POISSON MASH MODEL ALLOWING FOR UNWANTED VARIATION

1. Model setup. Suppose there are $j = 1, \ldots, J$ genes and $i = 1, \ldots, N$ cells. The observed single cell count matrix Y is $J \times N$, with its (j, i) element Y_{ji} denoting the count of gene j in cell i.

We assume that the N cells come from $r = 1, \ldots, R$ conditions, with n_r cells (indexed by $\mathcal{S}_r \subset \{1, \ldots, N\}$) coming from condition r. Further assume that the R conditions belong to $m = 1, \ldots, M$ subgroups ($1 \leq M < R$). For example, subgroups can be different cell types and conditions can be combinations of treatments and cell types. We are interested in comparing gene expression levels across conditions $r \in \mathcal{T}_m \subset \{1, \ldots, R\}$ within each subgroup m, e.g., comparing gene expression levels corresponding to different treatments within each cell type. To do so, a first step is to collapse the single cell count matrix Y into a condition level count matrix X, which is a $J \times R$ matrix with its (j, r) element $X_{jr} = \sum_{i \in \mathcal{S}_r} Y_{ji}$.

Let s_i denote the size factor of cell i, which can be calculated by taking the sum (or equivalently, mean) of counts over all genes in cell i, or using other more robust methods [1, 3]. Let $s_r = \sum_{i \in \mathcal{S}_r} s_i$ denote the size factor of condition r.

We assume the following model for the matrix of counts X collapsed over conditions:

$$X_{jr} \sim Pois(s_r \lambda_{jr}), \tag{1}$$

where λ_{jr} denotes the gene-specific, condition-specific intensity parameter. For each gene j, we are interested in comparing λ_{jr} across conditions $r \in \mathcal{T}_m$ within each subgroup m.

To i) model possible correlations in λ_{jr} across r, ii) allow over-dispersion in the count data, and iii) account for unwanted variation, for condition r which belongs to subgroup $m(r)$, we place the following prior on $\log(\lambda_{jr})$:

$$\log(\lambda_{jr}) = \mu_{jm(r)} + \beta_{jr} + \eta_{jr} + \sum_{d=1}^{D} \rho_{rd} f_{jd}, \tag{2}$$

$$\boldsymbol{\beta}_j \sim \sum_{k,l} \pi_{kl} N(\mathbf{0}, w_l U_k) \quad \text{where} \quad \sum_{k,l} \pi_{kl} = 1, \tag{3}$$

$$\boldsymbol{\eta}_j \sim N(\mathbf{0}, \psi_j^2 I_R). \tag{4}$$

In (2), $\mu_{jm(r)}$ represents the gene-specific, subgroup-specific underlying mean of $\log(\lambda_{jr})$, and the term $\sum_{d=1}^{D} \rho_{rd} f_{jd}$ represents the bias caused by unwanted variation, with F being a $J \times D$ matrix of unobserved factors and $\boldsymbol{\rho}$ being a $D \times R$ matrix of corresponding effects. Here we adopt a similar framework as in [2] to account for unwanted variation.

In (3), $\boldsymbol{\beta}_j$ is an $R \times 1$ vector modeling the gene-specific, condition-specific effects which is our *quantity of interest*, and has a mixture multivariate Gaussian prior involving a grid of scaling factors w_l ($l = 1, \ldots, L$) and a set of covariance matrices U_k ($k = 1, \ldots, K$) that include both canonical and data-driven ones. $\boldsymbol{\pi}$ is a $KL \times 1$ vector of weights for different prior covariances.

In (4), $\boldsymbol{\eta}_j$ is an $R \times 1$ vector of Gaussian random effect with a gene-specific prior covariance $\psi_j^2 I_R$, which is introduced to allow for possible over-dispersion of single cell data.

2. Model fitting with variational approximation. In (1) to (4), only X and s_r ($r = 1, \ldots, R$) are observed, and the grid of scaling factors w_l ($l = 1, \ldots, L$) can be chosen in a data-adaptive manner. The remaining quantities need to be estimated.

To fit the model described in Section 1, we first get an estimate $\hat{F}$ of F by running factor analysis on the single cell count matrix Y while accounting for condition-specific effects under a GLM model. This step can be performed using the R package "glmpca" [4].

With the plug-in estimate $\hat{F}$ for F, we now describe how to estimate the remaining quantities. Let $\boldsymbol{\Theta} := \left(\boldsymbol{\mu}, \boldsymbol{\rho}, \boldsymbol{\pi}, \boldsymbol{U}, \boldsymbol{\psi}^2\right)$ indicate the model parameters to be estimated from the data, where $\boldsymbol{\mu}$ is the $J \times M$ matrix of gene-specific, subgroup-specific mean parameters, $\boldsymbol{\psi}^2$ is the $J \times 1$ vector of gene-specific dispersion parameters, $\boldsymbol{\rho}$ is the $D \times R$ matrix of effects for unwanted variation, $\boldsymbol{\pi}$ is the $KL \times 1$ vector of prior weights, and $\boldsymbol{U}$ is the collection of prior covariance matrices. The data likelihood can be written as

$$L(\boldsymbol{\Theta};\, X, \boldsymbol{s}, \hat{F}) = \prod_j \left[\sum_{k,\,l} \pi_{kl}\, p\left(\boldsymbol{X}_j \mid \boldsymbol{\mu}_j, \boldsymbol{\rho}' \hat{\boldsymbol{f}}_j, w_l U_k, \psi_j^2\right) \right] \tag{5}$$

$$= \prod_j \left[\sum_{k,\,l} \pi_{kl} \int p\left(\boldsymbol{X}_j \mid \boldsymbol{\mu}_j, \boldsymbol{\beta}_j, \boldsymbol{\eta}_j, \boldsymbol{\rho}' \hat{\boldsymbol{f}}_j\right) p\left(\boldsymbol{\beta}_j \mid w_l U_k\right) p\left(\boldsymbol{\eta}_j \mid \psi_j^2\right) d\boldsymbol{\beta}_j\, d\boldsymbol{\eta}_j \right]. \tag{6}$$

As is commonly done when fitting mixture models, we introduce a $KL \times 1$ vector of latent indicator $\boldsymbol{z}_j$ for each gene j to facilitate model fitting, such that $\sum_{k,l} z_{jkl} = 1$ and

$$\boldsymbol{\beta}_j \mid (z_{jkl} = 1) \;\sim\; MVN(\boldsymbol{0},\, w_l U_k). \tag{7}$$

Let $\boldsymbol{Z}$ denote the collection of $\boldsymbol{z}_j$ for all j. With the introduction of latent indicator variables $\boldsymbol{z}_j$, the complete data log-likelihood is

$$\log L(\boldsymbol{\Theta};\, X, \boldsymbol{s}, \hat{F}, \boldsymbol{Z}) = \sum_j \sum_{k,\,l} z_{jkl} \left[\log \pi_{kl} + \log p\left(\boldsymbol{X}_j \mid \boldsymbol{\mu}_j, \boldsymbol{\rho}' \hat{\boldsymbol{f}}_j, w_l U_k, \psi_j^2\right) \right]. \tag{8}$$

Let $\boldsymbol{\theta}_j := \boldsymbol{\beta}_j + \boldsymbol{\eta}_j$ for each gene j, and $\boldsymbol{\theta}$ denote the collection of $\boldsymbol{\theta}_j$ for all j. We are interested in the joint posterior of $(\boldsymbol{\theta}, \boldsymbol{Z})$ which does not have a closed-form:

$$p\left(\boldsymbol{\theta}, \boldsymbol{Z} \mid X, \boldsymbol{\mu}, \boldsymbol{\rho}, \boldsymbol{\pi}, \boldsymbol{U}, \boldsymbol{\psi}^2\right) \propto p\left(X \mid \boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{\rho}\right)\; p\left(\boldsymbol{\theta}, \boldsymbol{Z} \mid \boldsymbol{\pi}, \boldsymbol{U}, \boldsymbol{\psi}^2\right) \tag{9}$$

$$\propto \prod_j \left\{ p\left(\boldsymbol{X}_j \mid \boldsymbol{\theta}_j, \boldsymbol{\mu}_j, \boldsymbol{\rho}' \hat{\boldsymbol{f}}_j\right) \prod_{k,\,l} \left[\pi_{kl}\, N\left(\boldsymbol{\theta}_j \mid \boldsymbol{0}, w_l U_k + \psi_j^2 I_R\right)\right]^{z_{jkl}} \right\}.$$

Therefore, we approximate the true joint posterior $p\left(\boldsymbol{\theta}_j, \boldsymbol{z}_j \mid \boldsymbol{X}_j, \boldsymbol{\mu}_j, \psi_j^2, \boldsymbol{\rho}, \boldsymbol{\pi}, \boldsymbol{U}\right)$ with $q(\boldsymbol{\theta}_j, \boldsymbol{z}_j)$, which is restricted to be a mixture of multivariate Gaussian distributions. That is, for each j,

$$q(\boldsymbol{\theta}_j, \boldsymbol{z}_j) = q(\boldsymbol{\theta}_j \mid \boldsymbol{z}_j)\, q(\boldsymbol{z}_j) = \prod_{k,\,l} \left[\zeta_{jkl}\, N(\boldsymbol{\theta}_j \mid \boldsymbol{\gamma}_{jkl}, V_{jkl})\right]^{z_{jkl}}, \tag{10}$$

where $\boldsymbol{\zeta}_j$ is a $KL \times 1$ vector of posterior weights for $\boldsymbol{z}_j$.

We estimate the model parameters $\boldsymbol{\mu}, \boldsymbol{\rho}, \boldsymbol{\pi}, \boldsymbol{U}, \boldsymbol{\psi}^2$ and the variational approximation parameters $\{\boldsymbol{\zeta}_j\}_j$, $\{\boldsymbol{\gamma}_{jkl}\}_{j,k,l}$, $\{V_{jkl}\}_{j,k,l}$ by maximizing the "overall" ELBO defined in (11):

$$F_{overall}\left(q(\boldsymbol{\theta}, \boldsymbol{Z}), \boldsymbol{\mu}, \boldsymbol{\rho}, \boldsymbol{\pi}, \boldsymbol{U}, \boldsymbol{\psi}^2; X, \boldsymbol{s}\right) \tag{11}$$

$$:= \log p\left(X \mid \boldsymbol{\mu}, \boldsymbol{\rho}, \boldsymbol{\pi}, \boldsymbol{U}, \boldsymbol{\psi}^2\right) - D_{KL}\left(q(\boldsymbol{\theta}, \boldsymbol{Z}) \parallel p\left(\boldsymbol{\theta}, \boldsymbol{Z} \mid X, \boldsymbol{\mu}, \boldsymbol{\rho}, \boldsymbol{\pi}, \boldsymbol{U}, \boldsymbol{\psi}^2\right)\right) \tag{12}$$

$$= \mathbb{E}_q\left[\log p(X, \boldsymbol{\theta}, \boldsymbol{Z} \mid \boldsymbol{\mu}, \boldsymbol{\rho}, \boldsymbol{\pi}, \boldsymbol{U}, \boldsymbol{\psi}^2)\right] - \mathbb{E}_q\left[\log q(\boldsymbol{\theta}, \boldsymbol{Z})\right] \tag{13}$$

$$= \mathbb{E}_q\left[\log p(X \mid \boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{\rho}) + \log p(\boldsymbol{\theta}, \boldsymbol{Z} \mid \boldsymbol{\pi}, \boldsymbol{\psi}^2, \boldsymbol{U})\right] - \mathbb{E}_q\left[\log q(\boldsymbol{\theta}, \boldsymbol{Z})\right] \tag{14}$$

$$= \sum_j \sum_{k,\,l} \zeta_{jkl} \left[\log \pi_{kl} + F\left(\boldsymbol{\gamma}_{jkl}, V_{jkl}, \boldsymbol{\mu}_j, \boldsymbol{\rho}' \hat{\boldsymbol{f}}_j, w_l U_k, \psi_j^2; \boldsymbol{X}_j\right) - \log \zeta_{jkl} \right], \tag{15}$$

where $F\left(\boldsymbol{\gamma}_{jkl}, V_{jkl}, \boldsymbol{\mu}_j, \boldsymbol{\rho}'\hat{\boldsymbol{f}}_j, w_l U_k, \psi_j^2; \boldsymbol{X}_j\right)$ is the "local" ELBO defined in (16):

(16)

$$\begin{aligned}
&F\left(\boldsymbol{\gamma}_{jkl}, V_{jkl}, \boldsymbol{\mu}_j, \boldsymbol{\rho}'\hat{\boldsymbol{f}}_j, w_l U_k, \psi_j^2; \boldsymbol{X}_j\right) \\
:=\ & \mathbb{E}_{q_{jkl}}\left[\log p(\boldsymbol{X}_j \mid \boldsymbol{\mu}_j, \boldsymbol{\theta}_j, \boldsymbol{\rho}'\hat{\boldsymbol{f}}_j)\right] - D_{KL}\left(N(\boldsymbol{\theta}_j \mid \boldsymbol{\gamma}_{jkl}, V_{jkl}) \parallel N(\boldsymbol{\theta}_j \mid \mathbf{0}, w_l U_k + \psi_j^2 I_R)\right) \\
=\ & \sum_r \left\{ X_{jr}\left(\log s_r + \mu_{jm(r)} + \sum_{d=1}^{D} \rho_{rd}\hat{f}_{jd} + \gamma_{jklr}\right) - s_r \exp\left(\mu_{jm(r)} + \sum_{d=1}^{D} \rho_{rd}\hat{f}_{jd} + \gamma_{jklr} + \frac{1}{2}V_{jkl,rr}\right) - \log(X_{jr}!)\right\} \\
& - D_{KL}\left(N(\boldsymbol{\theta}_j \mid \boldsymbol{\gamma}_{jkl}, V_{jkl}) \parallel N(\boldsymbol{\theta}_j \mid \mathbf{0}, w_l U_k + \psi_j^2 I_R)\right).
\end{aligned}$$

REFERENCES

[1] J. H. Bullard, E. Purdom, K. D. Hansen, and S. Dudoit, *Evaluation of statistical methods for normalization and differential expression in mrna-seq experiments*, BMC bioinformatics, 11 (2010), pp. 1–13.

[2] D. Gerard and M. Stephens, *Empirical bayes shrinkage and false discovery rate estimation, allowing for unwanted variation*, Biostatistics, 21 (2020), pp. 15–32.

[3] A. T. Lun, K. Bach, and J. C. Marioni, *Pooling across cells to normalize single-cell rna sequencing data with many zero counts*, Genome biology, 17 (2016), p. 75.

[4] F. W. Townes, S. C. Hicks, M. J. Aryee, and R. A. Irizarry, *Feature selection and dimension reduction for single-cell rna-seq based on a multinomial model*, Genome biology, 20 (2019), pp. 1–16.