is used. In Figure 3 (right) we give the corresponding curve for the free energy which likewise gives a peak at 5 clusters indicating that there are at least 5 principal subtypes of breast cancer.

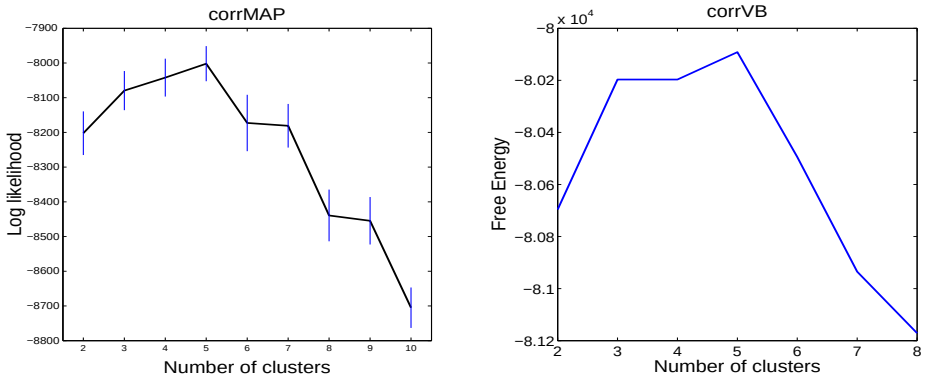

Figure 3: Estimated log likelihood on heldout data versus number of clusters for *corrMAP* (left) and free energy versus number of clusters for *corrVB* (right). The peaks in both curves indicate the most probable number of clusters (cancer subtypes) in the data. *CorrVB* does not use cross-validation data.

As remarked in section 2.2 we can assign sample d to cluster k using $k^* = \arg\max_k \gamma_{dk}$. Based on available survival data, we can therefore derive Kaplan Meier plots for the 5 indicated subtypes. These are given in Figure 4 where we see that subtype 3 is most aggressive, subtypes 4 and 5 less aggressive, and subtypes 1 and 2 are largely indolent.

We can also derive density estimates, quantifying the distribution of gene expression data values within subtypes. Using *corrMAP*, for example,and using mean μ_{gk} and standard deviation σ_{gk} for gene g within cluster k, we

present the density distributions for some genes in Figure 5. *FOXA1* and *FOXC1* have very distinctive distributions for the subtype labelled *Cl5*: while *FOXA1* underexpresses, *FOXC1* overexpresses within this subtype. *ERBB2* and *GRB7* overexpress in subtype *Cl3*: there is a well documented *ERBB2+* subtype of breast cancer [30].

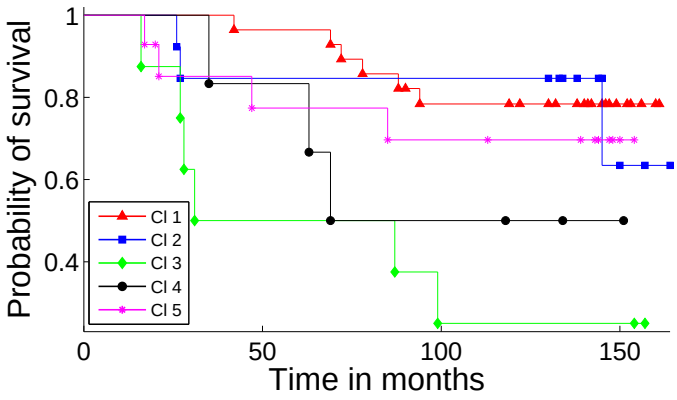

Figure 4: Kaplan Meier plot for the corrrespondence model (*corrMAP*) showing patient survival rate changing over time.

Abnormally expressed genes can be identified using a Fisher score $|\mu_{g_1k} - \mu_{g_2k}|/\sqrt{\sigma^2_{g_1k} + \sigma^2_{g_2k}})$, for example. However, this score tends to overlook genes with large spreads such as *FOXC1* in *Cl5* of Figure 5. Thus, we used a rank-based Mann-Whitney score instead to find genes abnormally expressing within one subtype relative to the other subtypes. In Table 1 we list the 20 top-ranked genes by significance for the 5 subtypes resolved by the correspondence model (*corrMAP*). The genes listed under *CL5* appear to be biologically significant. The X box-binding protein, *XBP1*, is believed to be regulated by *FOXA1* [11]. The biological importance of *FOXA1* is also apparent from some recent results reported in the literature: a substantial number of estrogen response elements (EREs) have associated binding sites for *FOXA1* [11, 20]. Similarly *GATA3* has associated co-expression with *XPBP1* and *ESR1* [19]. We also note that these genes have been previously identified and discussed by other authors [12, 13].

DOI: 10.2202/1544-6115.1441

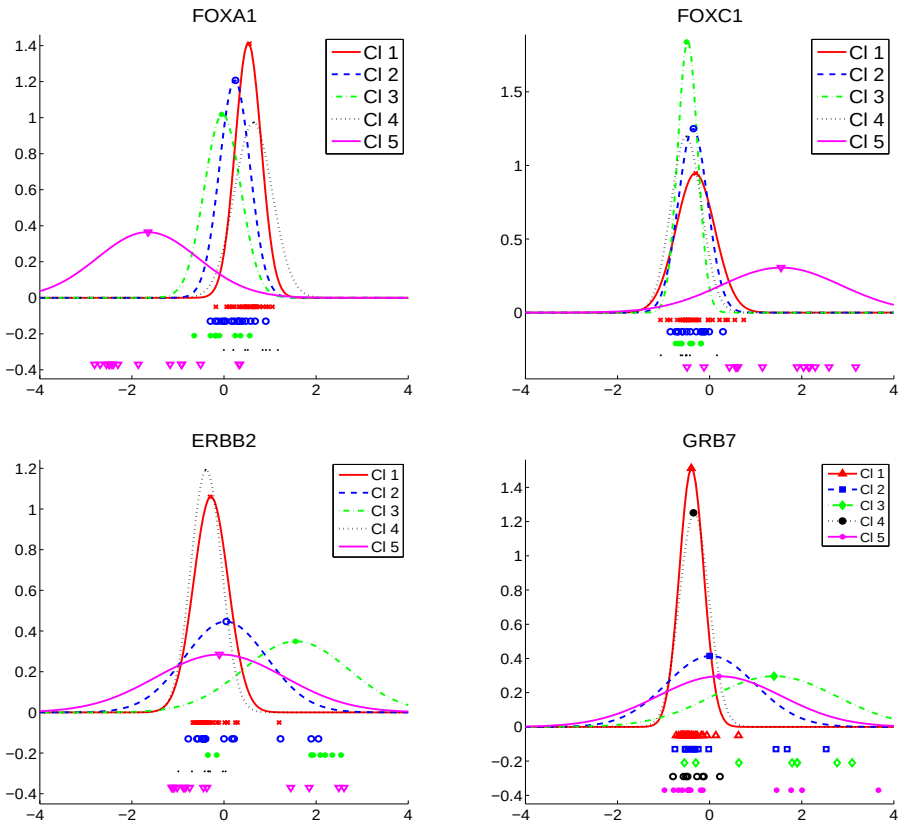

Figure 5: Density distribution plots for four genes using *corrMAP*. Gene expression values are given at the base of each plot. A sample d is assigned to the cluster k depending on the largest value of the confidence measure, γ_{dk}. The Gaussian distributions are derived from (μ_{gk}, σ_{gk}) in equations (8).

Next we need to determine if the genes listed in Table 1 are consistent with previous findings. In previous work we investigated a number of microarray datasets for breast cancer, and the results in Table 1 are consistent with these past findings. In Carrivick *et al* [10] we investigated four microarray datasets using a Bayesian variational method [26]. This analysis indicated 4 or 5 principal subtypes of breast cancer. It clearly showed a recognised *ERBB2+*, *ESR1-* subtype of breast cancer, typified by elevated expression of *ERBB2* and *GRB7* [30], associated with the aggressive *Cl3* cluster presented here. A second subtype (*Cl5*) has a clear connection with the *basaloid* subtype of breast cancer [30]. In our study [10] we used a variational Bayes method

[1, 3] to investigate 7 datasets for primary breast carcinoma ([30, 33, 35] and a composite dataset of 614 samples [14, 25, 34, 36] which all used the Affymetrix U133A chip (see [9] for full details). This gave the genetic signature of the basaloid subtype in Table 2 which has a good match to the signature under *Cl5* in Table 1.

Cl 1	Cl 2	Cl 5	Cl 4	Cl 3
UBE2C	*COL11A1*	**GATA3**	*CTGF*	*GSDML*
CDC20	*TIMP3*	**FOXC1**	*RARRES1*	*ORMDL3*
POSTN	*AEBP1*	*STARD10*	*C1S*	**ERBB2**
CYBRD1	*COL10A1*	**MLPH**	*PRKACB*	*STARD3*
OGN	*PLAU*	*TOB1*	*FBLN2*	*FGFR4*
ADH1B	*MFAP5*	**AGR2**	*TNC*	**ESR1**
ADH1A	*COL12A1*	**FBP1**	*ACTA2*	*PERLD1*
CYP4X1	*MMP11*	*GPR160*	*CR598488*	*CTXN1*
COL10A1	*FN1*	*C10orf116*	*COL6A1*	*DQ582071*
TIMP3	*SULF1*	*BCAS1*	*SPON1*	**GRB7**
TK1	*COL8A1*	*DEGS2*	*ASS1*	*RAP1GAP*
SH3BGRL	*POSTN*	**XBP1**	*FLNA*	*C1S*
SUSD3	*NBL1*	*CRYAB*	*PKIB*	*U79293*
MIA	*DCN*	*EEF1A2*	*SBEM*	*PRSS8*
CPA3	*OGN*	*SLC39A6*	*abParts*	*C17orf37*
PPP1R3C	*GJB2*	*KRT19*	*FLJ42258*	*MFAP2*
SFRP1	*THBS2*	*GALNT6*	*CRISPLD2*	**TFF1**
ATP1B1	*ACTA2*	**FOXA1**	*BAMBI*	**CA12**
SLC40A1	*TBC1D9*	**GABRP**	*SYT13*	*TBC1D9*
CILP	*LOXL2*	*NPNT*	*IGHA2*	*CAPS*

Table 1: Top ranked genes by the Mann Whitney score for each subtype in Figure 4 using *corrMAP*. Some genes are presented in boldface because they are commented in the text or feature in Table 2.

A very similar story emerges if we use the variational Bayes approach outlined in section 2.3. We likewise find a cluster with genes *FOXC1*, *AGR2*, *FOXA1*, *GATA3*, *TFF1*, *MLPH*, *XBP1*, *GABRP* ranked in the top 20. Thus the genes highlighted by our method appear to be consistent with previous studies, consistent between the two correspondence algorithms and also biologically significant. The advantage of our proposed correspondence models is that we now have additional information about the role of microRNA within given subtypes. Interestingly, the genetic signature reported in Table 2 derives

DOI: 10.2202/1544-6115.1441

purely from clustering on gene expression data whereas the corresponding signature in column 3 of Table 1 derives from the clustering structure dictated by microRNA expression. This indicates an intimate relation between gene and microRNA expression, at least for this subtype.

Sorlie *et al* [30]	West *et al* [35]	Van t' Veer *et al* [33]	Composite
TFF3	*CRIP1*	*VGLL1*	**FOXA1**
XBP1	**XBP1**	**AGR2**	**AGR2**
FOXA1	**FOXA1**	**TFF3**	**XBP1**
GATA3	*CEBPD*	*ESR1*	*MLPH*
B3GNT5	*HSPA8*	**CA12**	*FLJ20174*
GALNT10	**GATA3**	**DSC2**	**CA12**
FBP1	*RARA*	**NAT1**	**GATA3**
DSC2	*CRYAB*	*EST*	*AK127020*
FOXC1	**GATA3**	**CDH3**	**CA12**
FOXC1	**FBP1**	**FOXC1**	**CA12**
FLT1	*KRT18*	*SCUBE2*	**GATA3**
FOXC1	*MSN*	**AR**	**AR**
GATA3	*TCEAL1*	*Corf7*	**TFF3**
SLC11A3	*SCNN1A*	*SLC7A2*	*ABAT*
SLC11A3	*NSEP1*	**GABRP**	**FBP1**
MGC27171	**CDH3**	*EST*	**DSC2**
NAT1	*BF*	**XPB1**	**GATA3**
MRPS14	**TFF3**	*BCMP11*	**CA12**
LOC51313	*Hu. clone 23948*	*VAV3*	*TFF1*
MGC10710	*FSCN1*	*EST*	**GABRP**

Table 2: A list of the top-ranked 20 genes distinguishing the basaloid subtype of breast cancer from an earlier study [10], derived from seven separate studies for breast cancer. The composite dataset of 614 samples is taken from [14, 25, 34, 36], which all used the Affymetrix U133A chip (data was amalgamated after normalising to zero mean, unit standard deviation for each component dataset). Repeat gene names in a column derive from multiple probes for the same gene. Genes presented in boldface appear in more than one column.

We also used the Mann-Whitney score to rank microRNA features. In Table 3 we give the mean values, $\tilde{\mu}_{hk}$, for the 10 top-ranked microRNA expressions using the *corrMAP* model clusters. As with this Table, a plot of all microRNA expression values, averaged per cluster, indicates substan-

tial differences between the microRNA expression profiles between subtypes. The most aggressive subtype (Cl3 in Figure 4) appears to be linked with extensive abnormally high expression of microRNA, followed by Cl4 and Cl5 which have small subsets of microRNAs with abnormally high expression. A significant number of these microRNA have been previously associated with breast cancer. Thus *miR-214* can induce cell death resistance through targeting the PTEN/Akt pathway [17], *miR-21* is oncogenic [28] and *let-7* is listed as tumour-suppressive [22]. *miR-126* and *miR-335* have been associated with suppression of breast cancer metastasis[31, 37]. Sempere *et al* [27] reported altered expression of *miR-145* and *miR-21* in various epithelial cell subpopulations of breast cancer. Aberrant hypermethylation leading to inactivation of *miR-152* has been reported for breast cancer [23]. Finally, a number of these microRNAs, such as *miR-15a*, *miR-16*, *miR-21*, *miR-125b* and *miR-145*, have reported altered expression for other mammalian mammary carcinomas[8].

Cl1		Cl2		Cl3		Cl4		Cl5	
miR-505	0.38	miR-137	0.26	**miR-152**	1.13	miR-30b	0.66	miR-199a	0.62
miR-181c	0.37	miR-133a	0.19	miR-342	0.99	miR-15b	0.63	miR-99a	0.57
miR-142-5p	0.36	miR-9	0.19	miR-29a	0.98	**miR-15a**	0.60	miR-199b	0.55
miR-185	0.31	miR-9	0.18	miR-331	0.96	miR-30c	0.57	miR-199a	0.55
miR-203	0.31	miR-18a	0.08	**miR-214**	0.95	miR-195	0.55	miR-214	0.47
miR-200a	0.30	miR-128b	0.07	miR-199b	0.94	**miR-16**	0.49	miR-100	0.47
miR-183	0.29	miR-138	0.06	**miR-126**	0.90	**miR-21**	0.49	miR-130a	0.45
miR-509	0.29	miR-211	0.03	**miR-145**	0.89	miR-20a	0.45	miR-382	0.43
miR-107	0.29	**miR-335**	0.03	miR-24	0.89	miR-30a-3p	0.45	**miR-125b**	0.42
miR-93	0.29	miR-429	0.02	miR-27a	0.88	miR-210	0.44	**let-7b**	0.40

Table 3: Top ranked microRNA by Mann Whitney score using *corrMAP*. The numbers indicate the (normalized) mean microRNA expression for the samples in that cluster. Entries in boldface are discussed in the text.

4 Conclusion

In this paper we have introduced a correspondence model for unsupervised learning with two data types that are presumed to be related or dependent. Using a predictive likelihood estimate or a free energy term we can find the appropriate number of clusters in the data. The proposed methods can handle missing values. In Sections 3.1 we showed that clustering on amalgamated data gave inferior performance, and that modeling the two datasets independently gave worse results than the correspondence model. In addition, *corrMAP* proved superior on artifical datasets when compared to a joint mixture model.

DOI: 10.2202/1544-6115.1441

In Section 3.2 we gave an extended discussion of an application to breast cancer biology: the results for the correspondence model appeared consistent with previous findings and biologically plausible. Furthermore, by incorporating microRNA expression data in addition to gene expression data, the model may give possible new insights into dysregulation of microRNA expression associated with individual breast cancer subtypes.

The methods proposed here can be extended in various ways. Firstly, we have presented our models using two data sets of the same type: continuous valued data (e.g. gene expression data) which can be approximately modelled using a Gaussian distribution. However, the model can easily be adjusted to accommodate discrete data, using a multinomial or Poisson distribution to model one or both types of data. We could, of course, also use a variety of MCMC methods. Although MCMC proved to be too computationally intensive for determining the model complexity with the large expression array datasets here, it could certainly be usefully deployed for smaller datasets.

Appendices

A Lower Bound and Predictive Likelihood for corrMAP

In this appendix we outline the computation of the KL-divergence and predictive likelihood for the first correspondence model, *corrMAP*.

Lower bound (negative KL-divergence):

The lower bound for the log likelihood (denoted $\mathcal{L}$) equals the negative KL-divergence:

$$\mathcal{L} = \int \sum_{z,y} q(z,y,\theta) \log \frac{p(C,E,z,y,\theta|\Theta)}{q(z,y,\theta)} d\theta = -\mathrm{KL}\big(q(\theta,z,y) \| p(\theta,z,y|E,C,\Theta)\big).$$

Estimation of the log joint distribution gives:

$$\begin{aligned}\mathcal{L} = \; & D\Big[\log\Gamma(\textstyle\sum_k \alpha_k) - \sum_k \log\Gamma(\alpha_k)\Big] - \sum_d \Big[\log\Gamma(\textstyle\sum_k \gamma_{dk}) - \sum_k \log\Gamma(\gamma_{dk})\Big] \\ & + \textstyle\sum_{d,k} \Big[\big(\alpha_k - \gamma_{dk} + \sum_h R_{dhk}\big)\big(\psi(\gamma_{dk}) - \psi(\sum_j \gamma_{dj})\big)\Big] \\ & + \textstyle\sum_{d,h,k} \Big[R_{dhk}\Big(\log \mathcal{N}(c_{hd}|\tilde{\mu}_{hk}, \tilde{\sigma}^2_{hk}) - \log R_{dhk}\Big)\Big] \\ & + \textstyle\sum_{d,g,h} Q_{dgh}\Big[\big(\sum_k R_{dhk} \log \mathcal{N}(e_{gd}|\mu_{gk}, \sigma^2_{gk})\big) - \log Q_{dgh}\Big].\end{aligned}$$

Predictive likelihood:

First, marginalizing the joint probability given by equation (1) with respect to z gives

$$p(C,E,y|\Theta) = \prod_d \int_{\theta_d} \sum_z p(C_d,E_d,z_d,y_d,\theta_d|\Theta)d\theta_d$$
$$= \prod_d \int_{\theta_d} \prod_h \Big[\sum_k \theta_{dk}\mathcal{N}(C_{hd}|\tilde{\mu}_{hk},\tilde{\sigma}^2_{hk}) \prod_g \Big(\frac{1}{H}\mathcal{N}(E_{gd}|\mu_{gk},\sigma^2_{gk})\Big)^{y_{dg,h}}\Big]d\theta.$$

However, further marginalizing with respect to y will lead to very intensive computation since the dimension of expression gene g is usually large. Hence, we are forced to consider approximation of the test likelihood. To this end, we replace the untractable term $\prod_g \Big(\frac{1}{H}\mathcal{N}(E_{gd}|\mu_{gk},\sigma^2_{gk})\Big)^{y_{dg,h}}$ by its average

$$\prod_g \Big(\frac{1}{H}\mathcal{N}(E_{gd}|\mu_{gk},\sigma^2_{gk})\Big)$$

Consequently, we have the following approximation to the likelihood

$$p(C,E|\Theta) \approx \prod_d \int_{\theta_d} \prod_h \Big[\sum_k \theta_{dk}\mathcal{N}(C_{hd}|\tilde{\mu}_{hk},\tilde{\sigma}^2_{hk}) \prod_g \Big(\frac{1}{H}\mathcal{N}(E_{gd}|\mu_{gk},\sigma^2_{gk})\Big)\Big]d\theta$$

The integral with respect to θ can be further approximated by a sampling method described in Blei and Jordan [5].

B Updates for CorrVB

Here we list the update equations for the variational Bayesian inference for the correspondence model.

- For θ, we have that $q(\theta|\gamma) \propto \exp\Big(\mathbf{E}_{q(z,y,\Xi)}\big[\log p(C,E,\theta,z,y,\Xi|\alpha)\big]\Big)$ which yields
$$\gamma_{dk} = \alpha_k + \sum_h R_{dhk} \tag{13}$$

- For z, we have that

$$q(z|R) = \prod_{d,h} q(z_{dh}|R_{dh}) \propto \exp\Big(\mathbf{E}_{q(\theta,y,\Xi)}\big[\log p(C,E,\theta,z,y,\Xi|\alpha)\big]\Big).$$

DOI: 10.2202/1544-6115.1441

Consequently

$$\begin{aligned} R_{dh,k} \;\propto\; \exp\Bigg[& \psi(\gamma_{dk}) - \psi(\sum_j \gamma_{dj}) + \frac{1}{2}(\psi(a_{hk}) + \log b_{hk}) \\ & -\frac{1}{2}\left((C_{dh} - m_{hk})^2 + \frac{1}{v_{hk}}\right) a_{hk} b_{hk} + \frac{1}{2}\sum_g Q_{dgh}\psi(a_{gk}) \\ & + \log b_{gk} - a_{gk} b_{gk}\left((E_{dg} - m_{gk})^2 + \frac{1}{v_{gk}}\right)\Bigg] \end{aligned}$$

with the normalization $\sum_k R_{dh,k} = 1$ for any d, h.

- For the latent variable y, from the equation

$$q(y|Q) = \prod_{d,g} q(y_{dg}|Q_{dg}) \propto \exp\left(\mathbf{E}_{q(\theta,z,y,\Xi)}\left[\log p(C, E, \theta, z, y, \Xi|\alpha)\right]\right)$$

we get

$$\begin{aligned} Q_{dg,h} \;\propto\; \exp\Bigg[& \log\phi_{gh} + \frac{1}{2}\sum_k R_{dhk}\left(\psi(a_{gk}) + \log b_{gk}\right. \\ & \left. -\frac{a_{gk} b_{gk}}{v_{gk}}\left((E_{dg} - m_{gk})^2\right)\right)\Bigg] \end{aligned}$$

with the normalization $\sum_h Q_{dg,h} = 1$ for any d, g.

- For the latent variables $\widetilde{\mu}$ and $\widetilde{\beta}$, we have

$$\begin{aligned} q(\widetilde{\mu}|\widetilde{m}, \widetilde{v}) &= \prod_{h,k} \mathcal{N}(\widetilde{\mu}_{hk}|\widetilde{m}_{hk}, \widetilde{v}_{hk}) \\ &\propto \exp\left(E_{q(\theta,z,y,\mu,\beta,\widetilde{\beta})}\left[\log p\left(C, E, \theta, z, y, \Xi|\alpha\right)\right]\right), \end{aligned}$$

and

$$\begin{aligned} q(\widetilde{\beta}|\widetilde{a}, \widetilde{b}) &= \prod_{h,k} \Gamma(\widetilde{\beta}_{hk}|\widetilde{a}_{hk}, \widetilde{b}_{hk}) \\ &\propto \exp\left(E_{q(\theta,z,y,\mu,\beta,\widetilde{\mu})}\left[\log p\left(C, E, \theta, z, y, \Xi|\alpha\right)\right]\right) \end{aligned}$$

Consequently,

$$\widetilde{m}_{hk} = \frac{m_0 v_0 + \left(\sum_d C_{dh} R_{dhk}\right) \widetilde{a}_{hk} \widetilde{b}_{hk}}{\widetilde{v}_{hk}}, \quad \widetilde{v}_{hk} = v_0 + \widetilde{a}_{hk} \widetilde{b}_{hk} \sum_d R_{dhk} \quad (14)$$

$$\widetilde{a}_{hk} = a_0 + \frac{1}{2} \sum_d R_{dhk}, \quad \widetilde{b}_{hk}^{-1} = \frac{1}{b_0} + \frac{1}{2} \sum_d R_{dhk} \left[\left(C_{dh} - \widetilde{m}_{hk} \right)^2 + \frac{1}{\widetilde{v}_{hk}} \right] \quad (15)$$

- For μ and β we have that

$$\begin{aligned} q(\mu|m,v) &= \prod_{g,k} \mathcal{N}(\mu_{gk}|m_{gk}, v_{gk}) \\ &\propto \exp\left(E_{q(\theta,z,y,\beta,\widetilde{\mu},\widetilde{\beta})} \left[\log p\left(C, E, \theta, z, y, \Xi|\alpha\right) \right] \right) \end{aligned}$$

and

$$\begin{aligned} q(\beta|a,b) &= \prod_{g,k} \Gamma(\beta_{gk}|a_{gk}, b_{gk}) \\ &\propto \exp\left(E_{q(\theta,z,y,\mu,\widetilde{\mu},\widetilde{\beta})} \left[\log p\left(C, E, \theta, z, y, \Xi|\alpha\right) \right] \right) \end{aligned}$$

Consequently,

$$m_{gk} = \frac{m_0 v_0 + \left(\sum_{d,h} Q_{dg,h} R_{dh,k} \right) a_{hk} b_{hk}}{v_{gk}}, \quad (16)$$

$$v_{gk} = v_0 + a_{gk} b_{gk} \sum_{d,h} Q_{dg,h} R_{dh,k} \quad (17)$$

and

$$a_{gk} = a_0 + \frac{1}{2} \sum_{d,h} Q_{dg,h} R_{dh,k}, \quad (18)$$

$$b_{gk}^{-1} = \frac{1}{b_0} + \frac{1}{2} \sum_{d,h} Q_{dg,h} R_{dh,k} \left[\left(E_{dg} - m_{gk} \right)^2 + \frac{1}{v_{gk}} \right] \quad (19)$$

DOI: 10.2202/1544-6115.1441

In MAP type II, we also update the Dirichlet parameter by the following equation

$$\begin{aligned}\widehat{\alpha} &= \arg\max \left[D \log \Gamma \left(\sum_k \alpha_k \right) - D \sum_k \log\{\Gamma(\alpha_k)\} \right. \\ &\quad \left. + \sum_{d,k} (\alpha_k - 1) \left(\psi(\gamma_{dk}) - \psi(\sum_j \gamma_{dj}) \right) \right]\end{aligned}$$

C A Joint Mixture Model

For the joint mixture model (*JMM*) mentioned in Section 3.1, the functional relationship between the different data sets is modelled via a jointly clustering Dirichlet distribution. Samples in the different data sets are generated separately. This model is described as follows:

For a fixed data index d for both E ($G \times D$ matrix) and C ($H \times D$ matrix)

1. Prior distributions: $\theta_d \sim \mathrm{Dir}_K(\alpha)$
2. Generate C_d:
   (a) Choose cluster for C_{hd}: $\tilde{z}_{dh} \sim \mathrm{Multi}(\theta_d)$
   (b) Sample $C_{hd} \sim \mathcal{N}(C_{hd}|\tilde{\mu}_{h\tilde{z}_{dh}}, \tilde{\sigma}_{h\tilde{z}_{dh}})$ where $\mathcal{N}(C_{hd}|\tilde{\mu}, \tilde{\sigma}^2)$ denotes a normal distribution with mean $\tilde{\mu}$ and variance $\tilde{\sigma}^2$.
3. Generate E_d:
   (a) Choose cluster for E_{gd}: $z_{dg} \sim \mathrm{Multi}(\theta_d)$
   (b) Sample $E_{gd} \sim \mathcal{N}(E_{gd}|\mu, \sigma, z) = \mathcal{N}(E_{gd}|\mu_{gz_{dg}}, \tilde{\sigma}^2_{gz_{dg}})$

Using the notation $\Theta = \{\alpha, \mu, \sigma, \tilde{\mu}, \tilde{\sigma}\}$, the joint distribution for a given data index d is given by:

$$\begin{aligned}p(C_d, E_d, \tilde{z}_d, z_d, y_d, \theta_d|\Theta) &= p(\theta_d|\alpha) \textstyle\prod_h \left[p(\tilde{z}_{dh}|\theta_d) \mathcal{N}(C_{hd}|\tilde{\mu}_{hz_{dh}}, \tilde{\sigma}^2_{hz_{dh}}) \right] \\ &\quad \times \textstyle\prod_g p(z_{dg}|\theta_d) \mathcal{N}(E_{gd}|\mu_{gz_{dg}}, \tilde{\sigma}^2_{gz_{dg}})\end{aligned}$$

The overall joint distribution is then given by:

$$p(C, E, \tilde{z}, z, y, \theta|\Theta) = \prod_d p(C_d, E_d, \tilde{z}_d, z_d, y_d, \theta_d|\Theta) \tag{20}$$

References

[1] H Attias. Inferring parameters and structure of latent variable models by variational Bayes. In *Proceedings of the 15th Conference on Uncertainty in Artifical Intelligence*, pages 21–30. Morgan-Kaufmann, San Francisco, CA, 1999.

[2] H Attias. A variational bayesian framework for graphical models. *Advances in Neural Information Processing Systems 12*, pages 209–215, 2000.

[3] M J Beal and Z Ghahramani. The variational bayesian em algorithm for incomplete data: with application to scoring graphical model structures. *Bayesian Statistics*, 7:453–464, 2003.

[4] J Bi, T Zhang, and K Bennett. Column-generation boosting mthods for mixture of kernels. In *Proceedings of SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 521–526, 2004.

[5] David M. Blei and Michael I. Jordan. Modeling annotated data. In *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 127–134, New York, NY, USA, 2003. ACM Press.

[6] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[7] C Blenkiron et al. MicroRNA expression profiling of human breast cancer identifies new markers of tumour subtype. *Genome Biology*, 8:R214.1–R214.16, 2007.

[8] R M Boggs, Z M Wright, M J Stickney, W W Porter, and K E Murphy. MicroRNA expression in canine mammary cancer. *Mamm. Genome*, 19:561–9, 2008.

[9] L Carrivick. Probabilistic models in the biomedical sciences. PhD Thesis, Department of Engineering Mathematics, University of Bristol, UK, 2006.

DOI: 10.2202/1544-6115.1441

[10] L. Carrivick, S. Rogers, J. Clark, C. Campbell, M. Girolami, and C. Cooper. Identification of prognostic signatures in breast cancer microarray data using bayesian techniques. *Journal of the Royal Society: Interface*, 3:367–381, 2006.

[11] J Carroll et al. Chromosome-wide mapping of estrogen receptor binding reveals long-range regulation requiring the forkhead protein FOXA1. *Cell*, 122:33–43, 2005.

[12] A Dobra, B Jones, C Hans, J Nevins, and M West. Sparse graphical models for exploring gene expression data. *Journal of Multivariate Analysis*, 90:196–212, 2004.

[13] A Dobra and M West. Graphical model-based gene clustering and metagene expression analysis. Technical report, Duke University Technical Report, Statistics, 2004.

[14] P Farmer et al. Identification of molecular apocrine breast tumours by microarray analysis. *Oncogene*, 24:4660–4671, 2005.

[15] M Girolami and M Zhong. Data integration for classification problems employing gaussian process priors. In *Twentieth Annual Conference on Neural Information Processing Systems*, pages 465–472, 2006.

[16] Mark Girolami and Simon Rogers. Hierarchic bayesian models for kernel learning. In *ICML: 22nd International Conference on Machine Learning*, Bonn, Germany., August 2005.

[17] Y Hua et al. MicroRNA expression profiling in human ovarian cancer: miR-214 induces cell survival and cisplatin resistance by targeting PTEN. *Cancer Research*, 68:425–433, 2008.

[18] M Jordan, Z Ghahramani, T Jaakola, and L Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37:183–233, 1999.

[19] M Lacroix and G Leclerq. About GATA3, HNF3A and XBP1, three genes co-expressed with the oestrogen receptor-alpha gene (ESR1) in breast cancer. *Molecular and Cellular Endocrinology*, 219:1–7, 2004.

[20] J Laganiere et al. Location analysis of estrogen receptor α target promoters reveals that FOXA1 defines a domain of the estrogen response. *Proceedings National Academy Sciences*, 102:11651–11656, 2005.

[21] G R G Lanckriet, T De Bie, N Cristianini, M I Jordan, and W S Noble. A statistical framework for genomic data fusion. *Bioinformatics*, 20:2626–2635, 2004.

[22] Y S Lee and A Dutta. The tumor suppressor microRNA let-7 represses the HMGA2 oncogene. *Genes and Development*, 21:1025–1030, 2007.

[23] U Lehmann, B Hasemeier, D Romermann, M Muller, F Langer, and H Kreipe. Epigenetic inactivation of microRNA genes in mammary carcinoma. *Verh. Dtsch. Ges. Pathol.*, 91:214–20, 2007.

[24] M Middendorf, A Kundaje, C Wiggins, Y Freund, and C Leslie. Predicting genetic regulatory response using classification. In *Proceedings of the Twelfth International Conference on Intelligent Systems in Molecular Biology (ISMB 2004)*, page in press, 2004.

[25] Y Pawitan et al. Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts. *Breast Cancer Research*, 7:R953–R964, 2005.

[26] S Rogers, M Girolami, C Campbell, and R Breitling. The latent process decomposition of cDNA microarray datasets. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2:143–156, 2005.

[27] L F Sempere, M Christensen, A Silahtaroglu, M Bak, C V Heath, G Schwartz, W Wells, S Kauppinen, and C N Cole. Altered microRNA expression confined to specific epithelial cell subpopulations in breast cancer. *Cancer Research*, 67:11612–20, 2007.

[28] M-L Si et al. miR-21-mediated tumor growth. *Oncogene*, 26:2799–2803, 2007.

[29] S Sonnenburg et al. Large scale multiple kernel learning. *Journal of Machine Learning Research*, 7:1531–1565, 2006.

[30] T Sorlie et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings National Academy Sciences*, 98:10869–10874, 2001.

[31] S F Tavazoie, C Alarcon, T Oskarsson, D Padua, Q Wang, P D Bos, W L Gerald, and J Massague. Endogenous human microRNAs that suppress breast cancer metastasis. *Nature*, 451:147–52, 2008.

DOI: 10.2202/1544-6115.1441

[32] A Teschendorff et al. A variational bayesian mixture modelling framework for cluster analysis of gene-expression data. *Bioinformatics*, 21:3025–3033, 2005.

[33] L van 't Veer et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415:530–535, 2002.

[34] Y Wang et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *The Lancet*, 365:671–679, 2005.

[35] M West et al. Predicting the clinical status of human breast cancer using gene expression profiles. *Proceedings of the National Academy of Sciences*, 98:11462–11467, 2001.

[36] F Yang et al. Laser microdissection and microarray analysis of breast tumors reveal ER-α related genes and pathways. *Oncogene*, 25:1413–1419, 2006.

[37] J Zhang, Y Y Du, Y F Lin, Y T Chen, L Yang, H J Wang, and D Ma. The cell growth suppressor, mir-126, targets IRS-1. *Biochem Biophys Res Commun.*, 377:136–40, 2008.