A Variational Approach to Semi-Supervised Clustering

Peng Li†,* Yiming Ying and Colin Campbell
Department of Engineering Mathematics, University of Bristol,
Bristol BS8 1TR, United Kingdom

Abstract

We present a variational inference scheme for semi-supervised clustering in which data is supplemented with side information in the form of common labels. There is no mutual exclusion of classes assumption and samples are represented as a combinatorial mixture over multiple clusters. The method has other advantages such as the ability to find the most probable number of soft clusters in the data, the ability to avoid overfitting and the ability to find a confidence measure for membership of a soft cluster. We illustrate performance of the method on six data sets and find a positive comparison against constrained K-means clustering.

1 Introduction

With semi-supervised clustering the aim is to find clusters or meaningful partitions of the data, aided by the use of *side information*. This side information can be in the form of *must-links* (two samples must be in the same cluster) and *cannot-links* (two samples must not belong to the same cluster). A number of approaches have been proposed for semi-supervised clustering. Typically these methods have used the constraints from the side information to either alter the objective function or the distance measure used. Many methods have used modifications of pre-existing clustering schemes such as incremental clustering algorithms [13], K-means clustering [2, 7] or Hidden Markov Random Fields [3]. One problem with some of these approaches is that the method can work well if the correct number of clusters is already known: K-means clustering is an example. However, a principled approach to finding K is generally not given. Other potential disadvantages of some clustering approaches is that there is an implicit mutual exclusion of clusters assumption i.e. samples are assumed to uniquely belong to a particular cluster. In many cases this assumption may not be fully valid and it would be more appropriate for a sample to be associated with multiple soft clusters. A further issue with some clustering approaches is that there is no principled mechanism to avoid fitting to noise in the data. With Bayesian methods, however, we can incorporate a Bayes prior which penalizes such overfitting.

Our motivation for considering semi-supervised clustering comes from potential applications in cancer informatics. There have been a number of instances where unsupervised learning methods have been applied to cancer microarray data sets, for example, and clinically distinct subtypes of cancer have been indicated e.g. [1, 9]. However, in some cases a specific causative event is known and thus it is possible to give common labels to some samples. A specific example might be acute lymphoblastic leukemia [15]. This disease is known to have a number of subtypes with variable response to chemotherapy. An originating event for some of these subtypes is believed known, sometimes stemming from the creation of fusion genes through genetic rearrangement involving genes *BCR-ABL*, *E2A-PBX1*, *TEL-AML1* or rearrangements of the *MLL* gene. These rearrangements can be detected via FISH (Fluorescent in situ Hybridization), and thus we can assign common labels to certain samples and use semi-supervised clustering to improve characterization of unlabelled samples. In these cancer applications, the side information is typically in the form of *must-links* which will therefore be the focus of this paper.

We thus propose a probabilistic graphical model approach to semi-supervised clustering with the side information given as the addition of must-links between some samples. The proposed method provides an objective measure of the number of clusters present and can readily handle missing values. Samples are represented as combinatorial mixtures over a set of latent processes or soft clusters, so there is no

*†Current address: Dept. of Computer Science, University College London, Gower Street, London WC1E 6BT, United Kingdom

mutual exclusion of clusters assumption. In addition, by allowing a representation overlapping different clusters we can derive a confidence measure for cluster membership: a confidence measure for subtype membership is important in clinical cancer informatics applications. As a Bayesian methodology we can also modify the method to incorporate a Bayes prior penalizing over-complex models which would fit to noise in the data. In the next two sections we propose our semi-supervised probabilistic graphical model, followed by experiments in section 4.

2 The Methods

Our semi-supervised model utilizes the Latent Process Decomposition (LPD) model developed in [5, 12], and hence we will call this variant semi-supervised LPD or SLPD. For the natural numbers we adopt the notation $\mathbb{N}_m = \{1, \ldots, m\}$ for any $m \in \mathbb{N}$. For the data we use d as the sample index, g as the attribute index, and script letters $\mathcal{D}, \mathcal{G}$ to index the corresponding number of samples and attributes. The number of clusters is $\mathcal{K}$. The complete data set is $E = \{E_{dg} : d \in \mathbb{N}_{\mathcal{D}}, g \in \mathbb{N}_{\mathcal{G}}\}$. This notation stems from our cancer informatics motivation of using the expression value of gene g in sample d.
In our semi-supervised setting, we have additional block information $\mathcal{C}$ where each *block* c denotes a set of data points that is known to belong to a single class. In keeping with the standard Bayesian models, we also assume both blocks and the data points in each block are *i.i.d.* sampled. Specifically, this side information can be represented by a $\mathcal{D} \times C$ matrix δ with its entities δ_{dc} defined as follows

$$\delta_{dc} = \begin{cases} 1 & \text{if data } d \text{ is a member of block } c, \\ 0 & \text{otherwise.} \end{cases}$$

We can now describe our semi-supervised graphical model. In probabilistic terms, the data set E can be partitioned into $\mathcal{K}$-processes (soft clusters) described as follows. For a complete data set, a Dirichlet prior distribution for the distribution of processes is defined by $\mathcal{K}$-dimensional parameter α. For each known block c, a distribution θ_c over the set of mixture components indexed by k is drawn from a single Dirichlet distribution parameterized by α. Then, for all samples d in block c (i.e. $\delta_{dc} = 1$), the latent indicator variable Z_{dg} indicates which process k is chosen, with probability θ_{ck}, from the common block-specific distribution θ_c. The value E_{dg} for attribute g in sample d is then drawn from the kth Gaussian with mean μ_{gk} and deviation σ_{gk}, denoted as $\mathcal{N}(E_{dg}|\mu_{gk}, \sigma_{gk})$. We repeat the above procedure for each block in $\mathcal{C}$. The graphical model is illustrated in Figure 1 which is motivated by Latent Dirichlet Allocation [5].

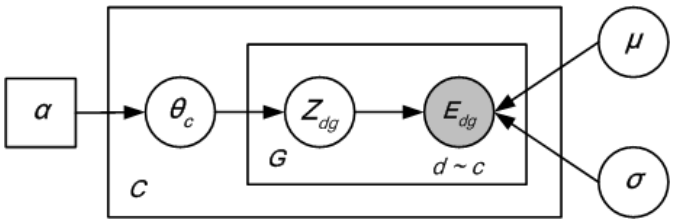

Figure 1: A graphical model representation of the method proposed in this paper. E_{dg} denotes the value of attribute g in sample d. μ and σ are model parameters. Z_{dg} is a hidden variable giving the process index of attribute g in sample d. θ_c gives the mixing over subgroups for sample d in block c denoted by $d \sim c$. The probability of θ_c is given by a Dirichlet distribution with hyper-parameter α.

The *model parameters* are $\Theta = (\mu, \sigma, \alpha)$ and we use the notation $d \sim c$ to denote sample d in block c. From the graphical model in Figure 1, we can formulate the block-specific joint distribution of the observed data E and the latent variables Z by

$$p(E, \theta, Z|\Theta, \mathcal{C}) = \prod_c p(\theta_c|\alpha) \prod_{d \sim c} p(E_d, Z|\theta_c, \Theta), \tag{1}$$

where $p(\theta_c|\alpha)$ is Dirichlet defined by $p(\theta_c|\alpha) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_k \theta_{ck}^{\alpha_k - 1}$. Using the block matrix δ, we further see that

$$\begin{aligned}\prod_{d\sim c} p(E_d, Z|\theta_c, \Theta) &= \prod_d \Big[p(Z_d|\theta_c)p(E_d|Z_d, \mu, \sigma)\Big]^{\delta_{dc}} \\ &= \prod_d \prod_g \Big[p(Z_{dg}|\theta_c)\mathcal{N}(E_{dg}|\mu_g, \sigma_g, Z_{dg})\Big]^{\delta_{dc}}.\end{aligned} \tag{2}$$

For notational simplicity, we regard Z_{dg} as a unit-basis vector $(Z_{dg,1}, \ldots, Z_{dg,\mathcal{K}})$ which transforms the process latent variable $Z_{dg} = k$ to the unique vector Z_{dg} given by $Z_{dg,k} = 1$ and $Z_{dg,j} = 0$ for $j \neq k$. Equivalently, the random variable Z_{dg} is distributed according to a multinomial probability defined by $p(Z_{dg}|\theta_c) = \prod_k \theta_{ck}^{Z_{dg,k}}$. Hence, the above equation can be rewritten as

$$p(E, \theta, Z|\Theta, \mathcal{C}) = \prod_c p(\theta_c|\alpha) \prod_d \prod_g \prod_k \Big[\theta_{ck}\mathcal{N}(E_{dg}|\mu_{gk}, \sigma_{gk})\Big]^{Z_{dg,k}\delta_{dc}}. \tag{3}$$

With these priors, the final data likelihood can be obtained by marginalizing out the latent variables θ and $Z := \{Z_{dg} : d \in \mathbb{N}_\mathcal{D}, g \in \mathbb{N}_\mathcal{G}\}$

$$p(E|\Theta, \mathcal{C}) := \int_\theta \sum_Z p(E, \theta, Z|\Theta, \mathcal{C}) d\theta. \tag{4}$$

In particular, we can see from equations (2) and (3) that

$$\begin{aligned}p(E|\Theta, \mathcal{C}) &= \prod_c \int_{\theta_c} \sum_Z \prod_d \prod_g \Big[p(Z_{dg}|\theta_c)\mathcal{N}(E_{dg}|\mu_g, \sigma_g, Z_{dg})\Big]^{\delta_{dc}} p(\theta_c|\alpha) d\theta_c \\ &= \prod_c \int_{\theta_c} \sum_{Z_{dg}, d\sim c} \prod_{d\sim c, g, k} \Big[\theta_{ck}\mathcal{N}(E_{dg}|\mu_{gk}, \sigma_{gk})\Big]^{Z_{dg,k}} p(\theta_c|\alpha) d\theta_c \\ &= \prod_c \int_{\theta_c} \prod_{d,g} \Big[\sum_k \theta_{ck}\mathcal{N}(E_{dg}|\mu_g, \sigma_g, Z_{dg})\Big]^{\delta_{dc}} p(\theta_c|\alpha) d\theta_c.\end{aligned} \tag{5}$$

We should mention that, without block information (i.e. $\delta = I_{\mathcal{D}\times\mathcal{D}}$), the above equation is the exact likelihood of Latent Process Decomposition given in [12].

3 Model inference and parameter learning

We now consider model inference and parameter estimation under SLPD. The main inferential goal is to compute the posterior distribution of the hidden variables $p(\theta, Z|E, \Theta, \mathcal{C})$. One direct method is to use Bayes rule $p(\theta, Z|E, \Theta, \mathcal{C}) = \frac{p(E,\theta,Z|\Theta,\mathcal{C})}{p(E|\Theta,\mathcal{C})}$. This approach is usually intractable since this involves computationally intensive estimation of multi-integrals in the final likelihood $p(E|\theta, \mathcal{C})$.
In this paper, we rely on *variational inference* methods [10, 11] which maximize a lower bound on the likelihood $p(E|\Theta, \mathcal{C})$ to estimate the model parameters Θ and approximate $p(\theta, Z|E, \Theta, \mathcal{C})$ in a *hypothesis family*. One common hypothesis family is the *factorized family* defined by $q(\theta, Z|\gamma, Q) = q(\theta|\gamma)q(Z|Q)$ with *variational parameters* γ, Q where, in the expression of the distribution q, the dependency on the $E, \Theta, \mathcal{C}$ is omitted. More specifically, in our model we assume that $q(\theta|\gamma) = \prod_c q(\theta_c|\gamma_c) = \prod_c \big(\frac{\Gamma(\sum_k \gamma_{ck})}{\prod_k \Gamma(\gamma_{ck})} \prod_k \theta_{ck}^{\gamma_{ck}-1}\big)$, and $q(Z|Q) = \prod_{d,g} q(Z_{dg}|Q_{dg}) = \prod_{d,g} \big(\prod_k Q_{dg,k}^{Z_{dg,k}}\big)$, among which γ, Q will be set as we describe below.
We can lower bound the log-likelihood by applying Jensen's inequality to equation (4):

$$\begin{aligned}\log p(E|\Theta, \mathcal{C}) &= \log \int_\theta \sum_Z p(E, \theta, Z|\Theta, \mathcal{C}) d\theta \\ &\geq \mathcal{L}(\gamma, Q; \Theta) := \mathbb{E}_q\big[\log p(E, \theta, Z|\Theta, \mathcal{C})\big] - \mathbb{E}_q\big[q(\theta, Z|\gamma, Q)\big].\end{aligned}$$

Consequently we can estimate the variational and model parameters by alternative coordinate ascent methods known as a variational EM algorithm. The details are summarized in the Appendix.

- E-step: maximize $\mathcal{L}$ with respect to the variational parameters γ, Q and get the updates

$$Q_{dg,k} = \frac{\mathcal{N}(E_{dg}|\mu_{gk},\sigma_{gk})\left[\prod_c \exp(\delta_{dc}(\Psi(\gamma_{ck}) - \Psi(\sum_k \gamma_{ck})))\right]}{\sum_k \mathcal{N}(E_{dg}|\mu_{gk},\sigma_{gk})\left[\prod_c \exp(\delta_{dc}\Psi(\gamma_{ck}) - \Psi(\sum_k \gamma_{ck}))\right]}, \tag{6}$$

  and

$$\gamma_{ck} = \alpha_k + \sum_{d,g} \delta_{dc} Q_{dg,k}. \tag{7}$$

- M-step: maximize $\mathcal{L}$ with respect to α, μ and σ and get that

$$\mu_{gk} = \frac{\sum_d Q_{dg,k} E_{dg}}{\sum_d Q_{dg,k}}, \quad \sigma^2_{gk} = \frac{\sum_d Q_{dg,k}(E_{dg} - \mu_{gk})^2}{\sum_d Q_{dg,k}}. \tag{8}$$

  and the parameter α is found using a Newton-Raphson method as given in the Appendix.

The above iterative procedure is run until convergence (plateauing of the lower bound or the estimated likelihood $p(E|\Theta,\mathcal{C})$, see the Appendix for details). Interpretation of the resultant model is very similar to Latent Process Decomposition [12]. When normalized over k, the parameter γ_{ck} gives the confidence that a sample belonging to block c (which share a common label) belongs to soft cluster k. For each soft cluster k the model parameters μ_{gk} and σ_{gk} give a density distribution for the attribute value g over all samples, see [6] for examples of use of these density estimators in application to interpreting breast cancer microarray data. If some values of E_{dg} are missing, we omit corresponding contributions in the M-step updates and the corresponding $Q_{dg,k}$.
The above argument is based on a maximum likelihood approach. We end this section with a comment about the maximum a posterior (MAP) solution. In this case μ_{gk} and σ_{gk} are produced by prior distributions $p(\mu) = \prod_{gk} p(\mu_{gk})$ and $p(\sigma^2) = \prod_{gk} p(\sigma^2_{gk})$ with $p(\mu_{gk}) \sim \mathcal{N}(0,\sigma_\mu), \quad p(\sigma^2_{gk}) \sim \exp(-\frac{s}{\sigma^2_{gk}})$. These priors for penalizing over-complexity are justified in [12]. In this case we only need to change the updates for μ and σ

$$\mu_{gk} = \frac{\sigma^2_\mu \sum_d Q_{dg,k} E_{dg}}{\sigma^2_{gk} + \sigma^2_\mu \sum_d Q_{dg,k})}, \quad \sigma^2_{gk} = \frac{\sum_d Q_{dg,k}(E_{dg} - \mu_{gk})^2 + 2s}{\sum_d Q_{dg,k}}. \tag{9}$$

The argument leading to the above updates is the same as before except we replace the likelihood $p(E|\Theta)$ by $p(E|\Theta)p(\mu)p(\sigma)$, and consequently the lower bound is replaced by $\mathcal{L}(Q;\Theta) + \log p(\mu) + \log p(\sigma)$.

4 Experimental Results

To validate the proposed approach, we investigated the ML solution applied to four datasets from the UCI Repository [8] and two cancer microarray data sets (see Table 1). The four data sets from the UCI Repository have known sample labels and thus we can achieve an objective performance measure. As mentioned in the introduction section, our interest in semi-supervised clustering stems from a potential use in cancer informatics. Thus the next two data sets we consider are for cancer. One is a leukemia microarray data [15]: in this case some labels are known since the causative events are observed genetic translocations or rearrangements. We have only used a subset of the original data with unambiguous sample labels. The second dataset is for lung cancer [4]. Again this consisted of histologically labelled samples derived from squamous cell lung carcinomas (21 samples), adenocarcinomas (139 samples), pulmonary carcinoids (20 samples), small-cell lung carcinomas (6 samples) and normal lung tissue (17 samples). The dimensionality of the both cancer datasets was reduced to 500 features based on largest variance.
In the evaluation of the methods given below, we investigated three issues. Firstly, we pursued a comparison of the proposed method with pre-existing semi-supervised clustering methods. As our principal comparison we will use constrained K-means clustering (CKM) [2, 7] since this method has been widely used (we will only compare against constrained K-means clustering with must-links). Secondly, we will consider the possible gains to be made by using sample label information. Finally, we also compared the unsupervised ULPD with semi-supervised SLPD to evaluate the gains made by using side information.
We use the Balanced Rand Index (BRI) as evaluation criterion. Let nTS (nTD) be the true number of pairs of data in the same (different) clusters and nPS (nPD) be the correctly predicted number of pairs of data in the same (different) clusters. BRI is defined as

Data Sets	Letter {I, J}	Wine	Iris	Digit {3, 6, 8}	Leukemia	Lung Cancer
Number of Samples	150	178	150	174	90	203
Number of Features	16	13	4	16	500	500
Number of Classes	2	3	3	3	6	5

Table 1: Data sets used in the experimental study.

Data	UKM	CKM		ULPD	SLPD	
Sets	0	25%	50%	0	25%	50%
Letter	0.501±0.005	0.502±0.010	0.501±0.009	0.519±0.025	0.521±0.031	**0.527±0.039**
Wine	0.877±0.052	0.885±0.051	0.893±0.047	0.930±0.032	0.926±0.032	**0.935±0.032**
Iris	0.824±0.036	0.825±0.035	0.828±0.041	0.872±0.037	0.910±0.043	**0.920±0.038**
Digit	0.751±0.068	0.758±0.069	**0.772±0.078**	0.736±0.046	0.747±0.045	0.755±0.045

Table 2: A comparison of constrained K-means clustering and SLPD. The entries are the BRI (mean ± standard deviation over 100 trials using 3-fold cross validation tests). Hypothesis testing indicates a statistically significant performance gain over CKM.

$$BRI = 0.5\left(\frac{nPS}{nTS} + \frac{nPD}{nTD}\right) \tag{10}$$

The standard Rand Index is defined as $(nPS + nPD)/(nTS + nTD)$, which favors assigning data points to different clusters [14]. BRI favors matched pairs and mismatched pairs equally, hence it is more suitable for evaluating clustering algorithms.

In Table 2 we tabulate performance for both constrained K-means clustering and SLPD using the BRI. For each study, the whole data set was divided into 3 folds by random partition. One fold of the data set was used as a test set and a subset of the remaining two folds was used for supervised training. Exactly the same sample allocations were used in the evaluation of both constrained K-means clustering and SLPD. For the training set we also imposed a degree of supervision to compare unsupervised learning with semi-supervised clustering using different levels of supervision. The fraction of the data set used for supervision was 0% (unsupervised), 25% and 50%. With random resampling of the data, this 3-fold cross validation procedure was repeated over 100 trials. As observed from Tables 2 and 3, SLPD compares favorably with CKM. In addition, by enforcing a degree of supervision (from 0% to 50%), we can observe performance improvements of both SLPD and CKM over LPD and unconstrained K-means clustering (UKM).

As a real-life application we tested SLPD on two cancer microarray datasets for leukemia and lung cancer. We selected these datasets since labels can be reliably assigned to a large proportion of the samples and thus we can evaluate performance of the proposed metethod. Thus, for the leukemia dataset, we used a reduced 90-sample, 6-class subset of the data where labels can be assigned in our study (balanced to give 15 samples per class). However, for the full dataset there are a small number of samples with unclear assignment to subtype: this subset would be a good target for semi-supervised clustering since we would want to use all available information to characterise these samples as defining novel subtypes or as having a relation to known subtypes. The results are tabulated in Table 3, where the values are the average values of 50 runs. For the leukemia dataset, both unsupervised LPD and semi-supervised LPD were trained on 45 samples from 3 classes, with clustering performance evaluated on the other 45 samples from the remaining 3 classes. For leukemia we find that the BRI index improves as we increase the extent of supervision. We also find that SLPD consistently outperforms K-means clustering. A similar picture is repeated for lung cancer.

Constrained K-means clustering is faster to execute compared to SLPD. However, SLPD has important advantages over constrained K-means. K-means clustering can work well if K is known but the method does not have an intrinsic mechanism for determining the correct model complexity. With SLPD we can determine the appropriate number of clusters by determining the estimated log-likelihood on hold-out data in a cross-validation study (see Appendix 6.2).

In Figure 2 we illustrate this procedure using the wine data set mentioned earlier. This data set has 3 class labels and 178 samples. To determine the appropriate model complexity we find the averaged log-likelihood on 28 hold-out samples using the likelihood estimate given in the Appendix (section 6.2, the curves are averages over 100 runs, error bars are omitted to simplify the Figure). For unsupervised learning the peak is at 4 (solid curve). This result appears reasonable since a dendrogram decomposition

Data Sets	UKM	CKM		ULPD	SLPD	
	0	25%	50%	0	25%	50%
Leukemia	0.786±0.065	0.782±0.061	0.798±0.062	0.838±0.053	0.846±0.049	**0.851±0.048**
Lung Cancer	0.578±0.030	0.583±0.033	0.599±0.041	0.660±0.033	0.665±0.032	**0.670±0.039**

Table 3: A comparison of constrained K-means clustering and SLPD. The entries are the BRI (mean $\pm$ standard deviation over the 50 trials of 3-fold cross validation tests). Hypothesis testing indicates a statistically significant performance gain over CKM.

of this data set (Figure 2 (right)) suggests this is an appropriate decomposition. If all the class labels are used we get a sharp peak at 3 (upper dashed curve) as expected. However, if we use the labels of one class and leave the other two classes unlabelled we get a shallow peak at 3 (lower dotted curve). This illustrates the point that the class labelling may not necessarily reflect the underlying cluster structure, potentially negating any gains made by using side information.

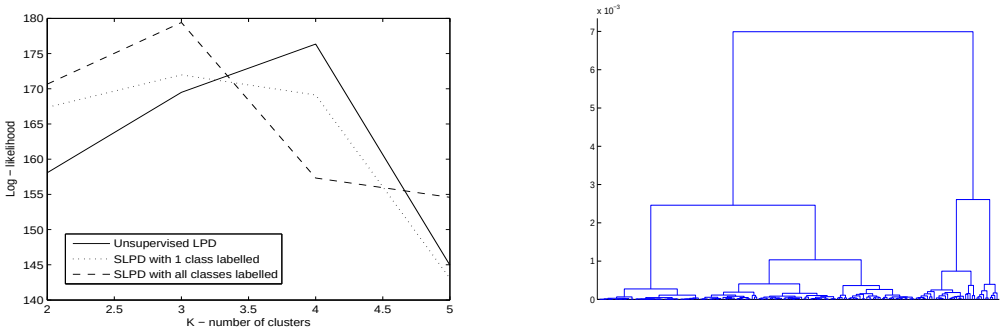

Figure 2: The left-hand Figure gives the Log-likelihood (y-axis) versus K for the wine data set for varying degrees of supervision (see text). The right-hand Figure gives a dendrogram partitioning of the same data set using a correlation distance measure and average linkage (only the upper branches of the dendrogram are illustrated).

5 Conclusion

In this paper we have presented a novel variational inference method for clustering with side information. The method provides a principled approach to model selection via determination of the log-likelihood on hold-out data, as illustrated in Figure 2. By contrast, many earlier methods for clustering with side information appear to lack a sound approach to establishing the correct model complexity. As a Bayesian methodology we could also implement prior beliefs which can improve model performance. Specifically, we can incorporate a Bayes prior penalizing overfitting to noise. Semi-supervised clustering methods have many important applications: we have mentioned cancer applications where partial labelling of the samples is sometimes possible. Determining the number of subtypes and avoiding a fit to the extensive noise present in many of these data sets are important issues in this application domain.

6 Appendix: derivation of the inference updates for SLPD

6.1 EM-updates

In the E-step, we compute the posterior distribution $q(E, \theta|\gamma, Q)$ and get the corresponding updates for γ, Q. To this end, we take the functional derivative of q, and get that

$$q(Z|Q) \propto \exp\big(\mathbb{E}_{q(\theta|\gamma)}\big[\log p(E, \theta, Z|\Theta, \mathcal{C})\big]\big), \tag{11}$$

and

$$q(\theta|\gamma) \propto \exp\big(\mathbb{E}_{q(Z|Q)}\big[\log p(E, \theta, Z|\Theta, \mathcal{C})\big]\big). \tag{12}$$

Note that the log joint likelihood can be expressed by

$$\begin{aligned}
&\log p(E,\theta,Z|\Theta,\mathcal{C}) \\
&= \textstyle\sum_{c,k}(\alpha_k-1)\log\theta_{ck} + C(\log\Gamma(\sum_k\alpha_k) - \sum_k\log\Gamma(\alpha_k)) \\
&+ \textstyle\sum_{c,d,g,k} Z_{dg,k}\delta_{dc}\log\theta_{ck} + \sum_{c,d,g,k} Z_{dg,k}\delta_{dc}\log\mathcal{N}(E_{dg}|\mu_{gk},\sigma_{gk}) \\
&= \textstyle\sum_{c,k}(\alpha_k-1)\log\theta_{ck} + C(\log\Gamma(\sum_k\alpha_k) - \sum_k\log\Gamma(\alpha_k)) \\
&+ \textstyle\sum_{c,d,g,k} Z_{dg,k}\delta_{dc}\log\theta_{ck} + \sum_{d,g,k} Z_{dg,k}\log\mathcal{N}(E_{dg}|\mu_{gk},\sigma_{gk}),
\end{aligned}$$

where we used the fact $\sum_c \delta_{dc} = 1$ for any d in the last equation. Note that $\mathbb{E}_{q(\theta_c|\gamma_c)}\big[\log\theta_{ck}\big] = \Psi(\gamma_{ck}) - \Psi(\sum_k\gamma_{ck})$ where Ψ is digamma function. Consequently, from the above log joint likelihood equation we have that

$$\begin{aligned}
&\mathbb{E}_{q(\theta|\gamma)}\big[\log p(E,\theta,Z|\Theta,\mathcal{C})\big] = \\
&\textstyle\sum_{d,g,k} Z_{dg,k}\Big(\sum_c\delta_{dc}(\Psi(\gamma_{ck}) - \Psi(\sum_k\gamma_{ck})) + \log\mathcal{N}(E_{dg}|\mu_{gk},\sigma_{gk}\Big),
\end{aligned}$$

and $\mathbb{E}_{q(Z|Q)}\big[\log p(E,\theta,Z|\Theta,\mathcal{C})\big] = \sum_{c,k}\log\theta_{ck}\big(\alpha_k - 1 + \sum_{d,g} Q_{dg,k}\delta_{dc}\big)$.
Therefore, putting the above equations into equations (11), (12) and normalizing Q we have the desired updates (6) and (7) in Section 3.
At the M-step, the updates (8) for μ,σ are easy to get by solving the stationary equations $\frac{\partial\mathcal{L}}{\partial\mu_{gk}} = 0$ and $\frac{\partial\mathcal{L}}{\partial\sigma^2_{gk}} = 0$. However, for the parameter α, observe that, in its derivative equation $\frac{\partial\mathcal{L}}{\partial\alpha_i} = C\Psi(\sum_k\alpha_k) - C\Psi(\alpha_i) + \sum_c(\Psi(\gamma_{ci}) - \Psi(\sum_k\gamma_{ck}))$, the variable α_i is mixed with other ones α_j with $j\neq i$, then we have to use iterative algorithm. Here, we employ a Newton-Raphson method to approximate the optimal α. For this purpose, we compute the Hessian as follows $H_{ij} = C(\Psi'(\sum_k\alpha_k) - \delta_{ij}\Psi'(\alpha_i))$. Hence, we have the iterative procedure

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

Because of the special form of the Hessian matrix (decomposable into diagonal and off-diagonal terms) we can avoid matrix inversion of H_{ij}, see Appendix of [5].

6.2 Computing the likelihood

Once the parameters have been estimated, we can calculate the likelihood using equation (5). By the law of large numbers in probability theory, we can estimate the expectation with respect to $p(\theta_c|\alpha)$ by averaging over large enough set of N samples:

$$p(E|\Theta,\mathcal{C}) = \prod_c\frac{1}{N}\sum_{n=1}^{N}\prod_d\left[\prod_g\sum_k\mathcal{N}(E_{dg}|\mu_{gk},\sigma_{gk})\theta_{ckn}\right]^{\delta_{dc}}.$$

where $\{\theta_{ckn}\}$ are N samples drawn from the Dirichlet distribution with parameter θ_c. Hence

$$\begin{aligned}
\log p(E|\Theta,\mathcal{C}) &= \sum_c\log\frac{1}{N}\sum_n\prod_d\left[\prod_g\sum_k\mathcal{N}(E_{dg}|\mu_{gk},\sigma_{gk})\theta_{ckn}\right]^{\delta_{dc}} \\
&= -C\log N + \sum_c\log\sum_n\prod_d\left[\prod_g\sum_k\mathcal{N}(E_{dg}|\mu_{gk},\sigma_{gk})\theta_{ckn}\right]^{\delta_{dc}}.
\end{aligned}$$

References

[1] A.A. Alizadeh et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403(3):503–511, February 2000.

[2] S. Basu, A. Banerjee, and R. J. Mooney. Semi-supervised clustering by seeding. In *Proceedings of International Conference on Machine Learning 2002*, pages 27–34, 2002.

[3] S. Basu, M. Bilenko, and R. J. Mooney. A probabilistic framework for semi-supervised clustering. In *KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 59–68, New York, NY, USA, 2004. ACM Press.

[4] A Bhattacharjee et al. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proceedings of the National Academy of Sciences*, 98:13790–13795, 2001.

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

[6] 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.

[7] A. Demiriz, K. Bennett, and M. Embrechts. Semi-supervised clustering using genetic algorithms, 1999.

[8] C.L. Blake D.J. Newman, S. Hettich and C.J. Merz. UCI repository of machine learning databases, http://www.ics.uci.edu/∼mlearn/mlrepository.html, 1998.

[9] E. Garber et al. Diversity of gene expression in adenocarcinoma of the lung. *Proceedings National Academy Sciences*, 98(24):12784–12789, 2001.

[10] Z. Ghahramani and M. J. Beal. *Graphical models and variational methods*. MIT Press, 2000.

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

[12] 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.

[13] K. Wagstaff and C. Cardie. Clustering with instance-level constraints. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 1103–1110, 2000.

[14] E. Xing, A. Ng, M. Jordan, and S. Russell. Distance metric learning, with application to clustering with side-information. In *Advances in NIPS*, number vol. 15, 2003.

[15] E-J Yeoh et al. Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell*, 1:133–143, 2002.