A note on a prior for multinomial vector pairs

Version 12/17/16, by MAN

Set up: In modeling single-cell RNASeq data, we imagine that cells fall into one of $K > 1$ cell-type classes, and further that we have cells from two source conditions, giving n_1 and n_2 cells respectively. Let $z^1 = \{z^1_k\}$ be multinomial counts recording the number of cells of each cell type in the first condition, and $z^2 = \{z^2_k\}$ be similar counts for the second condition. One task may be to infer something about possible equalities in the underlying probability vectors, which here I denote $\phi = \{\phi_k\}$ and $\psi = \{\psi_k\}$. I.e.

$$z^1 \sim \text{Multinomial}_K(n_1, \phi)$$

and

$$z^2 \sim \text{Multinomial}_K(n_2, \psi)$$

Here I describe a prior $p(\phi, \psi)$ that is conjugate to multinomial sampling but that also enables downstream gene-specific inferences about differential distribution when certain cell types do not differ in their expression distributions. To provide this functionality, we need to work with partitions of the cell-type classes, say $\pi = \{b\}$ of $\{1, 2, \cdots, K\}$. Here b indicates a block in the partition π, and of course different blocks do not overlap and they cover the whole set of classes: $\cup_{b\in\pi} b = \{1, 2, \cdots, K\}$. We recall there is a large number of such partitions π, say constituting the set Π of cardinality Bell(K). We'll carry along an example involving $K = 7$ cell types, and one three-block partition, shown in Figure 1, taken from the set of 877 possible partitions of $\{1, 2, \cdots, 7\}$.

Mixture prior: For our purposes, the prior will have a *spike-slab* structure that mixes over distinct patterns of equality of π-associated accummulated probabilities:

$$p(\phi, \psi) = \sum_{\pi\in\Pi} p(c_\pi)\, p(\phi, \psi | c_\pi)$$

where

$$c_\pi = \{(\phi, \psi) : \Phi_b = \Psi_b \,\forall b \in \pi\}$$

where $\Phi_b = \sum_{k\in b} \phi_k$ and $\Psi_b = \sum_{k\in b} \psi_k$. Figure 2 shows one pair (ϕ, ψ) within c_π, where π is from Figure 1.

Indeed, c_π is precisely the structure needed to address differential distribution DD_g at a given gene g from expression data. Briefly, if the class probabilities satisfy c_π, and if the class-specific expression distributions are constant among classes k within each $b \in \pi$, then

Figure 1: Example partition of 7 cell types, k, into 3 blocks, b

even if there are differential proportions $\phi_k \neq \psi_k$ on such classes, this will not lead to differential distribution. More specifically, we say cell class specific means $\mu = \{\mu_k\}$ satisfy expression pattern d_π if

$$\mu_j = \mu_k \quad \forall j, k \in b, \quad \forall b \in \pi$$

EBSeq computed over the identified cell classes computes the posterior probability of d_π given expression data, for all π and at each gene. With ED_g denoting *equivalent distribution*, which is the complement of DD_g,

Conjecture: At a given gene, equivalent distribution is

$$\text{ED}_g = \bigcup_{\pi \in \Pi} [c_\pi \cap d_\pi] \,.$$

The central point is that even allowing for cell-type probabilities to change between source conditions, there is no differential distribution at a given gene if the relevant cell types are not differentially expressed at that gene. Upon setting up a prior $p(\phi, \psi)$ that can mix over structures c_π, and by combining cell-type counts z^1 and z^2 with expression data x_g at a gene, we may compute $P(\text{ED}_g|\text{data})$ at each gene:

$$P(\text{ED}_g|\text{data}) = \sum_{\pi \in \Pi} P(c_\pi|z^1, z^2)\, P(d_\pi|x_g).$$

Side issue: Each structure c_π is a closed subset of the product simplex holding all possible pairs of multinomial vectors (ϕ, ψ). As constructed, structures corresponding to different partitions are not disjoint, though we could refine the definition to encode disjointness. Consider the case where the partition π has a refinement π^*, say (i.e. unions of blocks in π^* lead to blocks in π). Then $c_{\pi^*} \subset c_\pi$ because of the aggregation constraint. Essentially the overlap represents a lower dimensional subset. Our approach is to define $p(\phi, \psi|c_\pi)$ in such a way that points in lower dimensional subsets receive zero probability mass in that mixture component, so that we don't need to strictly enforce disjointness. It's analogous to the spike-slab issue of a parameter θ that either equals 0 or does not; conditional on the latter case we might uses a Gaussian prior for θ, which gives zero mass but positive density to the subset $\theta = 0$. Another approach might be to force disjointness of c_π (giving up closed sets) by removing from c_π the c_{π^*} corresponding to partition refinements of π. [needs to be worked out]

Mixture components: For notation, we use ϕ_b for the vector of values ϕ_k for $k \in b$, and similarly for ψ_b. Analogously, Φ_π and Ψ_π are vectors of accumulated class probabilities ϕ_b and ψ_b for all $b \in \pi$, respectively.

Initially, the multitude of $p(c_\pi)$'s will be preset constants. To complete the prior specification $p(\phi, \psi)$, consider further scalers $\alpha_k > 0$ for each class k and $\beta_b > 0$ for each potential block b. (Extending the notational convention, α_b is the vector of α_k for $k \in b$, and β_π is the vector of β_b for $b \in \pi$.) For any block b consider conditional probabilities

$$\tilde{\phi}_b = \frac{\phi_b}{\Phi_b} \qquad \tilde{\psi}_b = \frac{\psi_b}{\Psi_b}$$

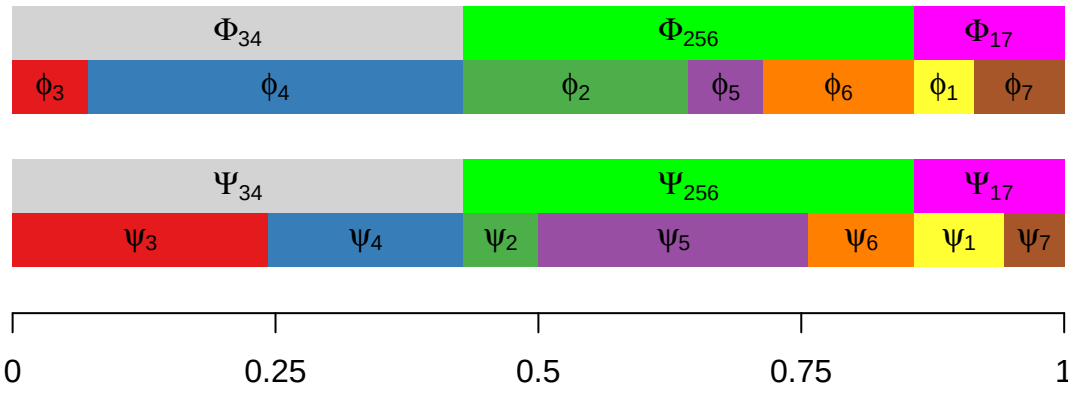


Figure 2: Multinomial probability vectors for two conditions arranged according to a fixed partition

which indicate the conditional probability of each class k given that the cell is of one of the types in b. Assume that conditional upon $(\phi, \psi) \in c_\pi$,

$$\Phi_\pi \sim \text{Dirichet}_{N(\pi)}[\beta_\pi]$$

where $N(\pi)$ is the number of blocks b in π, and further that accumulated probabilities are the same between the two source conditions: $\Phi_\pi = \Psi_\pi$. Finally, assume that for each $b \in \pi$,

$$\tilde{\phi}_b, \tilde{\psi}_b \sim_{\text{i.i.d.}} \text{Dirichlet}_{N(b)}[\alpha_b]$$

where $N(b)$ is the number of cell types in block b. In other words, if c_π is the active structure, then accumulated probability vectors Φ_π and Ψ_π are equal between the two source conditions, though the sub-block class-specific rates ϕ_k and ψ_k may differ, as would (re-normalized) independent Dirichlet-distributed vectors. Taken together,

$$p(\phi, \psi | c_\pi) = p(\Phi_\pi, \Psi_\pi | c_\pi) \prod_{b \in \pi} \left[p(\tilde{\phi}_b) p(\tilde{\psi}_b) \right]$$

with

$$p(\Phi_\pi, \Psi_\pi | c_\pi) = \frac{\Gamma(\sum_{b \in \pi} \beta_b)}{\prod_{b \in \pi} \Gamma(\beta_b)} \left[\prod_{b \in \pi} \Phi_b^{\beta_b - 1} \right] 1 \left[\Phi_\pi = \Psi_\pi \right]$$

and

$$p(\tilde{\phi}_b) = \frac{\Gamma(\sum_{k \in b} \alpha_k)}{\prod_{k \in b} \Gamma(\alpha_k)} \prod_{k \in b} \tilde{\phi}_k^{\alpha_k - 1}, \qquad p(\tilde{\psi}_b) = \frac{\Gamma(\sum_{k \in b} \alpha_k)}{\prod_{k \in b} \Gamma(\alpha_k)} \prod_{k \in b} \tilde{\psi}_k^{\alpha_k - 1}.$$

Distributional consequences: Using the Dirichlet-Multinomial conjugacy and the collapsing property of these distributions (cite), we get closed formulas for the predictive probability of cell-type counts z^1 and z^2. Fixing π, let $t_b^j = \sum_{k \in b} z_k^j$, for cell conditions $j = 1, 2$, record the total numbers of cells accumulated over all types in block b. And following our notation convention, t_π^j is the vector of these counts over $b \in \pi$. From the prior and model structure

$$p(z^1, z^2 | c_\pi) = p(z^1 | t_\pi^1)\, p(z^2 | t_\pi^2)\, p(t_\pi^1, t_\pi^2 | c_\pi).$$

Conditional independence of z^1 and z^2 given the block-level totals t^1_π and t^2_π on c_π reflects the possible differential class proportion structure within blocks but between cell conditions. For either cellular group $j = 1, 2$, we find, after some simplification, the following Dirichlet-Multinomial masses:

$$p(z^j|t^j_\pi) = \prod_{b\in\pi} \left\{ \left[\frac{\Gamma(t^j_b + 1)}{\prod_{k\in b}\Gamma(z^j_k + 1)} \right] \left[\frac{\Gamma(\sum_{k\in b}\alpha_k)}{\prod_{k\in b}\Gamma(\alpha_k)} \right] \left[\frac{\prod_{k\in b}\Gamma(\alpha_k + z^j_k)}{\Gamma(t^j_b + \sum_{k\in b}\alpha_k))} \right] \right\}$$

and

$$p(t^1_\pi, t^2_\pi|c_\pi) = \left[\frac{\Gamma(n_1 + 1)\Gamma(n_2 + 1)}{\prod_{b\in\pi}\Gamma(t^1_b + 1)\Gamma(t^2_b + 1)} \right] \left[\frac{\Gamma(\sum_{b\in\pi}\beta_b)}{\prod_{b\in\pi}\Gamma(\beta_b)} \right] \left[\frac{\prod_{b\in\pi}\Gamma(\beta_b + t^1_b + t^2_b)}{\Gamma(n_1 + n_2 + \sum_{b\in\pi}\beta_b)} \right].$$

Let's look at some special cases to dissect this result.

Check 1: If π has a single block equal to the entire set of cell types $\{1, 2, \cdots, K\}$, then $t^j_b = n_j$ for both $j = 1, 2$, and the second formula reduces, correctly, to $p(t^1_\pi, t^2_\pi|c_\pi) = 1$. Further,

$$p(z^j|t^j_\pi) = \left[\frac{\Gamma(n_j + 1)}{\Gamma(n_1 + \sum_{k=1}^K \alpha_k)} \right] \left[\frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)} \right] \left[\prod_{k=1}^K \frac{\Gamma(\alpha_k + z^j_k)}{\Gamma(z^j_k + 1)} \right]$$

which is the well-known Dirichlet-multinomial predictive distribution for counts z^j (cite). E.g, taking $\alpha_k = 1$ for all types k we get the uniform distribution

$$p(z^j|t^j_\pi) = \frac{\Gamma(n_j + 1)\Gamma(K)}{\Gamma(n_j + K)}.$$

Check 2: At the opposite extreme, π has one block b for each class k. Then $t^j_b = z^j_k$, and $p(z^j|t^j_\pi) = 1$, and further, assuming $\beta_b = \alpha_k$,

$$p(t^1_\pi, t^2_\pi|c_\pi) = \left[\frac{\Gamma(n_1 + 1)\Gamma(n_2 + 1)}{\prod_{k=1}^K \Gamma(z^1_k + 1)\Gamma(z^2_k + 1)} \right] \left[\frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)} \right] \left[\frac{\prod_{k=1}^K \Gamma(\alpha_k + z^1_k + z^2_k)}{\Gamma(n_1 + n_2 + \alpha_k)} \right].$$

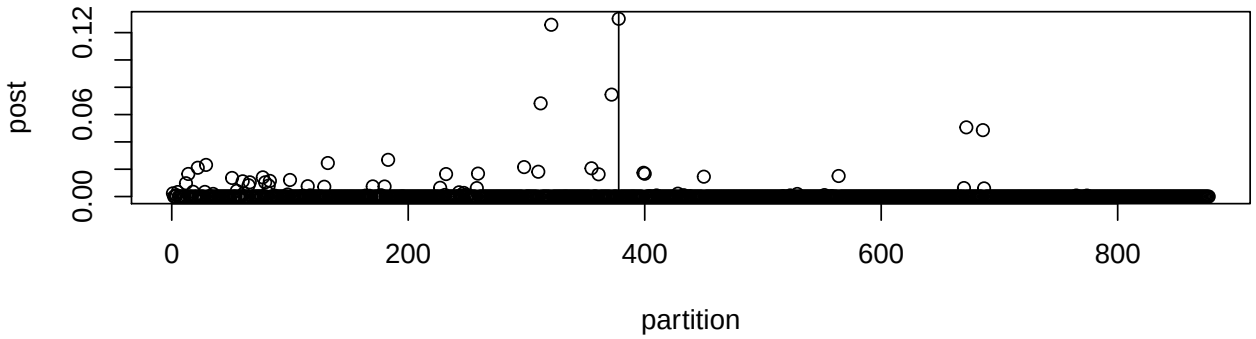
comment: cite a Bayes comparison of two multinomials

Regardless of the partition, log scale probabilities are readily evaluated given hyper-parameters $\{\alpha_k\}$ and $\{\beta_b\}$ and for cell-type counts z^1 and z^2.

Example

Here's an example using the probabilities ϕ and ψ from Figure 2; with $n_1 = n_2 = 500$, and $\alpha_k = 1$ for all k and $\beta_b = \sum_{k\in b}\alpha_k$.

```
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7]
## [1,]   38  183  110   28   77   27   37
## [2,]  128   93   46  117   46   46   24
```




```
## [1] 2 2 1 1 1 3 3
```

```
##        1   2  3
## [1,] 215 221 64
## [2,] 209 221 70
```