Chapter 4

A Two Step Bayesian Model for Matching Cell Line and Patient Genomic Profiles

4.1 Introduction

In a precision medicine paradigm, the patient's specific genetic architecture is assessed to propose a personalized treatment that is expected to be optimal for that individual. In this context, there is a trend in modern medicine to move from generalized treatment approaches towards the tailored treatment strategy that is dictated by their genomics or molecular profile. This paradigm shift accelerates the needs for advances in pharmacogenomics technology and associated analytical methods. In this paper, we develop methods to meet this demand, inclduing in particular novel priors for random structures. Briefly, we propose developing an integrative statistical framework, that merge multiplatform genomics ('omics, in short) profiles from multiple model systems (e.g. patients and cell-lines) for finding significant drug targets, preclinical models for appropriate drug discovery and repurposing and, finally, to calibrate therapeutic potential for future patients. By identifying these similarities (and differences) across model systems, we are able to gather more refined information about the patient than what is contained in their specific profile, while still proposing a personalized treatment that is strongly tied to the patient's profile - through appropriate integration of various data sources.

The objective of this paper is then to construct a novel Bayesian statistical approach for matching patient gene profiles with cell line profiles. Such inference is needed, among many other applications, for data integration, precision medicine and patient specific treatment assignment. The expansion of modern medicine and fast growth of research in health sciences have led to a great increase in available data on multiple sources/platforms such as The Cancer Genome Atlas (TCGA, `tcga-data.nci.nih.gov`), Cancer Cell Line Encyclopedia (CCLE, `portals.broadinstitute.org/ccle`), International Cancer Genome Consortium (ICGC, `icgc.org`) to name a few. A model-based approach for matching of a patient profile with data from other sources, such as from cell lines, allows us to access a wider range of information to predict a patients response to specific treatments. Important for the envisioned application, the matching should be carried out on the basis of a biologically meaningful signal only, putting aside mere noise.

A cell line is a culture of cells extracted from a tissue (e.g., cancer cells from a tumor in a human tissue) and grown in an in-vitro environment that simulates the environment of the tissue in the organism where it was extracted. Therefore, cell lines serve as models to study cancer biology. Information from the response to a drug or treatment applied to cell lines (cultivated *in-vitro*) is used to infer about the expected response *in vivo* (Goodspeed et al., 2016). Similarly, individuals can be grouped according to similarities between their profiles and observed profiles in a fixed set of cell lines. In such scenarios, the mapping of cell lines and patients opens the possibility to construct treatment recommendations based on results for the corresponding cell lines (Sinha et al., 2015). We propose a statistical approach that seamlessly combines the output of the Bayesian mixture model based on a proposal by Parmigiani et al. (2002) with a novel two-way Bayesian non-parametric (BNP) mixture model that is constructed as an extension of a BNP bi-clustering model of Lee et al. (2013).

Parmigiani et al. (2002) propose a Gaussian-uniform mixture model for probability of expression (POE). Later in the model construction we shall use the latent trinary signal of the POE model to carry out nested clustering of patient samples and the desired matching with cell lines. The uniform component in the POE models havier tails associated with genes that are over- and under-expressed, while the Gaussian term corresponds to regularly expressed genes. The authors argue that, by trichotomizing gene expressions into these 3 categorical levels, the POE approach smoothly removes uninformative biological and instrumental noise that is naturally present in genomic profile data, therefore strengthening downstream analysis.

The clustering of patient samples and the desired match with cell lines builds on a model developed of Lee et al. (2013), who present a Bayesian model (NoB-LoC) that identifies genes (columns) that are relevant for clustering of samples/individuals (rows). The identified genes are then partitioned in such a way that genes within the same subgroup (column-wise clusters) give rise to a common nested partition of individuals (row-wise clusters). The approach is motivated by the observation that high-dimensional protein profiles make it hard to find meaningful clusters of samples/individuals. Researchers therefore often restrict attention to groups of proteins that are expected to lead to more meaningful and interpretable results. NoB-LoC identies such groups in a seamless process, together with the nested clustering of samples. The NoB-LoC model conveniently allows for different clustering of samples with respect to different groups of proteins. In our context, this translates to association of cell lines and patients depending on the set of proteins in the profile.

Developing the outlined model consturction, this chapter makes two major contributions: the first one is the integration of POE with the two-way clustering building on the NoB-LoC model. The second, and perhaps more important contribution is the extension of the NoB-LoC model to allow for explicit probabilistic matching of profiles that could come from distinct sources (e.g., cell lines and patients). In short, in the proposed approach we first use the NoB-LoC model to partition the proteins according to a zero enriched Pólyia urn process where some proteins are set aside as inactive proteins, while the selected proteins are grouped into protein clusters (active proteins). Within each protein cluster, the samples are partitioned again, using a partition model that matches patients to cell lines. The motivation is that the usually high-dimension protein profiles make it hard to find similar samples to be clustered together, therefore restricting the attention to groups of proteins is expected to lead to more meaningful and interpretable results. This procedure also naturally allows for identification of co-expressed proteins in the form of protein clusters, i.e. a group of genes that are biologically correlated are also expected to have their expression levels ”tied together” along different samples. Finally, the NoB-LoC model conveniently allows for different clustering of samples depending on the subsample of proteins that is considered. In our problem, this translates to association of cell lines and patients depending on the set of proteins in the profile.

The real data used in the statistical analysis comes from an experiment using reverse phase protein arrays (RPPA) which record the expression of selected proteins simultaneously on multiple cell lines and patients samples (Charboneau et al., 2002). The dataset analyzed here consists of lung cancer protein expressions (233 proteins) measured in 687 patients and 124 cell lines. Data is batch corrected, i.e., they are also adjusted for the batch effect difference between cell line and patients' data).

4.2 POE Model

In this section we describe the POE (probability of expression) model defined in Parmigiani et al. (2002). We modified some of the priors in order to obtain analytical full-conditionals for as many parameters as possible, which facilitates the MCMC implementation in the larger, encompassing model (more details ahead and also in appendix D.1).

Each observation y_{sg} consists of expression levels for protein (gene) $g \in \{1, \ldots, G\}$ and sample $s \in \{1, \ldots, S\}$. Latent variables e_{sg} indicate high expression of gene g in sample t ($e_{sg} = 1$), normal expression ($e_{sg} = 0$) and under expression ($e_{sg} = -1$). Each possible value of e_{sg} determines a different distribution for the observed gene expressions according to the following Gaussian-Uniform mixture model

$$(y_{sg} \mid e_{sg}) \sim \begin{cases} Unif(\alpha_s + \mu_g,\ \alpha_s + \mu_g + k_g^+), & \text{if } e_{sg} = 1, \\ N(\alpha_s + \mu_g,\ \sigma_g^2), & \text{if } e_{sg} = 0, \\ Unif(\alpha_s + \mu_g - k_g^-,\ \alpha_s + \mu_g), & \text{if } e_{sg} = -1. \end{cases}$$

The lengths k_g^+ and k_g^- of the support of the uniform components should cover the tails of the corresponding gene expression distribution implying heavier tails than the Gaussian distribution. Under normality, the great majority of the samples (probability 0.997) concentrate within 3 standard deviations from the mean; therefore the constraints $k_g^+ > k_0\sigma_g,\ k_g^- > k_0\sigma_g$ imply heavier than Gaussian tails for fixed values of k_0 greater than, say, 3.

We now define the weights for each term in the mixture by the probability vectors $\boldsymbol{\pi}_g := (\pi_g^-, \pi_g^0, \pi_g^+),\ g \in \{1, \ldots, G\}$ where $\pi_g^+ = P(e_{sg} = 1 \mid \boldsymbol{\pi}_g)$, $\pi_g^0 = P(e_{sg} = 0 \mid \boldsymbol{\pi}_g)$ and $\pi_g^- = P(e_{sg} = -1 \mid \boldsymbol{\pi}_g)$. We assume $(\boldsymbol{\pi}_g \mid \boldsymbol{\eta}_\pi) \sim Dirichlet(\boldsymbol{\eta}_\pi)$.

Figure 4.1 illustrates the implied mixture model in the context of density estimation. The augmentation of the parameter space with inclusion of indicatior variables e_{st} allows for identification of up- and down-regulated genes that are not well captured bythe light tails of a single Gaussian component.

Posterior probabilities of differential expression are determined by Bayes

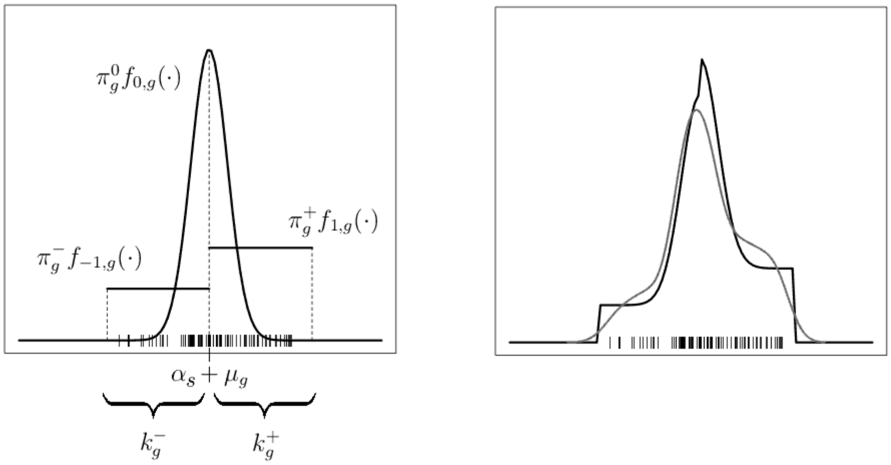


Figure 4.1: Left panel: Weighted components of the Gaussian-Uniform mixture model. Right panel: density estimates using Gaussian-Uniform mixture (black line) and using kernel estimate (gray line). Vertical bars represent data generated from the Gaussian-Uniform mixture.

rule as

$$\begin{aligned} p_{sg}^+ &:= P(e_{sg} = 1 \mid y_{sg}, \boldsymbol{\pi}_g, f_{1,g}, f_{0,g}) \\ &= \frac{\pi_g^+ f_{1,g}(y_{sg})}{\pi_g^+ f_{1,g}(y_{sg}) + \pi_g^0 f_{0,g}(y_{sg}) + \pi_g^- f_{1,g}(y_{sg})} \times \mathbb{1}(y_{sg} \in S_{f_{1,g}}) \\ &= \frac{\pi_g^+ f_{1,g}(y_{sg})}{\pi_g^+ f_{1,g}(y_{sg}) + \pi_g^0 f_{0,g}(y_{sg})} \times \mathbb{1}(y_{sg} \in S_{f_{1,g}}), \end{aligned} \tag{4.1}$$

where $\mathbb{1}(\cdot)$ is the indicator function and $S_{f_{1,g}}$ denotes the support of $f_{1,g}$. Analogously,

$$\begin{aligned} p_{sg}^- &:= P(e_{sg} = -1 \mid y_{sg}, \boldsymbol{\pi}_g, f_{-1,g}, f_{0,g}) \\ &= \frac{\pi_g^- f_{-1,g}(y_{sg})}{\pi_g^- f_{-1,g}(y_{sg}) + \pi_g^0 f_{0,g}(y_{sg})} \times \mathbb{1}(y_{sg} \in S_{f_{-1,g}}). \end{aligned}$$

Equations (4.1) and (**??**) are used for visualization of sample specific gene profiles. Since p^+_{sg} and p^-_{sg} are not simultaneously positive, the differences $d_{sg} := p^+_{sg} - p^-_{sg}$ will fall in the interval $[-1, 1]$, therefore serving as a unidimensional measure of gene expression ($d_{sg} \approx 1$ for highly expressed and $d_{sg} \approx -1$ for underexpressed genes).

and also for our sequential inference approach to serve as input for the followup application of the NoBloC model.

The model is completed by the prior specification $(\mu_g \mid \theta_\mu, \tau_\mu) \sim N(\theta_\mu, \tau_\mu)$, $(\alpha_s \mid \mu_\alpha, \tau_\alpha) \sim N(\mu_\alpha, \tau_\alpha)$ restricted to $\sum_{s=1}^{S} \alpha_s = 0$, $(\sigma^2_g \mid \gamma, \lambda) \sim InvGamma(\gamma, \lambda)$, $(k^+_g \mid \alpha_{k^+}, \beta_{k^+}) \sim InvGamma(\alpha_{k^+}, \beta_{k^+})$, $(k^-_g \mid \alpha_{k^-}, \beta_{k^-}) \sim InvGamma(\alpha_{k^-}, \beta_{k^-})$. We also chose prior models for hyperparameters as $\theta_\mu \sim N(m_\mu, s^2_\mu)$, $\tau_\mu \sim InvGamma(a_{\tau_\mu}, b_{\tau_\mu})$, $\alpha_{k^+} \sim Exp(\lambda_{\alpha_{k^+}})$, $\alpha_{k^-} \sim Exp(\lambda_{\alpha_{k^-}})$, $\beta_{k^+} \sim Gamma(a_{\beta_{k^+}}, b_{\beta_{k^+}})$, $\beta_{k^-} \sim Gamma(a_{\beta_{k^-}}, b_{\beta_{k^-}})$.

The motivation to propose Inverse Gamma priors for k^+_g, k^-_g and Dirichlet prior for $\boldsymbol{\pi}_g$ is to make use of conjugacy results in the full-conditional posterior of these parameters, which was not originally explored in Parmigiani et al. (2002).

4.2.1 Posterior inference for the POE model

We implement posterior inference by MCMC simulation. All full conditionals are available in closed form due to the choice of contditionally conjugate priors/hyperpriors; the only exceptions are α_{k^+} and α_{k^-}. We therefore implement Gibbs sampling transition probabilities for all parameters except

$(\alpha_{k^+} \mid \boldsymbol{y}, else)$ and $(\alpha_{k^-} \mid \boldsymbol{y}, else)$. For the latter we use Metropolis-Hastings transition probabilities with random walk proposal on $\log \alpha_{k^+}$ and $\log \alpha_{k^+}$ respectively. See appendix D.1 for more details.

To avoid numerical instability when sampling from truncated inverse gamma distributions (full conditional posterior distributions for k_g^+ and k_g^-), we used a variable augmentation scheme proposed in Damien and Walker (2001). The prior on the auxiliary variables introduced by the authors imply full-conditional posterior distributions for those variables, which are sampled together with the original parameters of the POE model within the full MCMC algorithm.

4.3 Nonparametric Bayesian Clustering with Patient and Cell Line Matching

4.3.1 A nested random partition and matching structure

Following posterior simulation for the POE model, the posterior estimated values d_{sg} become the inputs for model-based clustering of proteins and samples and the desired pairing with cell lines. This is implemented by the construction of a nested partition model that builds on the Nonparametric Bayesian local clustering (NoB-LoC) model defined in Lee et al. (2013). In this section, we relabel the data d_{sg} as d_{ig}^c if sample s corresponds to the i-th cell line or as d_{ig}^p if it is the i-th patient. The subindex g still denotes protein g. We will assume the dataset contains G proteins and S samples, including N^p patient samples and N^c cell line samples ($N^p + N^c = S$).

The model first partitions the proteins according to a zero enriched Pólya urn. One special cluster (corresponding to the zero-enrichment) is interpreted as "inactive proteins". The remaining ones are grouped into protein clusters (active proteins). Within each of these protein clusters, the samples are partitioned again by a second, nested partition model. The nested partition model includes also the desired pairing of each patient sample cluster with a matching cell line.

Consider the cluster membership indicator w_g for each protein $g = 1, \dots G$ and denote $\boldsymbol{w} = (w_1, \dots, w_G)$ the vector of protein cluster indicators. Let π_0 be the probability of inactivation and let $\alpha_0 > 0$ be the potential for creating a new cluster. Finally, define $n_k := \#\{g;\ w_g = k\}$ the number of proteins that fall into protein cluster k, for $k = 0, 1, \dots, K_{\boldsymbol{w}}$ with $k = 0$ denoting the cluster of inactive proteins and $K_{\boldsymbol{w}}$ denoting the total number of active clusters of proteins determined by $\boldsymbol{w}$. Then the zero enriched Pólyia urn defines $p(\boldsymbol{w} \mid \pi_0)$ as

$$p(\boldsymbol{w} \mid \pi_0) = \pi_0^{n_0}(1-\pi_0)^{G-n_0} \times \frac{\alpha_0^{K_{\boldsymbol{w}}} \prod_{k=1}^{K_{\boldsymbol{w}}} \Gamma(n_k)}{\prod_{g=1}^{G}(\alpha_0 + g - 1)}, \tag{4.2}$$

and for short we write $(\boldsymbol{w} \mid \alpha_0, \pi_0) \sim ZEPU(\alpha_0, \pi_0)$.

For each cluster of proteins defined by $\boldsymbol{w}$, two dependent partitions of the samples are defined, the first one involving patients only, and the second one (which is stochastically dependent on the partition of patients) includes only cell lines. The cluster membership indicators for the i-th patient and i-th cell line in k-th cluster of proteins are defined as $\delta_i^{p,k}$ and $\delta_i^{c,k}$, respectively. The two partitions are then determined by the cluster membership indicators for patients and for cell lines. We marginally model the cluster membership of patients $\boldsymbol{\delta}^{p,k} := (\delta_i^{p,k})_{i=1}^{N^p}$ as $\boldsymbol{\delta}^{p,k} \sim ZEPU(\alpha_{pk}, \pi_{pk})$. This implies a random number J_k of active clusters of patients within the k-th group of proteins.

Conditionally on $\boldsymbol{\delta}^{p,k}$, several choices of priors on $\boldsymbol{\delta}^{c,k}$ are possible, each of them representing one way of matching cell lines' to patients' profiles.

Discrete uniform prior: we assume a discrete uniform prior for $\boldsymbol{\delta}^{c,k}$ on the patient samples and the set of inactive samples: $\delta_i^{c,k} \mid \boldsymbol{\delta}^{p,k} \sim Uniform(\{0, 1, \ldots, J_k\})$.

Discrete uniform prior with at most ℓ cell lines per cluster of samples: the conditional prior on $(\boldsymbol{\delta}^{c,k} \mid \boldsymbol{\delta}^{p,k})$ has the p.m.f

$$p(\boldsymbol{\delta}^{c,k} \mid \boldsymbol{\delta}^{p,k}) \propto \mathbb{1}(\boldsymbol{\delta}^{c,k} \in \mathcal{B}_\ell^{c,k}) \tag{4.3}$$

with support

$$\mathcal{B}_\ell^{c,k} = \left\{ (\delta_i^{c,k})_{i=1}^{N^c} \in \{0, 1, \ldots, N^c\}^{N_c}; \quad \textstyle\sum_{i=1}^{N^c} \mathbb{1}(\delta_i^{c,k} = j) \leq \ell \ \forall j \in \{1, \ldots J_k\} \right\}.$$

In the special case $\ell = 1$ for example, the multiplicative normalization constant in equation (4.3) is $\sum_{n=0}^{\min\{J_k, N^c\}} \binom{J_k}{n}\binom{N^c}{n} n!$.

Conditional zero inflated Polya urn: a conditional zero inflated Polya urn prior distribution is assumed prior for $\boldsymbol{\delta}^{c,k} \mid \boldsymbol{\delta}^{p,k}$ as a continuation of the process of clustering patient samples. This means that the initial probability of allocation of cell lines to a patient cluster is proportional to the cluster size. Stochastically, this approach is equivalent to a joint zero inflated Polya urn for all samples.

We consider the Gaussian sampling model $(d^c_{ig} \mid \theta^c_{ig}, \sigma^2_g) \sim N(\theta^c_{ig}, \sigma^2_g)$ and $(d^p_{jg} \mid \theta^p_{ig}, \sigma^2_g) \sim N(\theta^p_{jg}, \sigma^2_g)$ for cell line i and patient j and protein g. The prior for θ^x_{ig} with x being either c or p is

$$\theta^x_{ig} \sim \begin{cases} I_{\theta^*_{jg}}, & \text{if } \delta^{x,w_g}_i = j > 0 \text{ and } w_g > 0 \text{ (active prot. and smpl.)} \\ N(\mu_{1g}, \sigma^2_{1g}), & \text{if } \delta^{x,w_g}_i = 0 \text{ and } w_g > 0 \text{ (active prot. inactive smpl.)} \\ N(\mu_{2g}, \sigma^2_{2g}), & \text{if } w_g = 0 \text{ (inactive prot.),} \end{cases} \tag{4.4}$$

where I_x denotes the point mass distribution (Dirac measure) at x. In (4.4) (first equation) we define the unique mean responses θ^*_{jg} for active proteins and samples. We denote by J_k the number of active sample clusters for all proteins g such that $w_g = k$. Notice that active cell lines and patients share the same mean response if they belong to the same sample cluster.

For the purpose of deriving the MCMC algorithm for posterior inference, we marginalize the patient specific (and cell line specific) means θ^x_{ig} in (4.4), which implies the following data distribution

$$d^x_{ig} \sim \begin{cases} N(\theta^*_{jg}, \sigma^2_g) & \text{if } \delta^{x,w_g}_i = j > 0 \text{ and } w_g > 0 \text{ (active prot. and smpl.)} \\ N(\mu_{1g}, \sigma^2_{1g} + \sigma^2_g), & \text{if } \delta^{x,w_g}_i = 0 \text{ and } w_g > 0 \text{ (active prot. inactive smpl.)} \\ N(\mu_{2g}, \sigma^2_{2g} + \sigma^2_g), & \text{if } w_g = 0 \text{ (inactive prot.).} \end{cases} \tag{4.5}$$

The prior for the unique mean response values is specified as $\theta^*_{jg} \sim N(\mu_{0g}, \sigma^2_{0g})$ with hyperpriors $\sigma_g^{-2} \sim Gamma(a_g, b_g)$, $\tau_{lg}^{-2} \sim Gamma(a_{lg}, b_{lg})$ and $\mu_{0g},\ \mu_{1g},\ \mu_{2g} \overset{iid}{\sim} N(m_0, s_0^2)$.

4.3.2 Summarizing the posterior nested partition

Point estimates of the cluster-membership indicators are obtained using the approach proposed by Dahl (2006). We run the MCMC algorithm and, after judging (practical) convergence, we evaluate for each pair $i < j$ of tumors within gene g, the pairwise co-clustering probability $\hat{p}_{ij} = \frac{1}{K}\sum_k p_{ijk}$, where K is the Monte Carlo sample size and p_{ijk} is an indicator for i and j being allocated to the same cluster, i.e., $p_{ijk} = 1 \Leftrightarrow e_{gi} = e_{gj}$ during iteration k. The dependence on g is omitted from the notation for clarity. The p_{ijk} and the $\hat{p}_{ij}$ are combined into $(I \times I)$ matrices $\boldsymbol{P}^{(k)} = [p_{ijk}]$ and $\hat{\boldsymbol{P}} = [\hat{p}_{ij}]$. We then report as posterior estimated $\bar{\delta}$ the partition corresponding to the co-clustering matrix $\boldsymbol{P}^{(k^*)}$ that minimizes $||\hat{\boldsymbol{P}} - \boldsymbol{P}^{(k)}||$. In other words,

$$\boldsymbol{P}^{(k^*)} = \arg\min_k ||\hat{\boldsymbol{P}} - \boldsymbol{P}^{(k)}||.$$

That is, k^* indexes the Monte Carlo sample whose co-clustering matrix is closest to $\hat{\boldsymbol{P}}$. The procedure for choosing k^* is done independently over each gene g.

4.4 Simulation

4.4.1 Simulation 1: POE

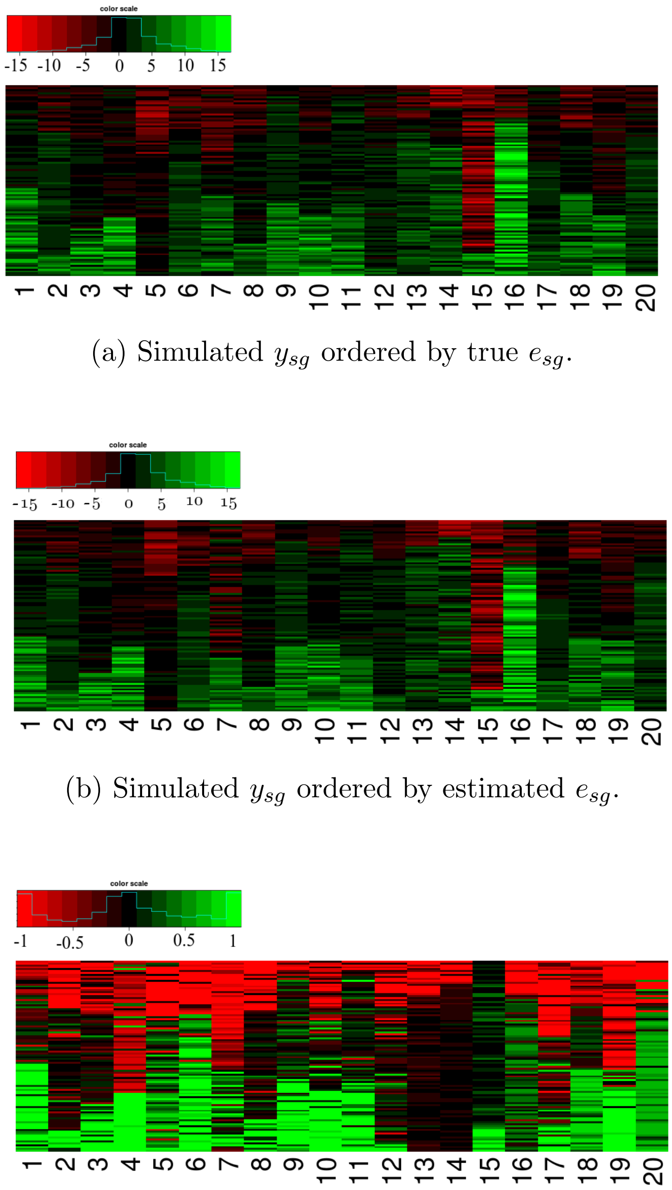
We carry out a first simulation to validate inference under the POE model.

We simulate a dataset with 100 samples and 20 genes assuming the POE model as the underlying truth. Some small changes were done in the simulation process that deviates slightly from the model in section 4.2. Namely, σ_g^2 was sampled from $\sigma_g^2 = N(0, 0.25)^2 + 1$ instead of an Inverse Gamma prior and k_g^+, k_g^- were both sampled from $\max(Gamma(8,1),\ 5\sigma_g)$ instead of $\max(InvGamma(8,1),\ 5\sigma_g)$. Hyperparameters were fixed as $\boldsymbol{\eta}_g = (1,1,1)$, $\mu_\alpha = 0$, $\tau_\alpha = 0.5$, $\theta_\mu = \tau_\mu = 1$.

To carry out the MCMC inference procedure, we fix $\eta_g = (1,1,1)$, $\mu_\alpha = 0$, $\tau_\alpha = 100$, $a_{\tau_\mu} = b_{\tau_\mu} = a_{\beta_{k+}} = b_{\beta_{k+}} = a_{\beta_{k-}} = b_{\beta_{k-}} = \lambda_{\alpha_{k+}} = \lambda_{\alpha_{k-}} = 0.01$, $\gamma = \lambda = 0.1$, $m_\mu = 0$ and $s_\mu^2 = 100$. Such values were chosen to represent weak prior information.

Figure 4.2 shows that the estimated cluster membership assignment of the observations reasonably recovers the simulation truth (compare pannels (a) and (b)). Panel (c) shows how the POE model removes noise from the data and highlights the biologically meaningful levels of protein activation (low, medium, high).

Figure 4.3 compares the true protein wise cluster assignment with the point estimates obtained with the methodology from Dahl (2006) as described also in 4.3.2. The cluster membership indicators are typically well recovered. However it should be noticed that proteins 13 and 14 seem to mix together the samples with low and medium protein expressions (proteins 13 and 14), while within protein 4 the model seems to mix the samples of high and medium expression. This probably happens when 2 components are enought to estimate the underlying density among the different samples. Also, the use of uniform components can in some cases exhibit high density near the center of the resulting mixed distribution therefore producing samples of medium expression even when the true is $e_{sg} = \pm 1$ in the simulation (see Figure 4.4)

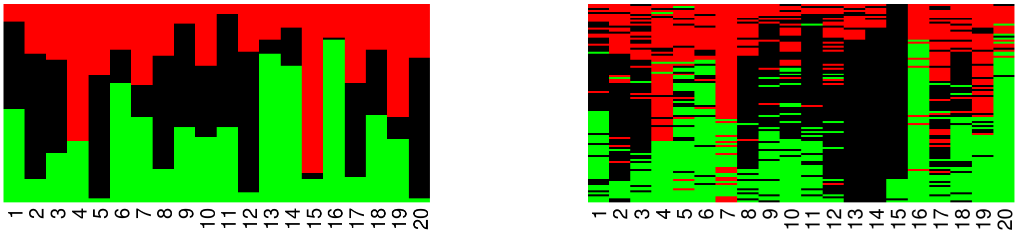


(a) Simulated y_{sg} ordered by true e_{sg}.

(b) Simulated y_{sg} ordered by estimated e_{sg}.

(c) d_{sg} ordered by true e_{sg}.

Figure 4.2: (a): Simulated data y_{sg}. Samples in each column are sorted by true e_{sg}. (b) Simulated data y_{sg}. Samples in each column are sorted by estimated $E\left(e_{sg} \mid \boldsymbol{y}\right)$. (c) Differences $d_{sg} = p^{+}_{sg} - p^{-}_{sg}$ with the same ordering as in panel (a). The ordering of samples change according to the protein (column) but is the same throughout the 3 panels.



(a) Simulated e_{sg}.

(b) Estimated e_{sg}.

Figure 4.3: (a) Simulation truth: $e_{sg} = 1$ (green), $e_{sg} = 0$ (black), $e_{sg} = -1$ (red). (b) Point estimates of e_{sg}. Rows represent tumors (samples) while columns represent genes (proteins). Each column is ordered in a different way within each heatmap. The ordering of samples change according to the protein (column) but is the same throughout the 2 panels.

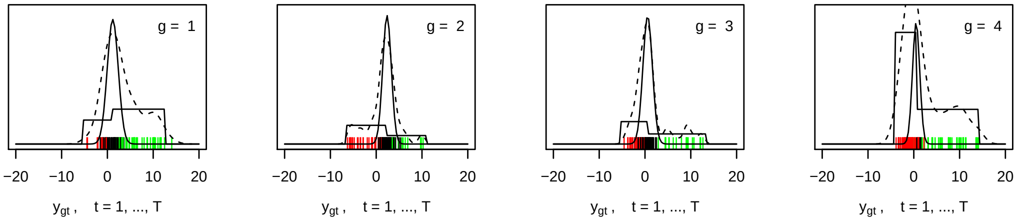
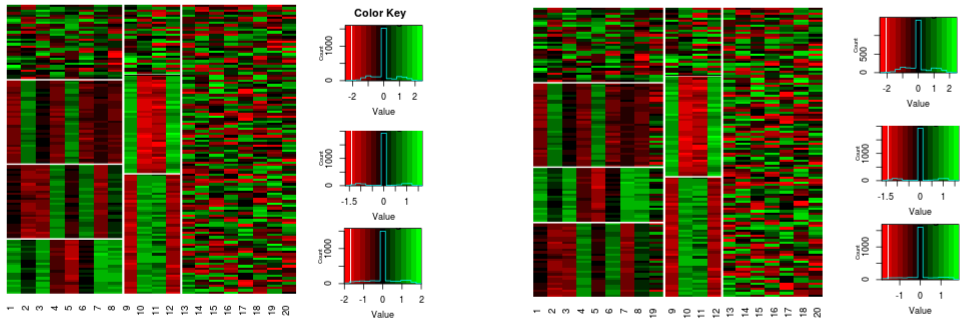


Figure 4.4: Posterior density estimates. Vertical bars represent gene expressions y_{sg}, $t = 1, ..., T$ for all genes g collored according to its estimated cluster membersip indicator (black = -1, red = 0, green = 1). Full lines represent the best fitting uniform and normal components of the mixture a posteriori multiplied by the respective weights. Dashed line corresponds to a kernel density estimate based on the vertical bars.

4.4.2 Simulation 2: nested partitions

In this section we describe the simulation to validate inference on the NobLoc model. We replicated the scenario in Lee et al. (2013), with 100 samples and 20 proteins. The simulation truth incorporates the local clustering feature of first partitioning proteins and then within protein cluster, partitioning the samples. However, instead of simulating the protein and sample partitions according to a zero inflated Plya urn, we fixed the partition of proteins upfront to have two active protein clusters, the first one containing proteins with 3 active sample clusters; and the second containing 4 proteins with 2 active sample clusters. The cluster-membership assignment of samples was made uniformly at random among the available sample clusters. The cluster specific means were fixed at the same values in Table 1 of Lee et al. (2013). Inactive samples and proteins were all sampled from $Unif(-0.8, 0.8)$. The standard deviation of the Gaussian sampling model for active samples was fixed at $\sigma_g = 0.1$ For an illustration, see Figure 4.5 panel (a).

In Figure 4.5 panel (b) we can see that the underlying cluster structure was reasonably captured by the NobLoc model. The only discrepancy is the inclusion of protein 19 in the first active cluster together with proteins 1 - 8, instead of classifying it as an inactive protein according to the simulation truth.



(a) Simulated y_{sg} ordered by true cluster assignments.

(b) Simulated y_{sg} ordered by estimated cluster assignments.

Figure 4.5: (a): Observations ordered according to the simulation truth. (b): Observations ordered according to estimated cluster membership indicators a posteriori. In both panels, rows represent samples while columns represent proteins.

4.5 Lung Cancer Dataset

4.5.1 The data

The dataset consists of protein profiles coming from an RPPA experiment on lung cancer samples. The data records 233 proteins that were preselected for their biological relevance to the study of this type of cancer. The data records samples from 687 patients and 124 cell lines. The objective is to identify groups of similar patients and cell lines with respect to subgroups of co-expressed proteins (we informaly say that proteins are co-expressed if their expressions are correlated). We therefore expect that the samples (patients and cell lines) can be partitioned in a different way depending on the group of co-expressed proteins that is considered.

4.5.2 Results

We describe here the results of a joint inference of the POE model and nested clustering by NoB-LoC. We start by analyzing the results of directly applying the NoB-LC model to the original lung data (without running POE first), which is illustrated in Figure 4.6 (a).

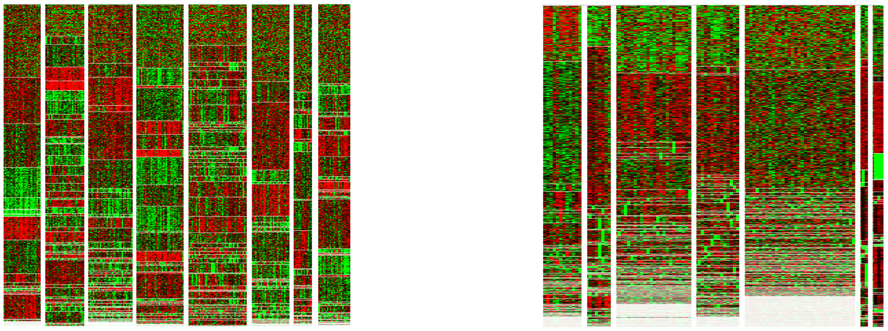


Figure 4.6: Observed protein expression arranged according to posterior estimated cluster structure under NoB-Loc. Only active proteins are displayed. Panel (a) shows the result of application of the NobLoc model on the original data and panel (b) shows the results after the application of POE.

Figure 4.7 shows one of the blocks in Figure 4.6 in more details highlighting the similarities between the cell lines and proteins in that block. Samples are reasonably homogeneous in terms of the expressions of the particular subgroup of proteins shown in the figure. Notice also that some proteins present typically high expression while others typically present low expression considering the particular group of samples.

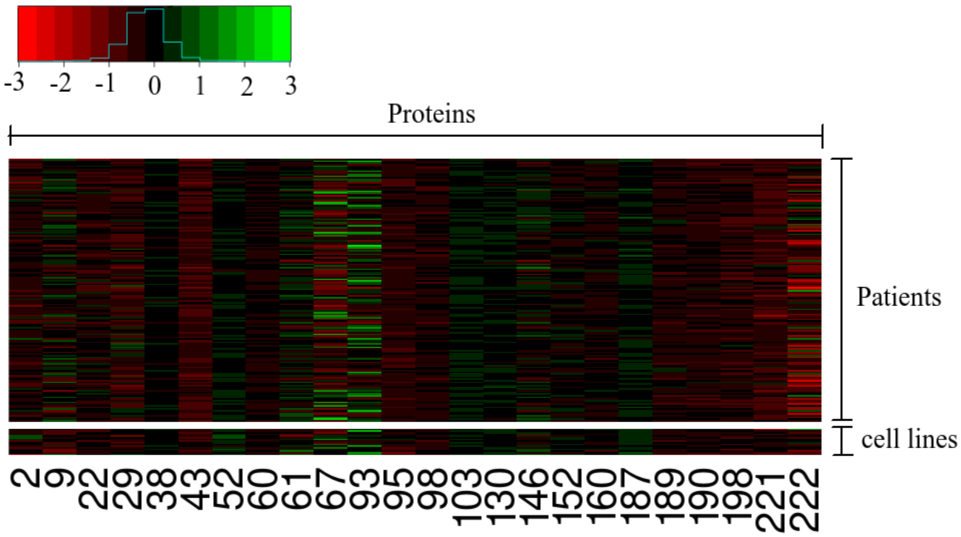


Figure 4.7: Protein expressions within one of the samples/proteins blocks of Figure 4.6 (b). The cell lines and patients exhibit very similar profiles when considering the subset of proteins that were clustered together by the model.

4.6 Discussion and Future Directions

Some innovations were introduced in this chapter: (i) The model naturally accounts for co-expression of genes and/or proteins in a manner that allows distinct clustering of samples depending on each estimated subgroup of co-expressed functional units or pathways; (ii) We develop efficient models for matching patients and cell line profiles. Development of such inferential tools is of critical relevance to implementing precision medicine, since it can be used for potential treatment assignment that is specific to a patient, and borrows information not only from that patient, but also from cell lines coming from external sources. (iii) Third, the approach does not need to be restricted to cell lines and patients but can be generalized to multiplatform omics profiles from diverse model systems such as patient-derived xenographs (PDX) models and organoids.

This chapter makes two important methodological contributions in Bayesian non-parametrics: (i) the seamless integration of the (modified) probability of expression (POE) model for noise reduction and the nested biclustering approach; (ii) the formal probabilistic modeling of co-clustering between proteins/genes and patients based on profile similarities via dependent priors on partition models.

There are some areas that need to be more thoroughly investigated. One of them is the need to extend the simulation studies where the underlying truth differs from the POE and NobLoc models in order to address the performance of the proposed methodology under model misspecification. In our results, we restricted the matching of cell lines and patients to a conditional zero inflated Plya urn (see section 4.3.1), therefore the investigation of the other models for matching information from cell lines to patients is also proposed as a future work.

Appendix D

Appendix for Chapter 4

D.1 Full Conditionals for the POE Model

Here we briefly describe how to sample from each one of the full conditionals. In this section, we use $\mathbb{1}(\cdot)$ to denote either an indicator function or th support of a truncated probability distribution. We also define the sets $\mathcal{P}_g^+ := \{1 \leq t \leq T : \ e_{sg} = 1\}$ and $\mathcal{P}_t^+ := \{1 \leq g \leq G : \ e_{sg} = 1\}$. The sets $\mathcal{P}_g^0$, $\mathcal{P}_g^-$, $\mathcal{P}_t^0$ and $\mathcal{P}_t^-$ are defined analogously.

Updating e_{sg}

$$p(e_{sg} = x \mid y, else) \propto \begin{cases} \frac{\pi_g^+}{k_g^+} \times \mathbb{1}(\alpha_s + \mu_g < y_{sg} < \alpha_s + \mu_g + k_g^+), & \text{if } x = 1, \\ \pi_g^0 \times N(y_{sg};\ \alpha_s + \mu_g,\ \sigma_g^2), & \text{if } x = 0, \\ \frac{\pi_g^-}{k_g^-} \times \mathbb{1}(\alpha_s + \mu_g - k_g^- < y_{sg} < \alpha_s + \mu_g), & \text{if } x = -1. \end{cases}$$

Updating k_g^+ and k_g^-

$$(k_g^+ \mid y, else) \sim InvGamma(\#\mathcal{P}_g^+ + \alpha_{k^+},\ \beta_{k^+}) \\ \mathbb{1}\left(k_g^+ > \max\left(\max_{t\in\mathcal{P}_g^+}\{y_{sg} - \alpha_s - \mu_g\},\ k_0\sigma_g\right)\right).$$

$$(k_g^- \mid y, else) \sim InvGamma(\#\mathcal{P}_g^- + \alpha_{k^-},\ \beta_{k^-}) \\ \mathbb{1}\left(k_g^- > \max\left(\max_{t\in\mathcal{P}_g^-}\{\alpha_s + \mu_g - y_{sg}\},\ k_0\sigma_g\right)\right).$$

Updating $\boldsymbol{\pi}_g$

$$(\boldsymbol{\pi}_g \mid y, else) \sim Dirichlet\left((\ \#\mathcal{P}_g^- + \alpha_\pi^-,\ \#\mathcal{P}_g^0 + \alpha_\pi^0,\ \#\mathcal{P}_g^+ + \alpha_\pi^+)\right).$$

Updating σ_g^2

$$(\sigma_g^2 \mid y, else) \sim InvGamma\left(\frac{\#\mathcal{P}_g^0}{2} + \gamma,\ \frac{1}{2}\sum_{t\in\mathcal{P}_g^0}(y_{sg} - \alpha_s - \mu_g)^2 + \lambda\right) \\ \mathbb{1}\left(\sigma_g^2 < \frac{\min\left(k_g^+, k_g^-\right)^2}{k_0^2}\right).$$

Updating μ_g

$$(\mu_g \mid y, else) \sim N(a_g, b_g)\mathbb{1}\left(\max\left(M_g^+, M_g^-\right) < \mu_g < \min\left(m_g^+, m_g^-\right)\right),$$

where

$$
\begin{aligned}
M_g^+ &= \max\{y_{sg} - \alpha_s;\ t \in \mathcal{P}_g^+\} - k_g^+;\\
M_g^- &= \max\{y_{sg} - \alpha_s;\ t \in \mathcal{P}_g^-\};\\
m_g^+ &= \min\{y_{sg} - \alpha_s;\ t \in \mathcal{P}_g^+\};\\
m_g^+ &= \min\{y_{sg} - \alpha_s;\ t \in \mathcal{P}_g^-\} + k_g^-;\\
b_g &= (\#\mathcal{P}_g^0 \sigma_g^{-2} + \tau_\mu^{-1})^{-1};\\
a_g &= b_g \times \left[\sigma_g^{-2} \sum_{t \in \mathcal{P}_g^0} (y_{sg} - \alpha_s) + \tau_\mu^{-1}\theta_\mu\right].
\end{aligned}
$$

Updating α_s

$$
(\alpha_s \mid y, else) \sim N(a_t, b_t)
$$

$$
\mathbb{1}\left(\max\left(M_t^+, M_t^-\right) < \alpha_s < \min\left(m_t^+, m_t^-\right)\right) \mathbb{1}\left(\sum_{t=1}^{T} \alpha_s = 0\right).
$$

where

$$
\begin{aligned}
M_t^+ &= \max\{y_{sg} - \mu_g - k_g^+;\ g \in \mathcal{P}_t^+\};\\
M_t^- &= \max\{y_{sg} - \mu_g;\ g \in \mathcal{P}_t^-\};\\
m_t^+ &= \min\{y_{sg} - \mu_g;\ g \in \mathcal{P}_t^+\};\\
m_t^+ &= \min\{y_{sg} - \mu_g + k_g^-;\ g \in \mathcal{P}_t^-\};\\
b_t &= \left(\sum_{g \in \mathcal{P}_t^0} \sigma_g^{-2} + \tau_\alpha^{-1}\right)^{-1};\\
a_t &= b_t \times \left[\sum_{t \in \mathcal{P}_t^0} \left(\frac{y_{sg} - \alpha_s}{\sigma_g^2}\right) + \tau_\alpha^{-1}\mu_\alpha\right].
\end{aligned}
$$

Updating θ_μ

$$(\theta_\mu \mid y, else) \sim N\left(\left(m_\mu s_\mu^{-2} + \tau_\mu^{-1}\sum_{g=1}^{G}\mu_g\right)(s_\mu^{-2} + G\tau_\mu^{-1})^{-1},\ (s_\mu^{-2} + G\tau_\mu^{-1})^{-1}\right)$$

Updating τ_μ

$$(\tau_\mu \mid y, else) \sim InvGamma\left(\frac{G}{2} + a_{\tau_\mu},\ \frac{1}{2}\sum_{g=1}^{G}(\mu_g - \theta_\mu)^2 + b_{\tau_\mu}\right)$$

Updating β_{k^+}

$$(\beta_{k^+} \mid y, else) \sim Gamma\left(G\alpha_{k^+} + a_{\beta_{k^+}},\ b_{\beta_{k^+}} + \sum_{g=1}^{G}\frac{1}{k_g^+}\right)$$

Updating β_{k^-}

$$(\beta_{k^-} \mid y, else) \sim Gamma\left(G\alpha_{k^-} + a_{\beta_{k^-}},\ b_{\beta_{k^-}} + \sum_{g=1}^{G}\frac{1}{k_g^-}\right)$$

Updating α_{k^+}

$p(\alpha_{k^+} \mid y, else)$ is not analytically available since

$$p(\alpha_{k^+} \mid y, else) \propto \Gamma(\alpha_{k^+})^{-G}\left(\frac{\beta_{k^+}^G}{\prod_{g=1}^{G} k_g^+}\right)^{\alpha_{k^+}} e^{-\alpha_{k^+}\lambda_{\alpha_{k^+}}}.$$

One way of (approximately) sampling from this distribution is through the Metropolis-Hastings scheme.

We specify a proposal $q(\alpha_{k^+}^{new} \mid \alpha_{k^+}^{old})$ corrected by the acceptance probability $\alpha(\alpha_{k^+}^{new} \mid \alpha_{k^+}^{old}) := \min\left\{1,\ r(\alpha_{k^+}^{new} \mid \alpha_{k^+}^{old})\right\}$, where

$$r(\alpha_{k^+}^{new} \mid \alpha_{k^+}^{old}) := \frac{q(\alpha_{k^+}^{old} \mid \alpha_{k^+}^{new}) p(\alpha_{k^+}^{new} \mid y, else)}{q(\alpha_{k^+}^{new} \mid \alpha_{k^+}^{old}) p(\alpha_{k^+}^{old} \mid y, else)}.$$

Our proposal is a random-walk on $\log \alpha_{k^+}$, i.e., $\log \alpha_{k^+}^{new} \sim N(\log \alpha_{k^+}^{old}, V^+)$ for some fixed $V^+ > 0$. which implies a LogNormal proposal density on the original scale with $q(\alpha_{k^+}^{new} \mid \alpha_{k^+}^{old}) = N(\alpha_{k^+}^{new}; \alpha_{k^+}^{old}, V^+) \times 1/\alpha_{k^+}^{new}$.

It is straightforward to verify that

$$\begin{aligned}\log r(\alpha_{k^+}^{new} \mid \alpha_{k^+}^{old}) = (\log \alpha_{k^+}^{old} - \log \alpha_{k^+}^{new}) - G\left[\log \Gamma(\alpha_{k^+}^{new}) - \log \Gamma(\alpha_{k^+}^{old})\right] + \\ + (\alpha_{k^+}^{new} - \alpha_{k^+}^{old}) \left[G \log \beta_{k^+} - \sum_{g=1}^{G} \log k_g^+ - \lambda_{\alpha_{k^+}}\right].\end{aligned}$$

Updating α_{k^-}

Updating α_{k^-} is entirely analogous to updating α_{k^+}.

D.1.1 Sampling from truncated distributions within MCMC

In this section we describe the Gibbs sampler augmentation scheme to asymptotically sample from the truncated inverse gamma and truncated normal distributions that appear in appendix D.1. Although sampling algorithms for truncated distributions can be easy derived, sometimes even by the inverse c.d.f. method, the resulting algorithm can often be numerically unstable (take the truncated Gaussian distribution for example). In such cases, it could be advantageous to use an approximate sampler if it is more robust to computational errors. In this regard, we follow the directions on Damien and Walker (2001).

The univariate normal sampling scheme can be seen as a particular instance of the algorithm for multivariate normals or as an extension of the sampling scheme for univariate standard normals that are both described in Damien and Walker (2001). The algorithm to sample from truncated inverse gamma is very similar to the one that samples from the truncated Gamma. We describe both sampling schemes here solely for the purpose of completeness.

D.1.1.1 Truncated normal

Suppose a truncated Gaussian distribution for the random variable X: $X \sim N(\mu, \sigma^2)\mathbb{1}(a < X < b)$, i.e., $f_X(x) \propto \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}\mathbb{1}(a < x < b)$, where we could have $a = -\infty$ or $b = +\infty$ to represent unilateral truncation. We define the auxiliary variable Y through the joint density

$$f_{X,Y}(x, y) \propto \mathbb{1}(0 < y < e^{-\frac{(x-\mu)^2}{2\sigma^2}})\mathbb{1}(a < x < b)$$

so that the implied marginal for X matches the original $N(\mu, \sigma^2)\mathbb{1}(a < X < b)$. The full conditional distributions are

$$(Y \mid X = x) \sim Unif\left(0,\ \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}\right), \tag{D.1}$$

$$(X \mid Y = y) \sim Unif\left(\max(a,\ \mu - \sqrt{-2\sigma^2 \log y}),\ \ \min(b,\ \mu + \sqrt{-2\sigma^2 \log y})\right). \tag{D.2}$$

Within the MCMC scheme described in section D.1, we include sampling from the auxiliary variables Y_{μ_g} and Y_{α_t} corresponding to the full conditional distributions of μ_g, and α_t respectively. The auxiliary variables are sampled from (D.1) while the original variables are sampled from (D.2), with the appropriate values of μ, σ^2, a and b.

D.1.1.2 Truncated inverse gamma

Suppose $X \sim InvGamma(\alpha, \beta)\mathbb{1}(a < x < b)$, i.e., $f_X(x) \propto x^{-\alpha-1}e^{-\frac{x}{\beta}}$ $\mathbb{1}(a < x < b)$. We define the joint density of X and Y:

$$f_{X,Y}(x, y) \propto x^{-\alpha-1}\mathbb{1}(0 < y < e^{-\frac{x}{\beta}})\mathbb{1}(a < x < b)$$

so that the implied marginal for X matches the original $InvGamma(\alpha, \beta)\mathbb{1}(a < X < b)$. The full conditional distributions are

$$(Y \mid X = x) \sim Unif(0, e^{-\frac{x}{\beta}}), \tag{D.3}$$

$$f_{X|Y}(x \mid y) \propto x^{-\alpha-1}\mathbb{1}(M(y) < x < b), \tag{D.4}$$

where $M(y) := \max\left(a,\ -\frac{\beta}{\log y}\right)$. The inverse c.d.f. method provides an efficient way to sample from (D.4): sample $U \sim Unif(0, 1)$ then evaluate the transformed variable

$$\frac{M(y)}{[U(\{M(y)/b\}^{\alpha} - 1) + 1]^{\frac{1}{\alpha}}},$$

which will be distributed as (D.4).

Within the MCMC scheme described in section D.1, we include sampling from the auxiliary variables $Y_{k_g^+}$, $Y_{k_g^-}$ and $Y_{\sigma_g^2}$ corresponding to the full conditional distributions of k_g^+, k_g^- and σ_g^2 respectively. The auxiliary variables are sampled from (D.4) while the original variables are sampled from (D.3), with the appropriate values of α and β.

D.2 Full Conditionals for Matching Cell Line and Patients Model

This appendix describes steps of the Gibbs sampler algorithm used to carry out posterior inference.

Define $S_j^{x,k} := \{i : \delta_i^{x,k} = j\}$ with x being either c or p. In the remainder of this appendix section, we will denote by s_j the single element in the set $S_j^{c,k}$ (it could even be $s_j = \emptyset$), omitting the superscripts for simplicity.

The full posterior (up to a normalizing constant depending solely on the data $\boldsymbol{d}$) can be factorized as

$$
\begin{aligned}
p(\boldsymbol{\Psi} \mid \boldsymbol{d}) \propto & \left[\prod_{g=1}^{G} p(\sigma_g^{-2}) p(\sigma_{1g}^{-2}) p(\sigma_{2g}^{-2})\right] p(\boldsymbol{w} \mid \pi_0, \alpha_0) \left[\prod_{k=1}^{K_{\boldsymbol{w}}} p(\boldsymbol{\delta}^{p,k}) p(\boldsymbol{\delta}^{c,k} \mid \boldsymbol{\delta}^{p,k})\right] \times \\
& \times \prod_{k=1}^{K_{\boldsymbol{w}}} \prod_{g:w_g=k} \prod_{j=1}^{J_k} p(\theta_{jg}^* \mid \mu_{0g}, \sigma_{0g}^2) \times \left[\prod_{g=1}^{G} p(\mu_{0g}) p(\mu_{1g}) p(\mu_{2g})\right] \times \\
& \times \prod_{k=1}^{K_{\boldsymbol{w}}} \prod_{g:w_g=k} \prod_{j=1}^{J_k} \left[p(d_{ig}^c \mid \theta_{jg}^*, \sigma_g^2)^{\mathbb{1}(S_j^{c,k}=\{i\}\neq\emptyset)} \prod_{i:\delta_i^{p,k}=j} p(d_{ig}^p \mid \theta_{jg}^*, \sigma_g^2)\right] \times \\
& \times \prod_{k=1}^{K_{\boldsymbol{w}}} \prod_{g:w_g=k} \left[\prod_{i:\delta_i^{c,k}=0} p(d_{ig}^c \mid \mu_{1g}, \sigma_g^2, \sigma_{1g}^2) \prod_{i:\delta_i^{p,k}=0} p(d_{ig}^p \mid \mu_{1g}, \sigma_g^2, \sigma_{1g}^2)\right] \times \\
& \times \prod_{g:w_g=0} \left[\prod_{i=1}^{N^p} p(d_{ig}^p \mid \mu_{2g}, \sigma_g^2, \sigma_{2g}^2) \prod_{i=1}^{N^c} p(d_{ig}^c \mid \mu_{2g}, \sigma_g^2, \sigma_{2g}^2)\right]. \qquad \text{(D.5)}
\end{aligned}
$$

Updating θ_{jg}^*:

$$
p(\theta_{jg}^* \mid \boldsymbol{d}, \boldsymbol{\Psi}_{-\theta_{jg}^*}) \propto N(\theta_{jg}^*;\ \mu_{0g}, \sigma_{0g}^2) \prod_{i \in S_j^{c,w_g}} N(d_{ig}^c;\ \theta_{jg}^*, \sigma_g^2) \prod_{i \in S_j^{c,w_g}} N(d_{ig}^p;\ \theta_{jg}^*, \sigma_g^2)
$$

$$
\begin{aligned}
(\theta_{jg}^* \mid \boldsymbol{d}, \boldsymbol{\Psi}_{-\theta_{jg}^*}) \sim N \Bigg(& \frac{(\sum_{i \in S_j^{c,w_g}} d_{ig}^c + \sum_{i \in S_j^{p,w_g}} d_{ig}^p)\sigma_g^{-2} + \mu_{0g}\sigma_{0g}^{-2}}{(|S_j^{c,w_g}| + |S_j^{p,w_g}|)\sigma_g^{-2} + \sigma_{0g}^{-2}}, \\
& \frac{1}{(|S_j^{c,w_g}| + |S_j^{p,w_g}|)\sigma_g^{-2} + \sigma_{0g}^{-2}} \Bigg).
\end{aligned}
$$

Updating $\boldsymbol{\delta}^{p,k}$:

We follow Bush and MacEachern (1996) and sample the cluster membership indicators $\delta^{p,k}$ within a Gibbs block marginalizing $\boldsymbol{\theta}^*$ out, i.e., by sampling $p(\delta_i^{p,k} = j \mid \boldsymbol{d}, \boldsymbol{\Psi}_{-(\delta_i^{p,k},\boldsymbol{\theta}^*)})$ for $i = 1, \dots, N^p$. These updates together with the previous one where we sampled $\theta_{jg}^* \sim p(\theta_{jg}^* \mid \boldsymbol{\Psi}_{-\theta_{jg}^*})$ for all j and g, asymptotically provides a blocked joint sample from $p(\boldsymbol{\theta}^*, \boldsymbol{\delta}^{p,k} \mid \boldsymbol{d}, \boldsymbol{\Psi}_{-(\boldsymbol{\delta}^{p,k},\boldsymbol{\theta}^*)})$.

Denote by $A_{p,k}^-$ the number of active patient samples within protein sample k excluding patient i and define $S_j^{p,k-} := S_j^{p,k} \setminus \{i\}$. Then $\boldsymbol{\delta}^{p,k} \sim ZEPU(\alpha_{p,k}, \pi_{pk})$ implies

$$P(\delta_i^{p,k} = j \mid \boldsymbol{\delta}_{-i}^{p,k}) = \begin{cases} \pi_{pk}, \quad j = 0 \\ (1 - \pi_{pk}) \frac{|S_j^{p,k-}|}{\alpha_{pk} + A_{p,k}^-}, \quad j = 1, \dots, J_k^- \\ (1 - \pi_{pk}) \frac{\alpha_{p,k}}{\alpha_{pk} + A_{p,k}^-}, \quad j = J_k^- + 1. \end{cases} \tag{D.6}$$

Using equation (D.6), we obtain

$$\begin{aligned} P(\delta_i^{p,k} = j \mid \boldsymbol{\delta}^{c,k}, \boldsymbol{\delta}_{-i}^{p,k}) &\propto p(\boldsymbol{\delta}^{c,k} \mid \boldsymbol{\delta}^{p,k}) P(\delta_i^{p,k} = j \mid \boldsymbol{\delta}_{-i}^{p,k}) \\ &\propto P(\delta_i^{p,k} = j \mid \boldsymbol{\delta}_{-i}^{p,k}) \sum_{n=0}^{\min\{J_k, N^c\}} \binom{J_k}{n} \binom{N^c}{n} n! \end{aligned} \tag{D.7}$$

analytically. Notice that J_k varies with $\delta_i^{p,k}$ so the summation term cannot by omitted from (D.7).

After marginalizing θ_{jg}^* out from $p(d_{ig}^p \mid \boldsymbol{d}_{-ig}^p, d_{s_j g}^c, \delta_i^{p,k} = j, \boldsymbol{\Psi}_{-\delta_i^{p,k}})$, we obtain

$$
\begin{aligned}
&p(d_{ig}^{p} \mid \boldsymbol{d}_{-ig}^{p}, d_{s_j g}^{c}, \delta_i^{p,k} = j, \boldsymbol{\Psi}_{-(\delta_i^{p,k}, \boldsymbol{\theta}^*)}) \\
&\quad = \sqrt{\frac{(|S_j^{p,k-}|\sigma_g^{-2} + \sigma_{0g}^{-2})(|S_j^{p,k-}|\sigma_g^{-2} + \sigma_{0g}^{-2} + \sigma_g^{-2})}{(2\pi\sigma_g^2)}} \times \\
&\quad \times \exp\left\{ -\frac{1}{2} \left[d_{ig}^{p\;2} \sigma_g^{-2} + (|S_j^{p,k-}|\sigma_g^{-2} + \sigma_{0g}^{-2}) \times \right.\right. \\
&\qquad\qquad \left.\left. \times \left(\sigma_g^{-2} \sum_{\ell \in S_j^{p,k-}} d_{\ell g}^{p} + d_{s_j g}^{c} \mathbb{1}(S_j^{c,k} \neq \emptyset) \sigma_g^{-2} + \mu_{0g}\sigma_{0g}^{-2} \right)^2 \right] \right\} \times \\
&\quad \times \exp\left\{ \frac{1}{2} (\sigma_g^{-2} + \sigma_{0g}^{-2} + |S_j^{p,k-}|\sigma_g^{-2})^{-1} \times \right. \\
&\qquad\qquad \left. \times \left[d_{ig}^{p} \sigma_g^{-2} + \left(\sum_{\ell \in S_j^{p,k-}} d_{\ell g}^{p} + d_{s_j g}^{c} \mathbb{1}(S_j^{c,k} \neq \emptyset) \right) \sigma_g^{-2} + \mu_{0g}\sigma_{0,g}^{-2} \right] \right\}.
\end{aligned}
$$

Using equation (D.7), we get

$$
\begin{aligned}
&P(\delta_i^{p,k} = j \mid \boldsymbol{\Psi}_{-(\delta_i^{p,k}, \boldsymbol{\theta}^*)}) \propto \\
&\quad \propto \begin{cases}
\left[\prod_{g: w_g = k} N(d_{ig}^{p} \mid \mu_{1g}, \sigma_g^2 + \sigma_{1g}^2)\right] \pi_{pk} \sum_{n=0}^{J_k^- + 1} \frac{1}{(N^c - n)!} b_j, & j = 0 \\[2ex]
\prod_{g: w_g = k} p(d_{ig}^{p} \mid \boldsymbol{d}_{-ig}^{p}, d_{s_j g}^{c}, \delta_i^{p,k} = j, \boldsymbol{\Psi}_{-(\delta_i^{p,k}, \boldsymbol{\theta}^*)}) \times & \\
\qquad\qquad \times (1 - \pi_{pk}) \frac{|S_j^{p,k} - \{i\}|}{\alpha_{pk} + A_{p,k}^-} \sum_{n=0}^{J_k^- + 1} \frac{1}{(N^c - n)!} b_j, & j = 1, \ldots, J_k^- \\[2ex]
\prod_{g: w_g = k} N(d_{ig}^{p} \mid \mu_{0g}, \sigma_g^2 + \sigma_{0g}^2) \times & \\
\qquad\qquad \times (1 - \pi_{pk}) \frac{\alpha_{p,k}}{\alpha_{pk} + A_{p,k}^-} \sum_{n=0}^{J_k^- + 2} \frac{1}{(N^c - n)!} b_j, & j = J_k^- + 1,
\end{cases}
\end{aligned}
$$

where J_k^- is the number of active clusters of patients within protein group k after removing patient i and $b_j := \mathbb{1}(|S_j^{c,k}| = 1)\mathbb{1}(|S_j^{p,k}| > 0) + \mathbb{1}(|S_j^{c,k}| = 0)$ is a binary variable that enforces non-empty active clusters of cell lines to contain at least one patient sample as well.

Updating $\boldsymbol{\delta}_i^{c,k}$:

$$p(\delta_i^{c,k} = j \mid \boldsymbol{d}, \boldsymbol{\Psi}_{-\delta_i^{c,k}}) \propto$$

$$\propto \begin{cases} \prod_{g:w_g=k} N(d_{ig}^c;\ \theta_{jg}^*, \sigma_{1g}^2), & j > 0,\ S_j^{p,k} \neq \emptyset,\ S_j^{c,k} = \emptyset \\ \prod_{g:w_g=k} N(d_{ig}^c;\ \mu_1, \sigma_g^2 + \sigma_{1g}^2), & j = 0 \\ 0, & \text{otherwise.} \end{cases}$$

We also define a Metropolis-Hastings algorithm to sample from $\boldsymbol{\delta}^{c,k}$ in a way that hopefully produces Markov Chains with better mixing properties. Here we omit the upper indexes c, k from $\boldsymbol{\delta}^{c,k}$ for clarity of exposition.

Recall that the Metropolis-Hastings algorithm produces a new sample $\boldsymbol{\delta}^{t+1}$ from $\boldsymbol{\delta}^t$ according to an auxiliary transition probability $q(\boldsymbol{\delta}^{t+1} \mid \boldsymbol{\delta}^t)$ that is irreducible and aperiodic. Then $\boldsymbol{\delta}^{t+1}$ is accepted with probability $\alpha(\boldsymbol{\delta}^{t+1} \mid \boldsymbol{\delta}^t) := \max\{1, r(\boldsymbol{\delta}^{t+1} \mid \boldsymbol{\delta}^t)\}$ where $r(\boldsymbol{\delta}^{t+1} \mid \boldsymbol{\delta}^t) := \frac{q(\boldsymbol{\delta}^t | \boldsymbol{\delta}^{t+1}) p(\boldsymbol{\delta}^{t+1})}{q(\boldsymbol{\delta}^{t+1} | \boldsymbol{\delta}^t) p(\boldsymbol{\delta}^t)}$.

We define two types of transitions and at each iteration we uniformly chose one of them at random.

Type I: We take an active cell line and switch it with one of the inactive cell lines. Under such proposal, $q(\boldsymbol{\delta}^{t+1} \mid \boldsymbol{\delta}^t) = \frac{1}{|S_0^{c,k}|(N^c - |S_0^{c,k}|)}$. Under this proposal, the acceptance ratio reduces to

$$r(\boldsymbol{\delta}^{c,k}(t+1) \mid \boldsymbol{\delta}^{c,k}(t)) = \prod_{g:w_g=k} \frac{N(d^c_{i_0(t+1)}; \mu_1, \sigma^2_g + \sigma^2_{1g}) N(d^c_{ig}; \theta^*_{jg}, \sigma^2_g)\big|_{j=\delta^{c,k}_{i_1(t+1)}}}{N(d^c_{i_0(t)}; \mu_1, \sigma^2_g + \sigma^2_{1g}) N(d^c_{ig}; \theta^*_{\ell g}, \sigma^2_g)\big|_{\ell=\delta^{c,k}_{i_1(t)}}},$$

where $i_1(x)$ and $i_0(x)$ respectively denote the active and inactive cell lines selected by the proposal at time x (before switching, when $x = t$; and after switching, when $x = t+1$).

Type II: Randomly pick an active cluster j. If $|S^{c,k}_j(t)| = 0$ (no active cell line in cluster j), assign an inactive cell line to cluster j uniformly at random. On the other hand, if $|S^{c,k}_j(t)| = 1$ we reassign the only active cell line $i \in S^{c,k}_j(t)$ from cluster j to the group of inactive cell lines by making $\delta^{c,k}_i(t+1) = 0$. Under such proposal, we have $q(\boldsymbol{\delta}^{t+1} \mid \boldsymbol{\delta}^t) = \frac{1}{J_k |S^{c,k}_0(t)|} \mathbb{1}(|S^{c,k}_j| = 0) + \frac{1}{J_k} \mathbb{1}(|S^{c,k}_j| = 1)$. Under this proposal,

$$r(\boldsymbol{\delta}^{c,k}(t+1) \mid \boldsymbol{\delta}^{c,k}(t)) = \begin{cases} \frac{N(d^c_{ig};\ \theta^*_{jg}, \sigma^2_g) |S^{c,k}_0(t)|}{N(d^c_{ig};\ \mu_1, \sigma^2_{0g} \sigma^2_g)}\Big|_{i \in S^{c,k}_j(t+1)}, & S^{c,k}_j(t) = \emptyset \\ \frac{N(d^c_{ig};\ \mu_1, \sigma^2_{0g} \sigma^2_g)}{N(d^c_{ig};\ \theta^*_{jg}, \sigma^2_g) |S^{c,k}_0(t+1)|}\Big|_{i \in S^{c,k}_j(t)}, & S^{c,k}_j(t) = \{i\} \neq \emptyset. \end{cases}$$