A Marginalized Variational Bayesian Approach to the Analysis of Array Data

Yiming Ying, Peng Li and Colin Campbell

Department of Engineering Mathematics, University of Bristol,
{enxyy, enxpl, C.Campbell}@bristol.ac.uk

Abstract. Bayesian unsupervised learning methods have many applications in the analysis of biological data. 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. Theoretically and experimentally we show that this gives a much better free-energy lower bound than a standard variational Bayesian approach. The algorithm is computationally efficient as demonstrated on cancer microarray data sets.

1 Introduction

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 [10] 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 of 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, 8, 4].

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 called marginalized variational Bayesian (MVB) inference for LPD. In this approach the latent process (cluster) variables and model parameters are allowed

to be *dependent* on each other. As we will show in Section 3, 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 VB [2, 4]. Moreover, the algorithm is computationally efficient and converges faster, as we demonstrate with experiments using expression array datasets.

2 LPD probabilistic model

We start by recalling LPD [8]. Let d index samples, g the genes (attributes) and k the processes (cluster). The numbers of processes, genes and samples are denoted $\mathcal{K}$, $\mathcal{G}$, and $\mathcal{D}$ respectively. For each data E_d, we have a multiple process 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 process 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.

3 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 $p(\theta, Z, \Theta|E) = \frac{p(E,\theta,Z|\Theta)p(\Theta)}{p(E)}$. Unfortunately, it involves computationally intensive estimation of the integral in the evidence $p(E)$. Hence, we approximate the posterior distribution in a *hypothesis family* whose element is 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 by

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

In this paper we study an alternative approach motivated by [11] 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 $\log p(E) = \mathbb{E}_q\big[\log \frac{p(E,Z|\Theta)}{q(Z)}\big] - \mathrm{KL}\big(q(\Theta\|p(\Theta))\big) + \mathrm{KL}\big(p(\theta|Z,\Theta)q(\Theta)q(Z)\|p(\theta,Z,\Theta)\big) = \mathbb{E}_q\big[\log \frac{p(E,Z|\Theta)}{q(Z)}\big] - \mathrm{KL}\big(q(\Theta\|p(\Theta))\big) + \mathrm{KL}\big(q(Z)q(\Theta)\|p(Z,\Theta)\big)$. Therefore, it is sufficient to maximize the lower bound

$$\mathcal{L}(q(Z),q(\Theta)) := \mathbb{E}_{q(\Theta)q(Z)}\Big[\log \frac{p(E,Z|\Theta)}{q(Z)}\Big] - \mathrm{KL}\big(q(\Theta\|p(\Theta)\big). \tag{4}$$

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 $\mathcal{L}(q(\theta),q(Z),q(\Theta)) \leq \mathcal{L}(q(Z),q(\Theta))$. As mentioned above, since θ can be dependent on Z,Θ and marginalized VB (MVB) yields a tighter lower bound for the likelihood than the standard VB approach in [4], thus potentially yielding better clustering results.

4 Model inference and learning

We 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 (4) 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 $Z^{\backslash dg}$ denote the random variables excluding Z_{dg}. For any d, g let Θ and $Z^{\backslash dg}$ 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:

$$q(Z_{dg}) \propto \exp\big(\mathbb{E}_{q^{\backslash dg}}\big[\log p(E, Z|\Theta)\big]\big). \tag{5}$$

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

$$q(\mu) \propto p(\mu)\exp\big(\mathbb{E}_{q^{\backslash \mu}}\big[\log p(E, Z|\Theta)\big]\big), q(\beta) \propto p(\beta)\exp\big(\mathbb{E}_{q^{\backslash \beta}}\big[\log p(E, Z|\Theta)\big]\big). \tag{6}$$

Marginalizing out θ in (1) yields

$$\begin{aligned} &p(E, Z|\Theta) := \textstyle\prod_d \big[p(Z_d|\alpha)\big] p(E_d|\mu_d, \beta_d, Z_d) \\ &= \textstyle\prod_d \big[\frac{\Gamma(\sum_k \alpha_k)}{\Gamma(\sum_k \alpha_k + \sum_{g,k} Z_{dg,k})} \prod_k \frac{\Gamma(\alpha_k + \sum_g Z_{dg,k})}{\Gamma(\alpha_k)}\big] \prod_{g,k} \big[\mathcal{N}(E_{dg}|\mu_{gk}, \beta_{gk})\big]^{Z_{dg,k}}. \end{aligned} \tag{7}$$

Estimating the expectations of the log likelihoods in equations (5) and (6), we derive the variational EM-updates (details are omitted due to limited space):
E-step: using equation (5) and denoting the digamma function by ψ, we have $r_{dg,k} \propto \big(\alpha_k + \sum_{g' \neq g} r_{dg',k}\big) \exp(N_{dg,k}) / \exp\big(\frac{\sum_{g' \neq g} r_{dg',k}(1 - r_{dg',k})}{2(\alpha_k + \sum_{g' \neq g} r_{dg',k})^2}\big)$ 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 equations (6): $v_{gk} = v_0 + a_{gk} b_{gk} \sum_d r_{dg,k}$, $\quad m_{gk} = \frac{1}{v_{gk}}\big[v_0 m_0 + a_{gk} b_{gk} \sum_d r_{dg,k} E_{dg}\big]$, and $a_{gk} = a_0 + 0.5 \sum_d r_{dg,k}$, $\quad \frac{1}{b_{gk}} = \frac{1}{b_0} + 0.5 \sum_d r_{dg,k}\big[(E_{dg} - m_{gk})^2 + \frac{1}{v_{gk}}\big]$.

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 process for the observed data point E_d at gene 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 processes based on a free energy lower bound of the marginalized VB. Experiments in the next section show that this approach is reasonable.

5 Experiments

The MVB was evaluated on a *lung-cancer* [6] and a *leukemia* data set [12]. The lung cancer microarray data [6] has 73 tissue samples and each sample has 918 features. Seven clusters has specified in [6] which are known classes, such as small cell lung cancer. For the leukemia microarray data [12], there are 248 samples and the six labels are known subtypes based on observed translocation or genomic rearrangements. We have only used a subset of the original data with unambiguous sample labels. The dimensionality of the leukemia data set was reduced from 12625 to 500 based on largest variance.

All the data sets are normalized to zero mean and standard deviation one. The same hyper-parameters m_0, v_0, a_0, and b_0 are used for both standard VB and

marginalized VB. m_0, v_0 are hyper-parameters of the Gaussian prior distribution over the mean of the data. It is reasonable to choose $m_0 = 0, v_0 = 1$ since data sets are normalized. 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 $a_0 b_0$. Without loss of generality, $a_0 = 20$ and $b_0 = 0.05$ were used throughout the experiments.

The free energy bounds of MVB and VB were evaluated firstly. MVB bounds at different iterations over 20 runs were illustrated in the first row of Figure 1. It is observed that the lower bound of MVB is higher than that of VB on both data sets. This shows that our proposed MVB provides better approximation to the likelihood than that of VB.

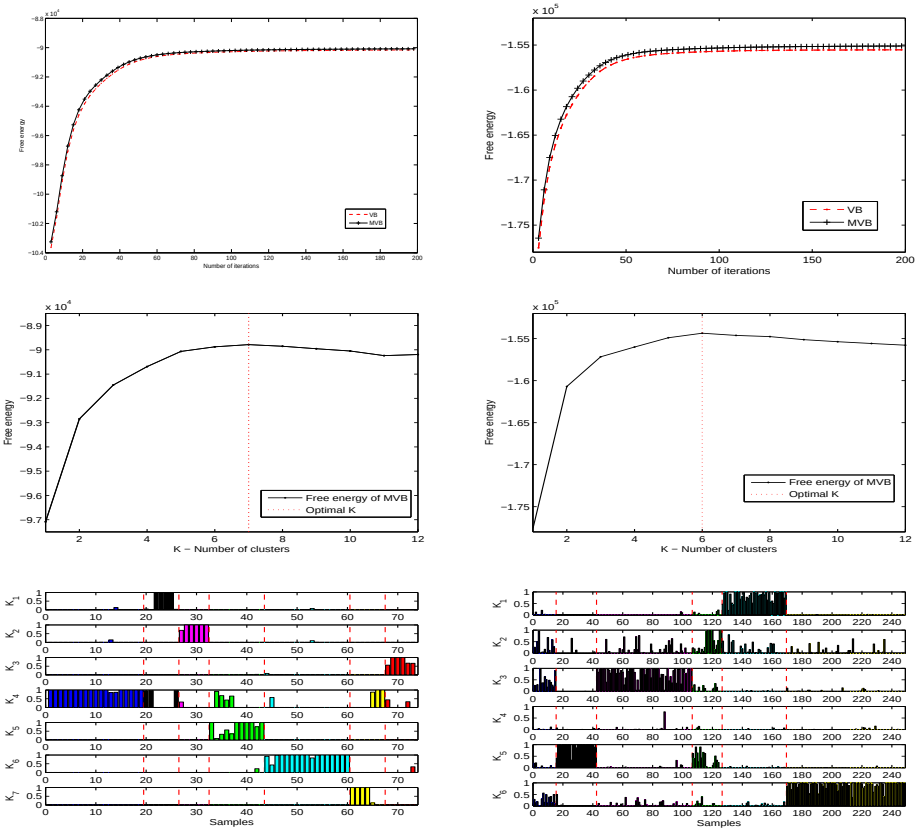

Fig. 1. Experimental results on *lung-cancer* (left column) and *leukemia* (right column) data sets. Top row: free energy bound comparison. Middle row: free energy (y-axis) of MVB versus $\mathcal{K}$. Bottom row: normalized $\sum_g r_{dg,k}$ which gives the confidence that a sample belongs to a given cluster. The samples in same color and separated by red dashed line are from same class.

In analogy to the standard VB, MVB can determine the appropriate number of soft clusters by using the free energy bound given by equation (4) in contrast

to the cross-validation of hold-out maximum likelihood in LPD [8]. To this end, the algorithm were run based on 20 random initializations. As shown in the middle row of Figure 1, MVB determines the correct processes for *lung-cancer* (seven principal subtypes) and *leukemia* (six known subtypes). For the last row of the figure, MVB shows quite promising clustering result using the normalized $\sum_g r_{dg,k}$, expressing the confidence allocation of dth sample to the kth process. As we will discuss elsewhere, these categories are clinically relevant.

6 Discussion

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. Meanwhile, the method provides a principled approach to model selection via the free energy bound. Quite promising clustering results are also reported on lung cancer and leukemia data sets. We are pursuing more experiments on cancer data sets. Extensions of this method to semi-supervised clustering could be a future avenue for further research.

References

1. Bishop, C. M.: Pattern recognition and machine learning. (Series: Information Science & Statistics), Springer, 2006.
2. Attias H.: A variational Bayesian framework for graphical models. NIPS, **12** (2000).
3. Blei, David M., Ng, Andrew Y., and Jordan, Michael I.: Latent Dirichlet Allocation. Journal of Machine Learning Research. **3** (2003) 993-1022.
4. Carrivick, L. and Campbell, C.: A Bayesian approach to the analysis of micoarray datasets using variational inference, Preprint, (2007).
5. Eisen, M. B. et al: Cluster analysis and display of genome-wide expression patterns. Proc. Natl Acad. Sci. USA, **95**(1998) 14863-14868.
6. Garber, E. et al: Diversity of gene expression in adenocarcinoma of the lung. Proceedings National Academy Sciences, **98** (2001) 12784–12789.
7. Jordan, Michael I., Ghahramani, Z., Jaakkola, T., and Saul, L. K.: An introduction to variational methods for graphical models. Machine Learning, **37**(1999) 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, **2** (2005) 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, **96**(1999) 2907-2912.
10. Tavazoie S. et al: Systematic determination of genetic network architecture. Nature Genetics, **22**(1999) 281-285.
11. Teh Y. W., Newman, D., and Welling, M.: A collapsed variational bayesian inference algorithm for latent dirichlet allocation. NIPS, **19** (2006).
12. E-J Yeoh et al.: Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. Cancer Cell, **1** (2002) 133–143.