A Marginalized Variational Bayesian Approach to the Analysis of Array Data

Yiming Ying*, Peng Li and Colin Campbell

Address: Department of Engineering Mathematics, University of Bristol, Bristol BS8 1TR, United Kingdom

Email: Yiming Ying*- enxyy@bris.ac.uk; Peng Li - enxpl@bris.ac.uk; Colin Campbell - C.Campbell@bris.ac.uk;

*Corresponding author

Abstract

Background: Bayesian unsupervised learning methods have many applications in the analysis of biological data. For example, for the cancer expression array datasets presented in this study, they can be used to resolve possible disease subtypes and to indicate statistically significant dysregulated genes within these subtypes.

Results: In this paper we outline a marginalized variational Bayesian inference method for unsupervised clustering. In this approach latent process variables and model parameters are allowed to be dependent. This is achieved by marginalizing the mixing Dirichlet variables and then performing inference in the reduced variable space. An iterative update procedure is proposed.

Conclusion: Theoretically and experimentally we show that the proposed algorithm gives a much better free-energy lower bound than a standard variational Bayesian approach. The algorithm is computationally efficient and its performance is demonstrated on two expression array data sets.

Background

Unsupervised clustering methods from machine learning are very appropriate in extracting structure from biological data sets. There has been extensive work in this direction using hierarchical clustering analysis [5], K-Means clustering [11] and self-organizing maps [9]. Bayesian methods are an effective alternative since they provide a mechanism for inferring the number of clusters. They can easily incorporate priors which penalise over-complexed models which would fit to noise and they allow probabilistic confidence measures for cluster membership. In this paper, we focus on Bayesian models which use Dirichlet priors. Examples of these models include Latent Dirichlet Allocation [3] (LDA) for use in text modeling and Latent Process Decomposition (LPD) [8] for analysis of microarray gene expression datasets. One appealing feature of the latter models is that each data point can be stochastically associated with multiple clusters. One approach to model inference is to use methods such as Markov Chain Monte Carlo and Gibbs sampling. However, for the large datasets which occur in many biomedical applications these methods can be too slow for certain tasks such as model selection. This motivates our interest in computationally efficient variational inference

methods [3, 4, 8].

Typically, these inference methods posit that all the latent variables and model parameters are *independent* of each other (i.e. a fully factorized family) which is a strong assumption. In this paper we propose and study an alternative inference method for LPD, which we call marginalized variational Bayesian (MVB). In this approach the latent process (cluster) variables and model parameters are allowed to be *dependent* on each other. As we will show in the next section, this assumption is made feasible by marginalizing the mixing Dirichlet variables, and then performing inference in the reduced variable space. This new approach to constructing an LPD model theoretically and experimentally provides much better free-energy lower bounds than standard a variational Bayes (VB) approach [2,4]. Moreover, the algorithm is computationally efficient and converges faster, as we demonstrate with experiments using expression array datasets.

Methods

The LPD probabilistic model

We start by recalling LPD [8]. Let d index samples, g the genes (attributes) and k the soft clusters (samples are represented as combinatorial mixtures over clusters). The numbers of clusters, genes and samples are denoted $\mathcal{K}$, $\mathcal{G}$, and $\mathcal{D}$ respectively. For each data E_d, we have a multiple process (cluster) latent variable $Z_d = \{Z_{dg} : g = 1, \ldots, \mathcal{G}\}$ where each Z_{dg} is a $\mathcal{K}$-dimensional unit-basis vector, i.e., choosing cluster k is represented by $Z_{dg,k} = 1$ and $Z_{dg,j} = 0$ for $j \neq k$, otherwise. Given the mixing coefficient θ_d, the conditional distribution of Z_d is given by $p(Z_d|\theta_d) = \prod_{g,k} \theta_{dk}^{Z_{dg,k}}$. The conditional distributions, given the latent variables, is given by $p(E_d|Z_d, \mu, \beta) = \prod_{g,k} \left[\mathcal{N}(E_{dg}|\mu_{gk}, \beta_{gk})\right]^{Z_{dg,k}}$, where $\mathcal{N}$ is the Gaussian distribution with mean μ and precision β.

Now we introduce conjugate priors over parameters θ, μ, β. Specifically, we choose $p(\theta_d) = \mathrm{Dir}(\theta_d|\alpha)$, and $p(\mu) \sim \prod_{g,k} \mathcal{N}(\mu_{gk}|m_0, v_0)$, and $p(\beta)$ distributed as $\prod_{gk} \Gamma(\beta_{gk}|a_0, b_0)$ where Γ is defined by $\Gamma(x|a_0, b_0) = x^{a_0-1} \exp(-\frac{x}{b})/b_0^{a_0}\Gamma(b_0)$. We assume the data is i.i.d. and let $\Theta = \{\mu, \beta\}$. The joint distribution is given by

$$p(E, \theta, Z|\Theta) = \prod_d p(\theta_d)p(Z_d|\theta_d)p(E_d|\mu, \beta, Z_d). \tag{1}$$

One can easily see that the marginal likelihood $p(E|\Theta)$ is the same as that in [8]. It is important to note that, in standard Gaussian mixture models [1], each data point is only related with a $\mathcal{K}$-dimensional latent variable which restricts the data to being in one cluster. Instead, in LPD each data point E_d is associated with multiple latent variables $Z_d = \{Z_{dg} : g = 1, \ldots, \mathcal{G}\}$, and thus E_d is stochastically associated with multiple clusters.

Marginalized variational Bayes

In this section we describe a marginalized variational Bayesian approach for LPD. The target of model inference is to compute the posterior distribution $p(\theta, Z, \Theta|E) = p(E, \theta, Z|\Theta)p(\Theta)/p(E)$. Unfortunately, this involves computationally intensive estimation of the integral in the evidence $p(E)$. Hence, we approximate the posterior distribution in a *hypothesis family* whose element are denoted by $q(\theta, Z, \Theta)$.

The standard variational bayesian method [2, 7] uses the equality:

$$\begin{aligned} \log p(E) &= \log \int \sum_Z p(E, \theta, Z, \Theta) d\theta d\Theta \\ &= \mathbb{E}_q\big[\log \tfrac{p(E,\theta,Z|\Theta)p(\Theta)}{q(\theta,Z,\Theta)}\big] + \mathrm{KL}(q(\theta, Z, \Theta)\|p(\theta, Z, \Theta))). \end{aligned} \tag{2}$$

Our optimization target is to maximize the free-energy: $\mathbb{E}_q\big[\log \frac{p(E,\theta,Z|\Theta)p(\Theta)}{q(\theta,Z,\Theta)}\big]$ which, equivalently, minimizes the KL-divergence. One standard way is to choose the hypothesis family in a factorized form $q(\theta, Z, \Theta) = q(\theta)q(Z)q(\Theta)$. In this setting, the free-energy lower bound (2) for the likelihood can be written as:

$$\mathcal{L}(q(\theta), q(Z), q(\Theta)) := \mathbb{E}_q \left[\log \frac{p(E, \theta, Z|\Theta)}{q(\theta)q(Z)} \right] - \mathrm{KL}(q(\Theta \| p(\Theta))). \tag{3}$$

In this paper we study an alternative approach motivated by [12] which only marginalizes the latent variable θ and do variational inference only with respect to the leftover latent variable Z. In essence, we assume that the latent variables θ can be dependent on Z, Θ and the hypothesis family is chosen in the form of $q(\theta, Z, \Theta) = q(\theta|Z, \Theta)q(Z)q(\Theta)$. Since the distribution $q(\theta|Z, \Theta)$ is arbitrary, let it be equal to $p(\theta|E, Z, \Theta) = \frac{p(E,\theta,Z,\Theta)}{p(E,Z,\Theta)}$. Putting this into equation (2) and observing that $\frac{p(E,\theta,Z|\Theta)}{p(\theta|E,Z,\Theta)} = p(E, Z|\Theta)$ gives

$$\begin{aligned} \log p(E) &= \mathbb{E}_q\big[\log \frac{p(E, Z|\Theta)}{q(Z)}\big] - \mathrm{KL}\big(q(\Theta \| p(\Theta))\big) + \mathrm{KL}\big(p(\theta|Z, \Theta)q(\Theta)q(Z) \| p(\theta, Z, \Theta)\big) && (4) \\ &= \mathbb{E}_q\big[\log \frac{p(E, Z|\Theta)}{q(Z)}\big] - \mathrm{KL}\big(q(\Theta \| p(\Theta))\big) + \mathrm{KL}\big(q(Z)q(\Theta) \| p(Z, \Theta)\big). && (5) \end{aligned}$$

Therefore, it is sufficient to maximize the lower bound

$$\mathcal{L}(q(Z), q(\Theta)) := \mathbb{E}_{q(\Theta)q(Z)}\Big[\log \frac{p(E, Z|\Theta)}{q(Z)} \Big] - \mathrm{KL}\big(q(\Theta \| p(\Theta)\big). \tag{6}$$

Observe that $\log \frac{p(E,Z|\Theta)}{q(Z)} \geq \int q(\theta) \log \frac{p(E,\theta,Z|\Theta)}{q(Z)q(\theta)} d\theta$. Consequently, we see that

$$\mathcal{L}(q(\theta), q(Z), q(\Theta)) \leq \mathcal{L}(q(Z), q(\Theta)). \tag{7}$$

As mentioned above, since θ can be dependent on Z, Θ, marginalized VB (MVB) yields a tighter lower bound for the likelihood than the standard VB approach in [4], thus potentially yielding better clustering results.

Model inference and learning

We now turn our attention to the derivation of updates for marginalized VB following the inference methodology [2,7]. For simplicity, let the posterior distribution $q(Z),\ q(\mu),\ q(\beta)$ be indexed by parameters. Specifically, we assume that $q(Z) = \prod_{d,g,k} r_{dg,k}^{Z_{dg,k}},\ \ q(\mu) = \prod_{g,k} \mathcal{N}(\mu_{gk}|m_{gk}, v_{gk})$, and $q(\beta) = \prod_{g,k} \Gamma(\beta_{gk}|a_{gk}, b_{gk})$. Correspondingly, the free-energy lower bound $\mathcal{L}(q(Z), q(\Theta))$ in equation (6) becomes a variational functional over these parameters, and hence we use $\mathcal{L}(R, \mu, \beta)$ later on. The model inference can be summarized by the following coordinate ascent updates.

Let $Z^{\setminus dg}$ denote the random variables excluding Z_{dg}. For any d, g let Θ and $Z^{\setminus dg}$ be fixed, then we take the functional derivative of the free-energy $\mathcal{L}(q(Z), q(\Theta))$ w.r.t. $q(Z_{dg})$ and obtain the update:

$$q(Z_{dg}) \propto \exp\big(\mathbb{E}_{q^{\setminus dg}}\big[\log p(E, Z|\Theta)\big]\big). \tag{8}$$

For the updates for $q(\Theta)$, we obtain

$$q(\mu) \propto p(\mu) \exp\big(\mathbb{E}_{q^{\setminus \mu}}\big[\log p(E, Z|\Theta)\big]\big), q(\beta) \propto p(\beta) \exp\big(\mathbb{E}_{q^{\setminus \beta}}\big[\log p(E, Z|\Theta)\big]\big). \tag{9}$$

Marginalizing out θ in (1) yields

$$\begin{aligned} &p(E, Z|\Theta) := \textstyle\prod_d \big[p(Z_d|\alpha)\big] p(E_d|\mu_d, \beta_d, Z_d) \\ &= \textstyle\prod_d \big[\frac{\Gamma(\sum_k \alpha_k)}{\Gamma\left(\sum_k \alpha_k + \sum_{g,k} Z_{dg,k}\right)} \prod_k \frac{\Gamma(\alpha_k + \sum_g Z_{dg,k})}{\Gamma(\alpha_k)}\big] \prod_{g,k} \big[\mathcal{N}(E_{dg}|\mu_{gk}, \beta_{gk})\big]^{Z_{dg,k}}. \end{aligned} \tag{10}$$

Estimating the expectations of the log likelihoods in equations (8) and (9), we derive the variational EM-updates as follows. Details are postponed to the Appendix.

E-step: using equation (8) and denoting the digamma function by ψ, we have

$$r_{dg,k} \propto \frac{\left(\alpha_k + \sum_{g' \neq g} r_{dg',k}\right) \exp(N_{dg,k})}{\exp\left(\frac{\sum_{g' \neq g} r_{dg',k}(1 - r_{dg',k})}{2(\alpha_k + \sum_{g' \neq g} r_{dg',k})^2}\right)} \tag{11}$$

where $N_{dg,k}$ is given by $0.5(\psi(a_{gk}) + \log b_{gk}) - 0.5 a_{gk} b_{gk}((E_{dg} - m_{gk})^2 + v_{gk}^{-1})$ and $r_{dg,k}$ should be normalized to one over k.

M-step: using equation (9):

$$v_{gk} = v_0 + a_{gk} b_{gk} \sum_d r_{dg,k}, \tag{12}$$

$$m_{gk} = \frac{1}{v_{gk}} \left[v_0 m_0 + a_{gk} b_{gk} \sum_d r_{dg,k} E_{dg} \right], \tag{13}$$

$$a_{gk} = a_0 + 0.5 \sum_d r_{dg,k}, \tag{14}$$

$$\frac{1}{b_{gk}} = \frac{1}{b_0} + 0.5 \sum_d r_{dg,k} \left[(E_{dg} - m_{gk})^2 + \frac{1}{v_{gk}} \right]. \tag{15}$$

We pursue the above iterative procedure until convergence of the lower bound $\mathcal{L}(R; \Theta)$ whose evaluation is given in the Appendix. Since $Z_{dg,k}$ determines the cluster for the observed data point E_d at attribute g and $r_{dg,k}$ is its expectation, we intuitively assign data E_d to cluster $\arg\max\{\sum_g r_{dg,k} : k = 1, \ldots, \mathcal{K}\}$. We can also do model selection over the number of clusters based on a free energy lower bound of the marginalized VB. Experiments in the next section show that this approach is reasonable.

Results

We ran marginalized VB on three data sets. The first was the wine data set from the UCI Repository [10]: this has 178 samples and each sample has 13 features. This data set was chosen for the purpose of validating the proposed method since there are 3 distinct clusters present (derived from 3 cultivars). As more biologically relevant examples we then selected two cancer expression array datasets. The first of these was a lung cancer data set [6] consisting of 73 samples and 918 features. The second was a leukemia data set [13] with 90 samples and 500 features. All the data sets were normalized to zero mean and unit variance and the hyper-parameters m_0, v_0, a_0, and b_0 were chosen to have the same values in both standard VB and marginalized VB. Since the datasets are normalized and m_0, v_0 are hyper-parameters of the Gaussian prior distribution over the mean for the data, it is reasonable to choose $m_0 = 0, v_0 = 1$. For similar reasons, given a_0, b_0 are hyper-parameters of the Gamma prior distribution over the precision (inverse variance) of the data and the mean of a Gamma distributed random variable is $a_0 b_0$, we chose $a_0 = 20$ and $b_0 = 0.05$ throughout these experiments.

First we compared the free energy lower bound of marginalized VB and standard VB based on 30 random initialization. In Figure 1 (top row) we observe an improvement in the free energy as a function of iteration step, for marginalized VB over standard VB. In analogy to standard VB, marginalized VB can determine the appropriate number of soft clusters by estimating the free energy bound given by equation (6) in contrast to the hold-out cross-validation procedure for a maximum likelihood approach to LPD [8]. To investigate the effectiveness of this approach to model selection, free energies were averaged over 20 runs from different random initializations. As shown in Figure 1 (middle row), marginalized VB performed well in determining the correct number of clusters (three) in the UCI wine data set. For the cancer array datasets, the peak in

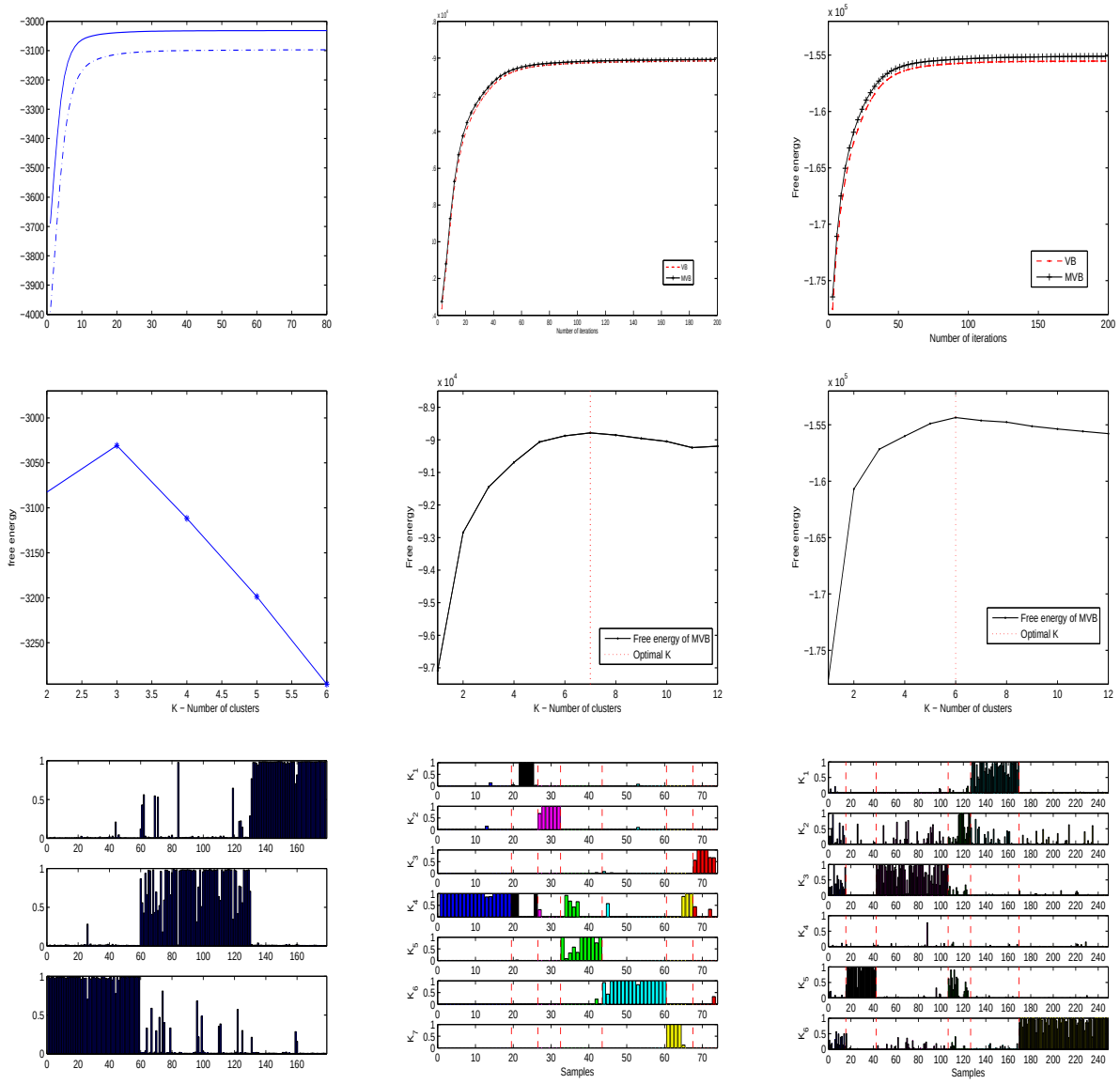

Figure 1: Results for the wine data set (left column), lung cancer data set (middle column) and leukemia data set (right column). Top row: free energy bounds comparison (upper curve:MVB, lower curve:VB). Middle row: free energy (y-axis) versus $\mathcal{K}$, the number of clusters. Bottom row: the normalized $\sum_g r_{dg,k}$ gives a confidence measure that sample d belongs to a cluster k. For the two cancer datasets, samples separated by dashed lines belong to an identified class e.g. adenocarcinoma samples or small cell lung cancer samples (middle column, bottom row, see text).

the averaged free energy is less marked with an indication of six soft clusters for the leukemia data set and seven clusters for the lung cancer data set.

In the bottom row of Figure 1, we see that marginalized VB shows quite promising clustering results using the normalized $\sum_g r_{dg,k}$: these peaks indicate the confidence in the allocation of the dth sample to the kth cluster and accord well with known classifications. The lung cancer dataset of Garber *et al* [6] (middle column, Figure 1) consisted of 73 gene expression profiles from normal and tumour samples with the tumours labelled as squamous, large cell, small cell and adenocarcinoma. The samples are in the order in which they are presented in the original paper [6] with the dashed lines showing their original principal sample groupings. As with Garber *et al* [6] we identified seven clusters in the data with the adenocarcinoma samples falling into three separate clusters with strong correlation with clinical outcomes. For their ordering (which we follow) samples 1-19 belong to adenocarcinoma cluster 1, samples 20-26 belong to adenocarcinoma cluster 2, samples 27-32 are normal tissue samples, samples 33-43 are adenocarcinoma cluster 3, samples 44-60 are squamous cell carcinomas, samples 61-67 are small cell carcinomas and samples 68-73 are from large cell tumours.

As our last example, we applied the proposed MVB method to an oligonucleotide microarray dataset from 360 patients with acute lymphoblastic leukemia (ALL) from Yeoh *et al* [13]. ALL is known to have a number of subtypes with variable responses to chemotherapy. In many cases fusion genes are implicated in the genesis of the disease. For the Yeoh *et al* [13] dataset, samples were drawn from leukemias with rearrangements involving *BCR-ABL*, *E2A-PBX1*, *TEL-AML1*, rearrangements of MLL gene, hyperdiploid karyotope (more than 50 chromosomes) and T lineage leukemias (*T-ALL*). The free energy is plotted in Figure 1 (right column, middle row) with a peak suggesting 6 subtypes. The dashed lines represent the original demarcations of groups according to known genetic rearrangement. Samples 1-15 are *BCR-ABL*, 16-42 are *E2A-PBX1*, 43-106 *Hyperdiploid*> 50, 107-126 *MLL*, 206-248 *T-ALL*, 249-327 *TEL-AML1*, 328-335 *Group23* and 127-205 are labelled as *Others*. Some groupings, such as *E2A-PBX1*, are very distinct. However, the overall groupings are not as well defined as with lung cancer.

Conclusion

We have proposed an efficient variational Bayesian inference method for LPD probabilistic models. By allowing the variables to be dependent on each other, the method can provide more accurate approximation than standard VB. Also, the method provides a principled approach to model selection via the free energy bound. Promising clustering results were also reported on lung cancer and leukemia data sets. Extensions of this method to semi-supervised clustering will be reported elsewhere.

References

1. Bishop CM: *Pattern recognition and machine learning.* (Series: Information Science & Statistics), Springer, 2006.

2. Attias H: **A variational Bayesian framework for graphical models**. *Advances in Neural Information Processing Systems*, 2002, **12:209-215**.

3. Blei DM, Ng AY and Jordan MI: **Latent Dirichlet Allocation**. *Journal of Machine Learning Research* 2003, **3:993-1022**.

4. Carrivick L and Campbell C: **A Bayesian approach to the analysis of micoarray datasets using variational inference**. *Journal submission* (2007).

5. Eisen MB *et al*: **Cluster analysis and display of genome-wide expression patterns**. *Proc. Natl Acad. Sci. USA* 1998, **95:14863-14868**.

6. Garber E *et al*: **Diversity of gene expression in adenocarcinoma of the lung**. *Proceedings National Academy Sciences* 2001, **98:12784-12789**.

7. Jordan MI, Ghahramani Z, Jaakkola T and Saul LK: **An introduction to variational methods for graphical models**. *Machine Learning* 1999, **37:183-233**.

8. Rogers S, Girolami M, Campbell C and Breitling R: **The Latent Process Decomposition of cDNA Microarray Datasets**. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 2005, **2:143-156**.

9. Tamayo P *et al*: **Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation**. *Proc. Natl Acad. Sci. USA* 1999, **96:2907-2912**.

10. Blake CL, Newman DJ, Hettich S and Merz, CJ: *UCI repository of machine learning databases*, http://www.ics.uci.edu/∼mlearn/mlrepository.html, 1998.

11. Tavazoie S *et al*: **Systematic determination of genetic network architecture**. *Nature Genetics* 1999, **22:281-285**.

12. Teh YW, Newman D and Welling M: **A collapsed variational bayesian inference algorithm for latent dirichlet allocation**. *Advances in Neural Information Processing Systems* 2006, **19:1353-1360**.

13. Yeoh E-J *etal*: **Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling**. *Cancer Cell* 2002, **1:133-143**.

Appendix

In this appendix we derive the EM-updates and free energy bound for MVB.

Derivation of updates

Noting that, for any d, g, $\sum_k Z_{dg,k} = 1$ and denoting the number of features by $\mathcal{G}$ we obtain from equation (10):

$$\begin{array}{l} \log p(E, Z|\Theta) = \mathcal{D} \log \Gamma(\sum_k \alpha_k) - \mathcal{D} \sum_k \log \Gamma(\alpha_k) - \mathcal{D} \log \Gamma(\sum_k \alpha_k + \mathcal{G}) \\ + \sum_{d',k} \log \Gamma\big(\alpha_k + \sum_{g'} Z_{d'g',k}\big) + \sum_{d',g',k} Z_{d'g',k} \log \mathcal{N}(E_{d'g'}|\mu_{g'k}, \beta_{g'k}). \end{array} \tag{16}$$

Since $\Gamma\big(\alpha_k + \sum_{g' \neq g} Z_{dg',k} + Z_{dg,k}\big) = \big(\alpha_k + \sum_{g' \neq g} Z_{dg',k}\big)^{Z_{dg,k}} \Gamma\big(\alpha_k + \sum_{g' \neq g} Z_{dg',k}\big)$, putting this observation into the $\log p(E, Z|\Theta)$ yields:

$$\begin{array}{ll} \mathbb{E}_{q \setminus dg}\big(\log p(E, Z|\Theta)\big) & = \sum_k Z_{dg,k} \big(\mathbb{E}_{q \setminus dg}\big[\log(\alpha_k + \sum_{g' \neq g} Z_{dg',k})\big] \\ & + \mathbb{E}_{q(\Theta)}\big[\log \mathcal{N}(E_{dg}|\mu_{gk}, \sigma_{gk})\big]\big) + \text{constant terms} , \end{array}$$

where *constant terms* are independent of $Z_{dg,k}$. Hence, substituting this into equation (8) we conclude that

$$r_{dg,k} \propto \exp\big(\mathbb{E}_{q(\Theta)}\big[\log \mathcal{N}(E_{dg}|\mu_{gk}, \beta_{gk})\big] + \log \mathbb{E}_{q \setminus dg}\big[\log(\alpha_k + \sum_{g' \neq g} Z_{dg',k})\big]\big). \tag{17}$$

To estimate the expectation of the Normal distribution, we use the fllowing observations (e.g. [2]) for the Gamma and Normal distributions:

$$\mathbb{E}_{q(\beta)}[\beta_{gk}] = a_{gk} b_{gk}, \quad \mathbb{E}_{q(\beta)}[\log \beta_{gk}] = \psi(a_{gk}) + \log b_{gk},$$

and

$$\mathbb{E}_{q(\mu)}[\mu_{gk}^2] = m_{gk}^2 + v_{gk}^{-1}, \ \mathbb{E}_{q(\mu)}[\mu_{gk}] = m_{gk}.$$

Consequently, simple manipulation yields:

$$\mathbb{E}_{q(\Theta)}\big[\log \mathcal{N}(E_{dg}|\mu_{gk}, \beta_{gk})\big]$$

equals, up to a constant term:

$$0.5(\psi(a_{gk}) + \log b_{gk}) - 0.5 a_{gk} b_{gk}((E_{dg} - m_{gk})^2 + v_{gk}^{-1}).$$

We also use approximating methods [12] to estimate $\log \mathbb{E}_{q \setminus dg}\big[\log(\alpha_k + \sum_{g' \neq g} Z_{dg',k})\big]$. For this purpose, we observe, for any positive random variable x, that

$$\mathbb{E}\big(\log(\alpha_k + x)\big) \approx \log(\alpha_k + \mathbb{E}x) - \frac{\mathrm{Var}(x)}{2(\alpha_k + \mathbb{E}x)^2}, \tag{18}$$

and $\mathbb{E}_{q \setminus dg}\big[\sum_{g' \neq g} Z_{dg',k}\big] = \sum_{g' \neq g} r_{dg',k}$, $\quad \mathrm{Var}(\sum_{g' \neq g} Z_{dg',k}) = \sum_{g' \neq g} r_{dg',k}(1 - r_{dg',k})$. Plugging the above observations into equation (17) yields the desired E-step updates.

For the updates for $q(\Theta)$, the updates are essentially the same as those in [2,4] since the associated terms with variables with Θ in $\mathbb{E}_{q \setminus \mu}[\log p(E, Z|\Theta)]$ are exact the same, that is, Θ only appears in the Normal distribution. Hence, noting that the product of two Gamma (Normal) distributions is a Gamma (Normal) distribution, we can obtain, from equations (16) and (9), the M-step updates.

Free energy bound

The free-energy lower bound of marginalized VB is defined by equation (6):

$$\mathcal{L}(R;\Theta) = \mathbb{E}_q\big[\log p(E,Z|\Theta)\big] - \mathbb{E}_{q(Z)}\big[q(Z)\big] - \mathrm{KL}\big(q(\mu)\|p(\mu)\big) - \mathrm{KL}\big(q(\beta)\|p(\beta)\big).$$

From the fact that $\Gamma(x+1) = x\Gamma(x)$ for any $x > 0$, we know that $\Gamma\big(\alpha_k + \sum_g Z_{dg,k}\big) = \Gamma(\alpha_k)\prod_{g=1}^{\mathcal{G}}\big(\alpha_k + \sum_{j=g+1}^{\mathcal{G}} Z_{dj,k}\big)^{Z_{dg,k}}$, where we use the convention $\sum_{j=\mathcal{G}+1}^{\mathcal{G}} = 0$. Putting this equation into the expression (10) of log likelihood, we obtain:

$$\begin{array}{l}\mathbb{E}_q\big[\log p(E,Z|\Theta)\big] = \mathcal{D}\log\Gamma(\sum_k \alpha_k) - \mathcal{D}\sum_k \log\Gamma(\alpha_k) - \mathcal{D}\log\Gamma(\sum_k \alpha_k + \mathcal{G}) \\ + \sum_{d,k}\mathbb{E}_q\big[\log\Gamma(\alpha_k + \sum_g Z_{dg,k})\big] + \sum_{d,g,k} r_{dg,k}\log\mathcal{N}(E_{dg}|\mu_{gk},\sigma_{gk}) \\ = \mathcal{D}\log\Gamma(\sum_k \alpha_k) - \mathcal{D}\log\Gamma(\sum_k \alpha_k + \mathcal{G}) \\ + \sum_{d,g,k} r_{dg,k}\big(\mathbb{E}_q\big[\log\big(\alpha_k + \sum_{j\geq g+1} Z_{dj,k}\big)\big] + \log\mathcal{N}(E_{dg}|\mu_{gk},\sigma_{gk})\big).\end{array}$$

Since we used the convention $\sum_{j\geq\mathcal{G}+1} Z_{dj,k} = 0$, $\mathbb{E}_q\big[\log\big(\alpha_k + \sum_{j\geq\mathcal{G}+1} Z_{dj,k}\big)\big] = \log\alpha_k$. It remains to estimate the term $\mathbb{E}_q\big[\log\big(\alpha_k + \sum_{j\geq g+1} Z_{dj,k}\big)\big]$ for $g = 1,\ldots,\mathcal{G}-1$. To this end, we use the approximation (18) again to get:

$$\mathbb{E}_q\big[\log\big(\alpha_k + \sum_{j\geq g+1} Z_{dj,k}\big)\big] \approx \log\big(\alpha_k + \sum_{j\geq g+1} r_{dj,k}\big) - \frac{\sum_{j\geq g+1} r_{dj,k}(1-r_{dj,k})}{2(\alpha_k + \sum_{j\geq g+1} r_{dj,k})^2}.$$

Consequently, we conclude:

$$\begin{array}{l}\mathbb{E}_q\big[\log p(E,Z|\Theta)\big] = \mathcal{D}\log\Gamma(\sum_k \alpha_k) - \mathcal{D}\log\Gamma(\sum_k \alpha_k + \mathcal{G}) \\ + \sum_{d,g,k} r_{dg,k}\big[\log\big(\alpha_k + \sum_{j\geq g+1} r_{dj,k}\big) - \frac{\sum_{j\geq g+1} r_{dj,k}(1-r_{dj,k})}{2(\alpha_k + \sum_{j\geq g+1} r_{dj,k})^2}\big] \\ + \sum_{d,g,k} r_{dg,k}\big[-0.5\log 2\pi + 0.5(\psi(a_{gk}) + \log b_{gk}) \\ -0.5 a_{gk} b_{gk}(\frac{1}{v_{gk}} + (E_{dg} - m_{gk})^2)\big].\end{array} \tag{19}$$

where the convention $\sum_{g\geq\mathcal{G}+1}^{\mathcal{G}} = 0$ is used again.

In addition,

$$\mathbb{E}_{q(Z)}\big[\log q(Z)\big] = \sum_{d,g,k} r_{dg,k}\log r_{dg,k}.$$

For the KL divergences, we have:

$$\mathrm{KL}(q(\mu)\|p(\mu)) = \sum_{g,k} 0.5\log\frac{v_{gk}}{v_0} + 0.5 v_0[m_{gk} - m_0]^2 + 0.5\left(\frac{v_0}{v_{gk}} - 1\right),$$

and

$$\begin{array}{rl}\mathrm{KL}(q(\beta\|p(\beta))) & = \sum_{g,k}(a_{gk} - a_0)\psi(a_{gk}) - \log b_{gk} - a_{gk} - \log\Gamma(a_{gk}) \\ & + \log\Gamma(a_0) + a_0\log b_0 - (a_0 - 1)(\psi(a_{gk}) + \log b_{gk}) + \frac{a_{gk} b_{gk}}{b_0}.\end{array}$$