Supporting Information for "Integrating Sample Similarities into Latent Class Models: A Tree-Structured Shrinkage Approach" by Li, Park, Aziz, Liu, Price and Wu

Mengbing Li[1], Daniel Park[3], Maliha Aziz[3], Cindy M Liu[3], Lance Price[3], and Zhenke Wu[1,2*]

[1]Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA; E-mail: *zhenkewu@umich.edu

[2]Michigan Institute for Data Science (MIDAS), University of Michigan, Ann Arbor, MI 48019, USA;

[3]Environmental and Occupational Health, Milken Institute School of Public Health, The George Washington University, Washington, DC 20052, USA

A1 Details of the Variational Inference Algorithm

Step 0. Initialize relevant moments that define the variational distributions at iteration $t = 0$: $E_{q_t}[\alpha_{vk} \mid s_u = 1]$, $V_{q_t}[\alpha_{vk} \mid s_u = 1]$, $E_{q_t}[\gamma_{jk}]$, $V_{q_t}[\gamma_{jk}]$, $E_{q_t}[s_u]$, ψ_{jk}, $\phi_k^{(v)}$ and the hyperparameters to be optimized τ_{1kl}, τ_{2jk}, for $j = 1, \ldots, J$, $k = 1, \ldots, K-1$, $l = 1, \ldots, L$; calculate an initial ELBO. In particular, we compute the first and second moments as follows:

$$E_{q_t}[\eta_{vk}^2] = \sum_{u \in a(v)} \left\{ p_{tu}(\sigma^2_{\alpha_{uk,1}} + (1-p_{tu})\mu^2_{\alpha_{uk,1}}) \right\} + E^2_{q_t}[\eta_{vk}],$$
$$E_{q_t}[\alpha^2_{uk}] = p_{tu}(\sigma^2_{\alpha_{uk,1}} + \mu^2_{\alpha_{uk,1}}) + (1-p_{tu})\sigma^2_{\alpha_{uk,0}},$$

$E_{q_t}[\gamma^2_{jk}] = \sigma^2_{\gamma_{jk,1}} + \mu^2_{\gamma_{jk,1}}$, $E_{q_t}[\eta_{vk}] = \sum_{u \in a(v)} E_{q_t}[\xi_{uk}]$, $E_{q_t}[\xi_{uk}] = E_{q_t}[s_u \alpha_{uk}] = p_{tu}\mu_{\alpha_{uk,1}}$ from the initialized moments.

At Step $t+1$, iterate between Step 1a-1c:

Step 1a. Update $q_{t+1}(\boldsymbol{Z}_i^{(v)})$, $i = 1, \ldots, n_v$, $v \in \mathcal{V}_L$ by a multinomial with probabilities $\boldsymbol{r}_i^{(v),t+1}$:

$$
\begin{aligned}
&r_{ik}^{(v),(t+1)} \propto \\
&\exp\Bigg[\sum_{j=1}^{J} \log\sigma(\psi_{jk}) + \left[X_{ij}^{(v)} E_{q_t}(\gamma_{jk}) - \psi_{jk}\right]/2 - g(\psi_{jk})\left\{E_{q_t}(\gamma_{jk}^2) - \psi_{jk}^2\right\} \\
&+ \sum_{m<k}\left(\log\sigma(\phi_m^{(v)}) + \left[-E_{q_t}(\eta_{vm}) - \phi_m^{(v)}\right]/2 - g(\phi_m^{(v)})\left\{E_{q_t}(\eta_{vm}^2) - \phi_m^{(v),2}\right\}\right) \\
&+ \mathbf{1}\{k < K\}\left(\log\sigma(\phi_k^{(v)}) + \left[E_{q_t}(\eta_{vk}) - \phi_k^{(v)}\right]/2 - g(\phi_k^{(v)})\left\{E_{q_t}(\eta_{vk}^2) - \phi_k^{(v),2}\right\}\right)\Bigg].
\end{aligned}
$$

Step 1b. Find q that maximizes $\mathcal{E}_t^*(q)$, which results in the variational update for $\boldsymbol{\gamma}$ and $(s_u, \boldsymbol{\alpha}_u)$. We first do the update for the root node $u = u_0$ when $\boldsymbol{\gamma}$ gets updated; for non-root nodes, the $\boldsymbol{\gamma}$ is not updated. In particular, we have

$$
\begin{aligned}
&\log q_{t+1}(\boldsymbol{\gamma}, s_u, \boldsymbol{\alpha}_u) = E_{q_t(-(\boldsymbol{\gamma}, s_u, \boldsymbol{\alpha}_u))}\left[\log\{h(\boldsymbol{\psi}, \boldsymbol{\gamma}, \mathbf{Z}) h(\boldsymbol{\phi}, \boldsymbol{s}, \boldsymbol{\alpha}, \mathbf{Z}) \pi(\boldsymbol{s}, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\varrho})\}\right] + const \\
&= -\sum_{j=1}^{J}\sum_{k=1}^{K}\left\{\frac{1}{2\sigma_{\gamma_{jk},1}^{2,(t+1)}}\left(\gamma_{jk} - \mu_{\gamma_{jk}}^{(t+1)}\right)^2 + \frac{1}{2}\log\left[2\pi\sigma_{\gamma_{jk},1}^{2,(t+1)}\right]\right\} \\
&\quad - s_u \sum_{k=1}^{K-1}\left\{\frac{1}{2\sigma_{\alpha_{uk},1}^2}(\alpha_{uk} - \mu_{\alpha_{uk}})^2 + \frac{1}{2}\log[2\pi\sigma_{\alpha_{uk},1}^2]\right\} \\
&\quad - (1 - s_u)\sum_{k=1}^{K-1}\left\{\frac{1}{2\tau_{1kl_u} h_u}\alpha_{uk}^2 + \frac{1}{2}\log\{2\pi\tau_{1kl_u} h_u\}\right\} + s_u w_u^{(t+1)} + const
\end{aligned}
$$

where const does not depend on s_u, $\boldsymbol{\gamma}$ and $\boldsymbol{\alpha}_u$; $\mu_{\gamma_{jk}}^{(t+1)} = B_{jk}^{(t)}/A_{jk}^{(t)}$, $\sigma_{\gamma_{jk},1}^{2,(t+1)} = 1/A_{jk}^{(t)}$, $\mu_{\alpha_{uk}}^{(t+1)} =$

$D_{uk}^{(t)}/C_{uk}^{(t)}$, $\sigma^{2,(t+1)}_{\alpha_{uk},1} = 1/C_{uk}^{(t)}$ where

$$A_{jk}^{(t)} = \frac{1}{\tau_{2jk}} + 2\sum_{v\in\mathcal{V}_L}\sum_{i=1}^{n_v} r_{ik}^{(v),t} g(\psi_{jk}) \underline{X_{ij}^{(v),2}}, \tag{S1}$$

$$B_{jk}^{(t)} = \sum_{v\in\mathcal{V}_L}\sum_{i=1}^{n_v}\left\{r_{ik}^{(v),t}X_{ij}^{(v)}/2\right\}, \tag{S2}$$

$$C_{uk}^{(t)} = \frac{1}{\tau_{1kl_u}h_u} + 2\sum_{v\in\mathcal{V}_L\cap d(u)}\sum_{i=1}^{n_v}\sum_{m=k}^{K} r_{im}^{(v),t} g(\phi_k^{(v)}),\ \text{for } k=1,\ldots,K-1, \tag{S3}$$

$$D_{uk}^{(t)} = \sum_{v\in\mathcal{V}_L\cap d(u)}\sum_{i=1}^{n_v}\frac{1}{2}r_{ik}^{(t)} - \frac{1}{2}\sum_{m=k+1}^{K} r_{im}^{(t)} - 2\left(\sum_{m=k}^{K} r_{im}^{(t)} g(\phi_k^{(v)})\sum_{w\in a(v)\setminus u} E_{q_t}[s_w\alpha_{wk}]\right), \tag{S4}$$

$$w_u^{(t+1)} = E_{q_t}\left[\log\frac{\rho_{l_u}}{1-\rho_{l_u}}\right] - \frac{1}{2}\sum_{k=1}^{K-1}[\log(\tau_{1kl_u}h_u) + \log(C_{uk}^{(t)})] + \sum_{k=1}^{K-1}\frac{D_{uk}^{(t),2}}{2C_{uk}^{(t)}}. \tag{S5}$$

It is readily recognized that $q_{t+1}(\boldsymbol{\gamma}, s_u, \boldsymbol{\alpha}_u)$ is a two-component mixture distribution:

$$\underbrace{\prod_{j=1}^{J}\prod_{k=1}^{K}\mathcal{N}\left(\boldsymbol{\gamma}_{jk}\mid\mu_{\gamma_{jk}}^{(t+1)}, \sigma^{2,(t+1)}_{\gamma_{jk},1}\right)}_{(I)}$$
$$\cdot\underbrace{\prod_{k=1}^{K-1}\mathcal{N}\left(\alpha_{uk}\mid s_u\mu_{\alpha_{uk}}^{(t+1)}, s_u\sigma^{(t+1)}_{\alpha_{uk},1} + (1-s_u)\sigma^{(t+1)}_{\alpha_{uk},0}\right)\cdot \mathsf{Bernoulli}\left(s_u; p_u^{(t+1)}\right)}_{(II)}, \tag{S6}$$

where $\sigma^{2,(t+1)}_{\gamma_{jk},0} = \tau_{2jk}$, $\sigma^2_{\alpha_{ukl_u}0} = \tau_{1kl_u}h_u$; and $p_u^{(t+1)}$ satisfies $\log\left\{\frac{p_u^{(t+1)}}{1-p_u^{(t+1)}}\right\} = w_u^{(t+1)}$.

Of note, because given s_u the update is additive with respect to j and k, we have induced factorization $\prod_{j=1}^{J}\prod_{k=1}^{K} q_{t+1}(\gamma_{jk})\prod_{k=1}^{K-1} q_{t+1}(\alpha_{uk}\mid s_u)$. In the variational family, we did not assume this factorization. However, we do obtain updates that factorize. This phenomenon is determined by the underlying generative model (the graph structure that determines the joint distribution) and the variational assumption (which determines blocks of parameters to iteratively update) (see, e.g., Bishop, 2006, Section 10.2.5). We update $q_{t+1}(\boldsymbol{\gamma})$ according to component (I) in (S6) and $q_{t+1}(s_u, \boldsymbol{\alpha}_u)$ according to (II) in (S6) when u is the root node; for a non-root node, we only update $q_{t+1}(s_u, \boldsymbol{\alpha}_u)$ according to component (II) in (S6).

Step 1c. Update $q_{t+1}(\rho_l)$, $l=1,\ldots,L$ by $\mathsf{Beta}(e_l^{(t+1)}, f_l^{(t+1)})$, where $e_l^{(t+1)} = \sum_{u\in\mathcal{V}:l_u=l} E_{q_t}[s_u] + a_\ell$ and $f_l^{(t+1)} = \sum_{u\in\mathcal{V}:l_u=l} E_{q_t}[1-s_u] + b_l$.

For every d steps above, do Step 2-4:

Step 2. Update variational parameters ψ_{jk} and $\phi_k^{(v)}$ by optimizing the lower bound $\mathcal{E}^*(q)$ with a generic q which leads to the updates:

$$\psi_{jk}^{(t+1)} = \sqrt{E_{q_t}[\gamma_{jk}^2]}, \qquad \phi_k^{(v),(t+1)} = \sqrt{E_{q_t}[\eta_{vk}^2]}. \tag{S7}$$

Step 3. We update the hyper-parameters τ_{1kl} by

$$\tau_{1kl} = \frac{1}{\sum_{u:l_u=l} 1} \sum_{u \in \mathcal{V}:l_u=l} E_{q_t}[\alpha_{uk}^2/h_u], \tag{S8}$$

and update τ_{2jk} by $\tau_{2jk} = E_{q_t}[\gamma_{jk}^2]$.

Step 4. Compute the ELBO value $\mathcal{E}^*(q_{t+1})$ at step $t+1$ by plugging in the update q_{t+1} in Step 1-3, and the updated local variational parameters $\boldsymbol{\psi}^{(t+1)}$, $\boldsymbol{\phi}^{(v),(t+1)}$ and hyperparameters $(\boldsymbol{\tau}_1, \boldsymbol{\tau}_2)$. Stop the iteration once the absolute change in ELBO is less than a tolerance `tol=1e-8`. The hyperparameter updates are often slower than the variational parameters to converge in terms of the ELBO. In practice, we can separate the tolerance levels for the hyperparameter (`hyper_tol=1e-4`) and VI parameter updates (e.g., `tol=1e-8`). One may adjust the size of `update_freq_hyper` to explore ways to speed up the convergence.

The ELBO $\mathcal{E}^*(q)$ is computed according to Appendix A1.1.

Of note, in Step 1a, the update is a function of the moments of unknown quantities, e.g., $E_{q_t}[Z_{ik}^{(v)}]$, $E_{q_t}[\gamma_{jk}]$, $v \in \mathcal{V}_L$, $u \in \mathcal{V}$, which can be computed from the previous iteration's $q_t(\boldsymbol{Z}_t^{(v)})$ and the current update in Equation (S6) involving terms for nodes $w \in a(v) \setminus u$. In Step 1b, the update depends on the moment $E_{q_t}[s_u]$. In Step 1c, the update $q_{t+1}(\boldsymbol{Z}_i^{(v)})$ depends only on the moments $E_{q_t}(\gamma_{jk})$, $E_{q_t}(\gamma_{jk}^2)$, $E_{q_t}(\eta_{vk})$ and $E_{q_t}(\eta_{vk})^2$. In Step 2, 3 and 4, the updates for variational and hyper-parameters are also only dependent upon the moments. We initialize and iteratively update the moments until ELBO convergence. We access the approximate posterior densities when needed by plugging in the converged values of the moments.

A1.1 Computing ELBO

$$\mathcal{E}^*(q; \boldsymbol{\phi}, \boldsymbol{\psi}, \boldsymbol{\tau}_1, \boldsymbol{\tau}_2) =$$

$$= \sum_{v \in \mathcal{V}_L} \sum_{i=1}^{n_v} \sum_{k=1}^{K} r_{ik}^{(v)} \left\{ \sum_{m<k} \left(\log \sigma(\phi_m^{(v)}) + \left[(-1) E_{q_t}(\eta_{vm}) - \phi_m^{(v)} \right] /2 - g(\phi_m^{(v)}) \left\{ E_{q_t}(\eta_{vm}^2) - \phi_m^{(v),2} \right\} \right) \right.$$

$$+ \mathbf{1}\{k < K\} \left(\log \sigma(\phi_k^{(v)}) + \left[E_{q_t}(\eta_{vk}) - \phi_k^{(v)} \right] /2 - g(\phi_k^{(v)}) \left\{ E_{q_t}(\eta_{vk}^2) - \phi_k^{(v),2} \right\} \right)$$

$$\left. + \sum_{j=1}^{J} \log(\sigma(\psi_{jk})) + E_q \left[\{ X_{ij}^{(v)} \beta_{vjk} - \psi_{jk} \}/2 \right] - g(\psi_{jk}) E_q \left(\{ X_{ij}^{(v)} \beta_{jk} \}^2 - \psi_{jk}^2 \right) \right\} \quad \text{(S9)}$$

$$+ \sum_{u \in \mathcal{V}} E_q[\log \rho_{l_u}] E_q[s_u] + E_q[\log(1 - \rho_{l_u})] E_q[1 - s_u] \quad \text{(S10)}$$

$$+ \sum_{l=1}^{L} (a_l - 1) E_q[\log \rho_l] + (b_l - 1) E_q[\log(1 - \rho_l)] - \log Beta(a_l, b_l) \quad \text{(S11)}$$

$$- \sum_{u \in \mathcal{V}} \sum_{k=1}^{K-1} \left(\frac{E_q[\alpha_{uk}^2]}{2\tau_{1kl_u} h_u} + \frac{1}{2} \log(2\pi \tau_{1kl_u} h_u) \right) \quad \text{(S12)}$$

$$- \sum_{j=1}^{J} \sum_{k=1}^{K} \left(\frac{E_q[\gamma_{jk}^2]}{2\tau_{2jk}} + \frac{1}{2} \log(2\pi \tau_{2jk}) \right) \quad \text{(S13)}$$

$$+ \frac{1}{2} \sum_{j=1}^{J} \sum_{k=1}^{K} E_q[(\gamma_{ujk} - \mu_{ujk})^2 / \sigma_{\gamma_{ujk},1}^2] + \log(2\pi \sigma_{\gamma_{ujk,1}}^2) \quad \text{(S14)}$$

$$- \sum_{u \in \mathcal{V}} E_q[s_u] \log(p_u) + E_q[1 - s_u] \log(1 - p_u) \quad \text{(S15)}$$

$$+ \frac{1}{2} \sum_{u \in \mathcal{V}} \sum_{k=1}^{K-1} E_q[s_u (\alpha_{uk} - \mu_{\alpha_{uk}})^2 / \sigma_{\alpha_{uk},1}^2] + E_q[s_u] \log(2\pi \sigma_{\alpha_{uk,1}}^2) \quad \text{(S16)}$$

$$+ \frac{1}{2} \sum_{u \in \mathcal{V}} \sum_{k=1}^{K-1} [E_q[1 - s_u] + E_q[1 - s_u] \log(2\pi \tau_{1kl_u} h_u)] \quad \text{(S17)}$$

$$- \sum_{l=1}^{L} \{ (e_l - 1) E_q[\log(\rho_l)] + (f_l - 1) E_q[\log(1 - \rho_l)] - \log \mathsf{Beta}(e_l, f_l) \} \quad \text{(S18)}$$

$$- \sum_{v \in \mathcal{V}_L} \sum_{i=1}^{n_v} \sum_{k=1}^{K} r_{ik}^{(v)} \log(r_{ik}^{(v)}) \quad \text{(S19)}$$

A2 Additional Details of Simulation Setup

For the small tree, we set three leaf groups Group 1: $\{6, 7, 8\}$, Group 2 $\{9, 10, 11\}$ and Group 3 $\{12, 13, 14, 15, 16\}$. The true grouping of leaf nodes is obtained once we only set $s_1 = s_2 = s_3 = 1$ and zero otherwise. We have set $\boldsymbol{\pi}_v = (0.356, 0.416, 0.229)$, $(0.803, 0.164, 0.033)$ and $(0.6, 0.3, 0.1)$ for $v \in G_1, G_2, G_3$, respectively. For the larger tree, we use the tree based on the phylogenetic tree in the data analysis in Section 6 of the Main Paper. In particular, we only let internal nodes $3, 4, 76, 120$ to have non-zero s_u, resulting in four true groups of 73, 45, 14 and 1 leaf nodes; see a zoomed-in version of the larger tree in Appendix Figure S3. The group-specific true class probabilities are $(0.356, 0.536, 0.108)$, $(0.803, 0.183, 0.014)$, $(0.069, 0.373, 0.557)$ and $(0.6, 0.3, 0.1)$. In both cases, we set the response probability profiles to be $\underbrace{(0.9, \ldots, 0.9)}_{J}$, $\underbrace{(0.5, \ldots, 0.5)}_{J}$ and $\underbrace{(0.1, \ldots, 0.1)}_{J}$ for three classes.

A3 Appendix Figures

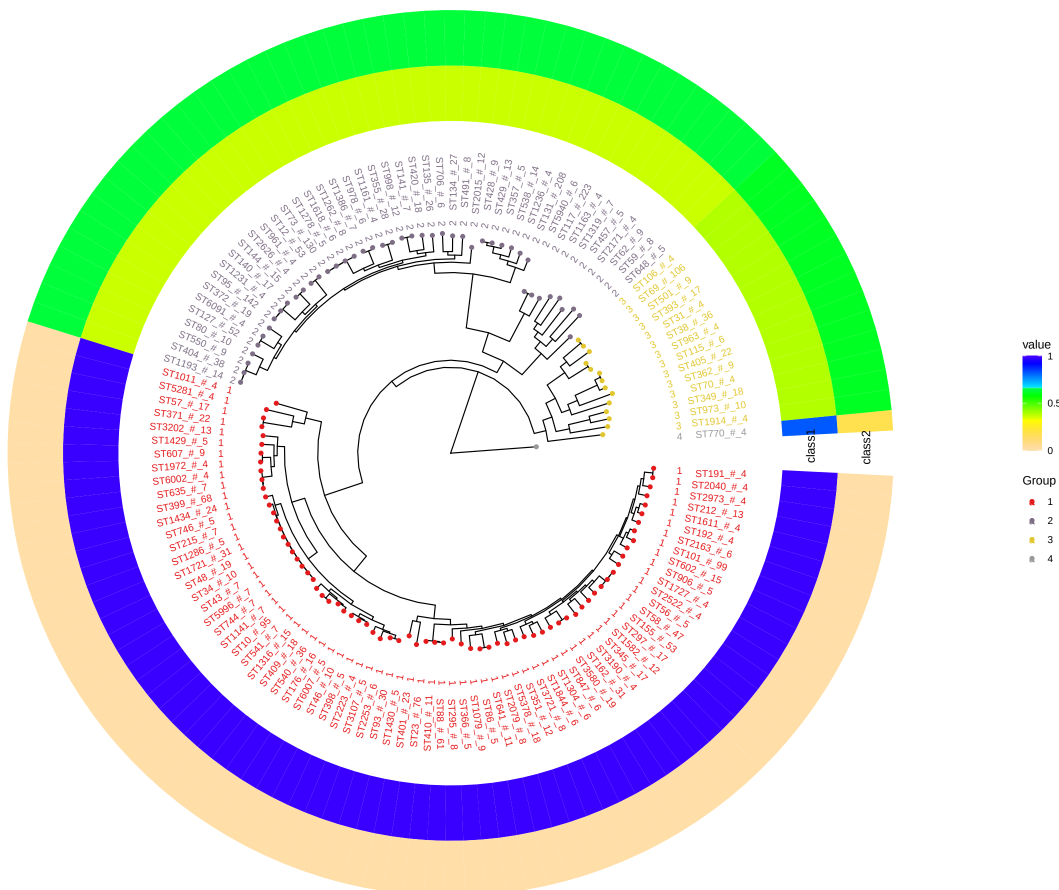

Figure S1: Data results: Estimated groups and class proportions based on fixed ad hoc leaf groupings.

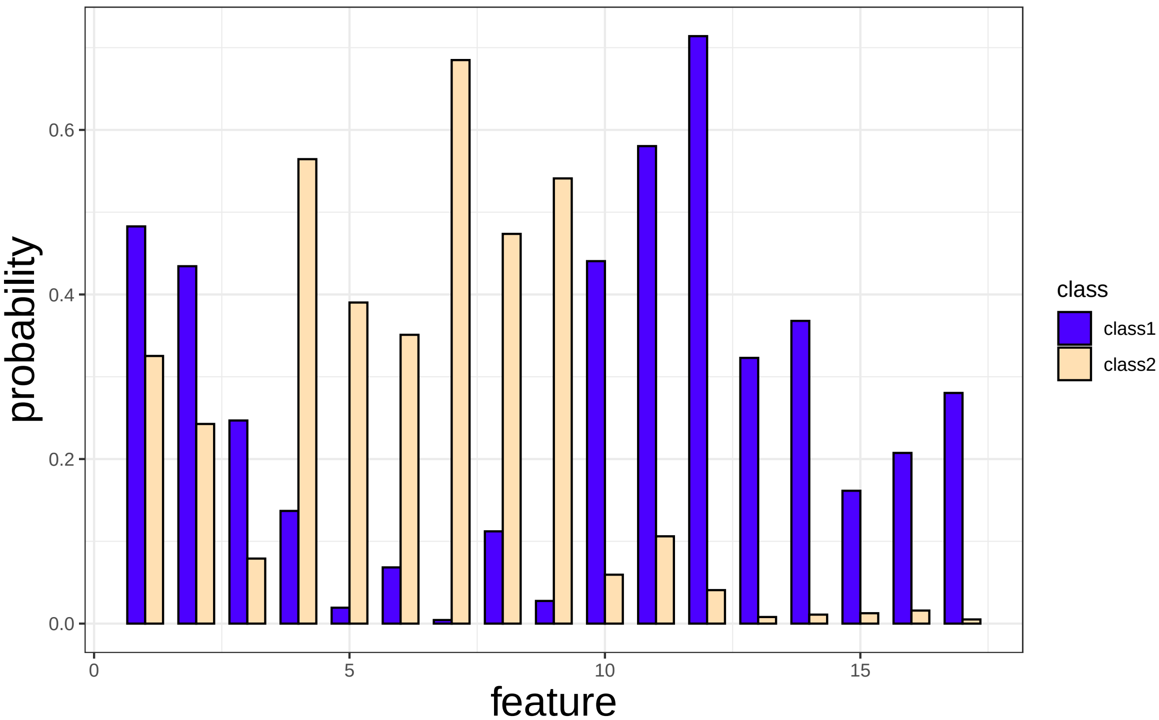

Figure S2: Data results: Estimated class-specific response profiles based on fixed ad hoc leaf groupings..

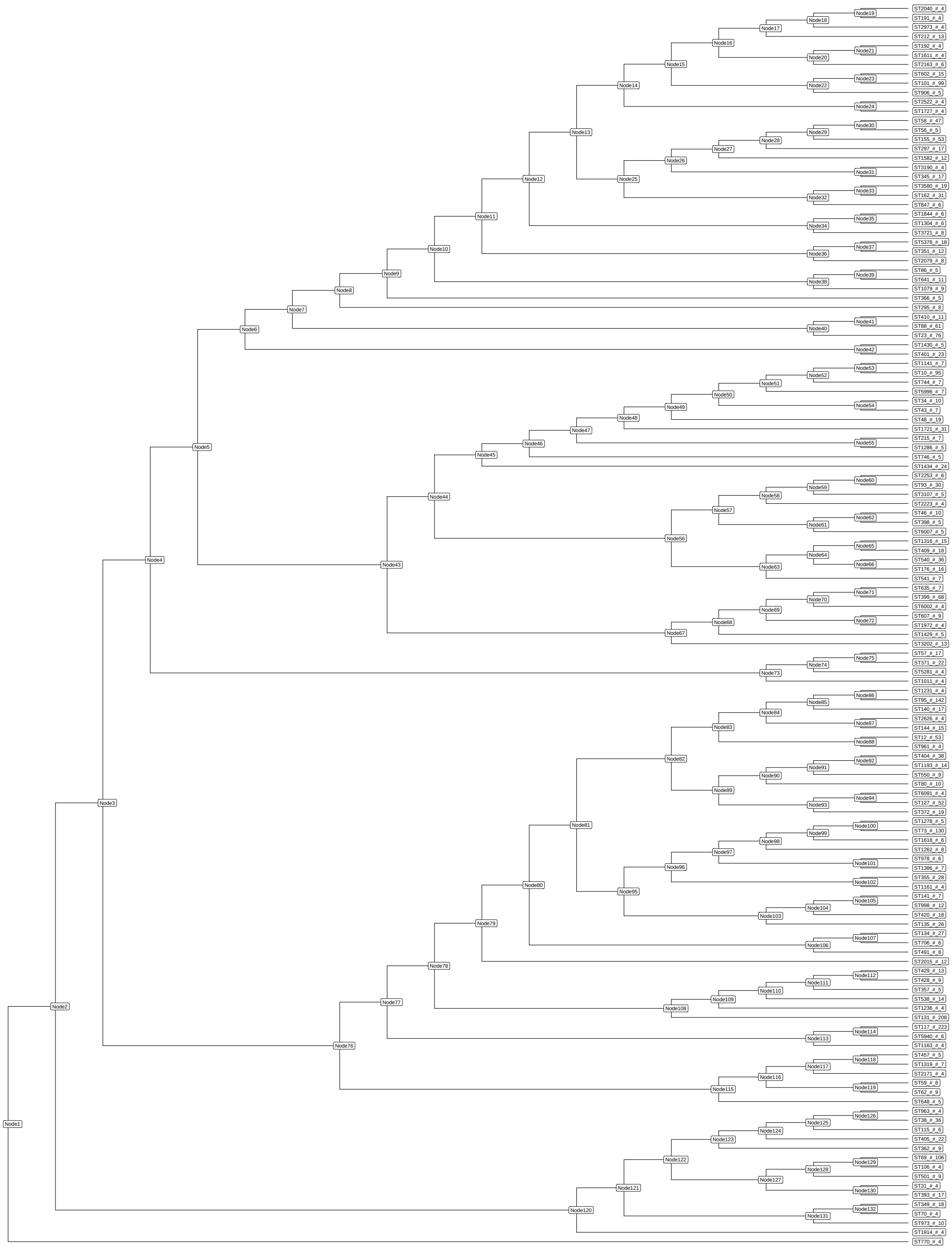

Figure S3: Zoomed-in version of the larger tree in Figure 3b of the Main Paper.

References

Bishop, C. M. (2006). *Pattern recognition and machine learning*. springer.