Bayesian Analysis (0000) **00**, Number 0, pp. 1

Supplementary Material for "Adaptive shrinkage in Pólya tree type models"

Li Ma*

S1. Technical proofs

Proof of Lemma 1. The existence of a collection of PACs for G follows immediately from the definition of PACs by letting $\theta(A) = G(A_l)/G(A)$ for each A such that $G(A) > 0$ and $\theta(A) = 0$ otherwise. The uniqueness follows because $\mathcal{A}^{(\infty)}$, along with the empty set, forms a π-system that generates the Borel σ-algebra. So by the extension theorem, two distributions with the same PACs on all As such that $G(A) > 0$ must be the same up to a set of μ-measure 0. □

Proof of Lemma 2. If $n(A) = 0$ then by definition $\xi_A(i, \boldsymbol{\phi}) = 1$. If A has no children, then also by definition $\xi_A(i, \boldsymbol{\phi}) = q_0(\boldsymbol{x}|A)$. If $n(A) = 1$, then $\xi_A(i, \boldsymbol{\phi})$ becomes the conditional prior predictive density on A valued at x, which is just $q_0(x|A)$ since the Markov-APT conditional A is still an Markov-tree and by Theorem 1 its predictive density is Q_0. Finally we consider the case when A has children and $n(A) \geq 2$. For $A \in \mathcal{A}^{(\infty)} \backslash \Omega$,

$$\begin{aligned}
\xi_A(i, \boldsymbol{\phi}) &= \int q(\boldsymbol{x}|A)\pi(dq \,|\, \boldsymbol{\phi}, C(A_p) = i) \\
&= \int q(\boldsymbol{x}|A)\pi(dq \,|\, \boldsymbol{\phi}, C(A) = i')\gamma_{i,i'}(A) \\
&= \sum_{i'=1}^{I} \int q(\boldsymbol{x} \,|\, A)\pi(dq \,|\, \boldsymbol{\phi}, C(A) = i')\gamma_{i,i'}(A) \\
&= \sum_{i'=1}^{I} \gamma_{i,i'}(A) \int \theta(A)^{n(A_l)}(1 - \theta(A))^{n(A_r)}\pi(\theta(A) \,|\, C(A) = i') \times \\
&\qquad\qquad \int q(\boldsymbol{x} \,|\, A_l)q(\boldsymbol{x} \,|\, A_r)\pi(dq \,|\, \boldsymbol{\phi}, C(A) = i') \\
&= \sum_{i'=1}^{I} \gamma_{i,i'}(A) M_A^{i'}(\boldsymbol{\theta}_0)\xi_{A_l}(i', \boldsymbol{\phi})\xi_{A_r}(i', \boldsymbol{\phi}).
\end{aligned}$$

For $A = \Omega$, the derivation follows similarly with $\gamma_{i,i'}(A)$ replaced by $\gamma_{i'}(A)$. □

Proof of Theorem 1. This theorem follows by two applications of Bayes rule. For $A \in \mathcal{A}^{(\infty)} \backslash \Omega$, the posterior transition probability is

$$\tilde{\gamma}_{i,i'}(A) = P(C(A) = i' \,|\, C(A_p) = i, \boldsymbol{x}(A))$$

*Department of Statistical Science, Duke University, Durham, NC 27708, USA. li.ma@duke.edu

 DOI: 0000

$$= \int q(\boldsymbol{x} \mid A)\gamma_{i,i'}(A)\pi(dq \mid C(A) = i') \Big/ \int q(\boldsymbol{x} \mid A)\pi(dq \mid C(A_p) = i)$$

$$= \begin{cases} \gamma_{i,i'}(A)M_A^{i'}(\boldsymbol{\theta}_0)\xi_{A_l}(i', \boldsymbol{\phi})\xi_{A_r}(i', \boldsymbol{\phi})/\xi_A(i, \boldsymbol{\phi}) & \text{if } A \text{ has children} \\ \gamma_{i,i'}(A)M_A^{i'}(\boldsymbol{\theta}_0)/\xi_A(i, \boldsymbol{\phi}) & \text{otherwise.} \end{cases}$$

Therefore the state transition probability matrix is

$$\tilde{\boldsymbol{\gamma}}(A) = \boldsymbol{D}'(A)^{-1}\boldsymbol{\gamma}(A)\boldsymbol{D}''(A).$$

For $A = \Omega$, the proof for the initial state probability vector is similar. The expression follows because the overall marginal likelihood is $\xi_\Omega(1, \boldsymbol{\phi})$. □

Proof of Theorem 2. Given $\boldsymbol{\nu}$, Q has a PT distribution with mean Q_0. That is,

$$\mathrm{E}(Q(B)|\boldsymbol{\nu}, Q_0) = Q_0(B).$$

The result follows immediately by the law of iterated expectation. □

Proof of Theorem 3. Let $Q^{(k)}$ be the level-k truncated version of Q. That is, $Q^{(k)}(A) = Q(A)$ for all $A \in \mathcal{A}^{(k)}$ and $Q^{(k)}(\cdot|A) = Q_0(\cdot|A)$ for all $A \in \mathcal{A}^k$. By the same argument as in the proof of Theorem 1 in Wong and Ma (2010) (with Q_0 replacing μ), we know that $Q^{(k)}$ converges in total variational distance to Q as $k \to \infty$. By construction, $Q^{(k)} \ll Q_0$ for all k. Now for any set B such that $Q(B) > 0$, then there must exist some k such that $Q^{(k)}(B) > 0$, and therefore $Q_0(B) > 0$. Hence $Q \ll Q_0$. □

Proof of Theorem 4. Let $\tilde{q} = q/q_0$ and $\tilde{g} = g/q_0$ where $q_0 = dQ_0/d\mu$. Our goal is to prove that for any $\tau > 0$,

$$P\left(\int |\tilde{q} - \tilde{g}| dQ_0 < \tau\right) > 0.$$

First we assume that $\tilde{g}$ is continuous and bounded, and let M be a finite upperbound of $\tilde{g}$. For any $\sigma > 0$, there exists a compact set E such that there is a partition $\Omega = \cup_i A_i$ such that the diameter of each $A_i \cap E$ is less than σ. By the absolute continuity of G w.r.t Q_0, there exists $\beta(\sigma) > 0$ such that $G(E^c) < \beta(\sigma)$ if $Q_0(E^c) < \sigma$ and $\beta(\sigma) \downarrow 0$ as $\sigma \downarrow 0$. We define the modulus of continuity of $\tilde{g}$ on E as

$$\delta_E(\epsilon) = \sup_{x,y \in E: |x-y| < \epsilon} |\tilde{g}(x) - \tilde{g}(y)|.$$

Note that by the continuity of $\tilde{g}$ and the compactness of E, $\delta_E(\epsilon) \downarrow 0$ as $\epsilon \downarrow 0$. Now we approximate $\tilde{g}$ by a step function $\tilde{g}^*(x) = \sum_i \tilde{g}_i^* I_{A_i}$ where $\tilde{g}_i^* = \int_{A_i \cap E} \tilde{g} dQ_0 / Q_0(A_i \cap E)$. Let $D_\epsilon(\tilde{g})$ be the set of step functions $h(\cdot) = \sum_i h_i I_{A_i}(\cdot)$ such that $\sup_i |h_i - \tilde{g}_i^*| < \delta_E(\epsilon) + M\sigma$.

Suppose $h \in D_\epsilon(\tilde{g})$. For any $B \in \mathcal{B}$, the Borel sets, we have $B_i = B \cap A_i$. Then

$$\left| \int_B (h - \tilde{g}) dQ_0 \right| \le \sum_i |h_i - \tilde{g}_i^*| Q_0(B_i) + \sum_i \left| \tilde{g}_i^* \, Q_0(B_i) - \int_{B_i} \tilde{g} dQ_0 \right|$$

$$\leq(\delta_E(\epsilon)+M\sigma)Q_0(B)+\sum_i\left|\tilde{g}_i^*\,Q_0(B_i\cap E)-\int_{B_i\cap E}\tilde{g}dQ_0\right|+\sum_i\left|\tilde{g}_i^*\,Q_0(B_i\cap E^c)-\int_{B_i\cap E^c}\tilde{g}dQ_0\right|$$
$$\leq(\delta_E(\epsilon)+M\sigma)Q_0(B)+\sum_i r_i+2M\cdot Q_0(E^c)$$
$$<(\delta_E(\epsilon)+M\sigma)Q_0(B)+\sum_i r_i+2M\sigma$$

where

$$\begin{aligned}r_i&=Q_0(B_i\cap E)\left|\frac{\int_{A_i\cap E}\tilde{g}dQ_0}{Q_0(A_i\cap E)}-\frac{\int_{B_i\cap E}\tilde{g}dQ_0}{Q_0(B_i\cap E)}\right|\\&=Q_0(B_i\cap E)\left|\frac{\int_{A_i\cap E}(\tilde{g}(x)-\tilde{g}(x_i))\,q_0(x)\,dx}{Q_0(A_i\cap E)}-\frac{\int_{B_i\cap E}(\tilde{g}(x)-\tilde{g}(x_i))\,q_0(x)\,dx}{Q_0(B_i\cap E)}\right|\end{aligned}$$

for some $x_i\in B_i$. Thus

$$|r_i|<2\delta_E(\epsilon)Q_0(B_i)$$

and so

$$\left|\int_B(h-\tilde{g})dQ_0\right|<3\delta_E(\epsilon)Q_0(B)+3M\sigma\quad\text{for all } B\in\mathcal{B}.$$

Therefore by taking $B=\{x:h>\tilde{g}\}$ and $B=\{x:h\leq\tilde{g}\}$, we get

$$\int|h-\tilde{g}|dQ_0<3\delta_E(\epsilon)+6M\sigma.$$

Now we let $Q^{(k)}$ be the level-k truncated version of Q. That is, $Q^{(k)}(A)=Q(A)$ for all $A\in\mathcal{A}^{(k)}$ and $Q^{(k)}(\cdot|A)=Q_0(\cdot|A)$ for all $A\in\mathcal{A}^k$. By the conditions in the theorem, we have for $\tilde{q}^{(k)}=q^{(k)}/q_0$ where $q^{(k)}=dQ^{(k)}/d\mu$,

$$P\left(\tilde{q}^{(k)}\in D_\epsilon(\tilde{g})\text{ for all large }k\right)>0.$$

Thus

$$P\left(\int|\tilde{q}^{(k)}-\tilde{g}|dQ_0<3\delta_E(\epsilon)+6M\sigma\ \text{ for all large }k\right)>0.$$

But since

$$P\left(\int|\tilde{q}^{(k)}-\tilde{q}|dQ_0\to0\right)=1,$$

combining these we get

$$P\left(\int|\tilde{q}-\tilde{g}|dQ_0<4\delta_E(\epsilon)+6M\sigma\right)>0.$$

The result follows by letting $\epsilon\downarrow0$ and $\sigma\downarrow0$.

Finally, if $\tilde{g}$ is not continuous and bounded, then since Q_0 is a probability measure, $\tilde{g}$ can be approximately arbitrarily well in L_1 w.r.t Q_0 by a continuous bounded density. □

Proof of Theorem 5. Let $p_0 = dP_0/d\mu$, $q_0 = dQ_0/d\mu$, $\tilde{p}_0 = dP_0/dQ_0$, and for any $Q \ll Q_0$, $\tilde{q} = dQ/dQ_0$. Let M be a finite upperbound of $\tilde{p}_0$. Then the Kullback-Leibler (K-L) distance between p_0 and q is given by

$$\mathrm{KL}_\mu(p_0, q) = \int p_0 \log(p_0/q) d\mu = \int \tilde{p}_0 \log(\tilde{p}_0/\tilde{q}) dQ_0 = \mathrm{KL}_{Q_0}(\tilde{p}_0, \tilde{q}).$$

By Lusin's theorem we have a compact $E \subset \Omega$ with $Q_0(E^c) < \epsilon'$ such that $\tilde{p}_0$ is continuous on E. This E can be chosen such that for every $\epsilon > 0$, there exists a partition, $\Omega = \cup_i A_i$ with all $A_i \in \mathcal{A}^{(k)}$ for some k, such that the diameter of each $A_i \cap E$ is less than ϵ. We define

$$\delta_E(\epsilon) = \sup_{x,y \in E: |x-y|<\epsilon} |\tilde{p}_0(x) - \tilde{p}_0(y)| \quad \text{and} \quad d_i = \max\left(\sup_{A_i \cap E} \tilde{p}_0(x) + \delta_E(\epsilon), \epsilon' \right)$$

and let $D_\epsilon(\tilde{p}_0)$ be the collection of step functions $g(x) = \sum_i g_i \mathbf{1}_{A_i}(x)$ with $d_i \le g_i < d_i + \delta_E(\epsilon)$. For every $g \in D_\epsilon(\tilde{p}_0)$, let $\tilde{g}$ be the normalized version of g, that is $\tilde{g} = g/\int g dQ_0$. Then

$$\int_E (g - \tilde{p}_0) dQ_0 - \int_{E^c} |g - \tilde{p}_0| dQ_0 \le \int (g - \tilde{p}_0) dQ_0 \le \int_E (g - \tilde{p}_0) dQ_0 + \int_{E^c} |g - \tilde{p}_0| dQ_0,$$

and so

$$\delta_E(\epsilon) - (2M + \epsilon')\epsilon' \le \int (g - \tilde{p}_0) dQ_0 \le 3\delta_E(\epsilon) + (2M + \epsilon')\epsilon'.$$

Thus for any fixed ϵ, when ϵ' is small enough, we have $\int (g - \tilde{p}_0) dQ_0 \ge 0$, and thus,

$$\log\left(\int g dQ_0 \right) = \log\left(1 + \int (g - \tilde{p}_0) dQ_0 \right) \le 3\delta_E(\epsilon) + (2M + \epsilon')\epsilon'.$$

Now,

$$\begin{aligned} 0 \le \mathrm{KL}_{Q_0}(\tilde{p}_0, \tilde{g}) &= \int \tilde{p}_0 \log(\tilde{p}_0/\tilde{g}) dQ_0 \\ &= \int_E \tilde{p}_0 \log(\tilde{p}_0/g) dQ_0 + \int_{E^c} \tilde{p}_0 \log(\tilde{p}_0/g) dQ_0 + \log\left(\int g dQ_0 \right) \\ &\le M \log(M/\epsilon')\epsilon' + 3\delta_E(\epsilon) + (2M + \epsilon')\epsilon'. \end{aligned}$$

By first choosing ϵ' and then ϵ small enough, we can make $\mathrm{KL}_{Q_0}(\tilde{p}_0, \tilde{g})$ arbitrarily small. So p_0 lies in the K-L support of π. Therefore, by Schwartz's theorem, we have posterior consistency at p_0 under the weak topology. □

S2. Numerical evaluation of the $M_A^i(\boldsymbol{\theta}_0)$ terms

Because ν is one-dimensional, it is easy to evaluate the $M_A^i(\boldsymbol{\theta}_0)$ terms numerically. Specifically, we evaluate $f^i = dF^i/d\mu$ on a grid of different ν values $\nu_{(1)}, \nu_{(2)}, \ldots, \nu_{(H)}$ covering the support of F^i, and approximate $M_A^{(i)}(\boldsymbol{\theta}_0)$ using a finite Riemann integral.

This approximation becomes particularly straightforward when F^i is uniform on an interval $[a(i), b(i))$ under some transformed scale of ν, such as $\log_{10}(\nu)$, which we adopt in all of our numerical examples. In this case one can choose the grid points to be of equal distance on the transformed scale and $M_A^i(\boldsymbol{\theta}_0) \approx \frac{1}{H}\sum_{h=1}^{H} M_A(\boldsymbol{\theta}_0, \nu_{(h)})$. The marginal density usually changes relatively slowly in the shrinkage parameter ν and hence we found that in all of our examples adopting $H = 5$ is sufficient in producing numerically accurate results.

S3. $\pi(\mathcal{C}, \mathcal{A}^{(\infty)} \mid \boldsymbol{\phi}, \boldsymbol{x})$ in multivariate cases with adaptive partitioning

One can sample from the posterior of $(\mathcal{C}, \mathcal{A}^{(\infty)})$ using generative inductive procedure based on a forward-summation-backward-sampling algorithm similar to those described in Section 2.4. The procedure randomly generates $A^{(\infty)}$ and $\mathcal{C}$ together based on the joint posterior. Starting from $A = \Omega$, the generative procedure grows the partition tree $\mathcal{A}^{(\infty)}$ level by level while at the same time generates the latent state $C(A)$ on each nodade that has been grown on the partition tree. On each node A that has been grown on the tree, the procedure generates $(C(A), J(A))$ by first drawing $C(A)$ based on a state transition given the state on A's parent based on a posterior transition probability matrix $\tilde{\tilde{\boldsymbol{\gamma}}}(A)$, computable by a forward-backward recursion described below.

The forward-backward algorithm is a modified version of the same algorithm in Section 2.4. First, now that $\mathcal{A}^{(\infty)}$ is not fixed but can take one of a range of possible values, for any A that can potentially be a node in $\mathcal{A}^{(\infty)}$, we redefine for $i = 1, 2, \ldots, I$,

$$\xi_A(i, \boldsymbol{\phi}) := \begin{cases} \int q(\boldsymbol{x}|A)\pi(dq \mid \boldsymbol{\phi}, C(A_p) = i, A \in \mathcal{A}^{(\infty)}) & \text{if } A \neq \Omega \\ \int q(\boldsymbol{x}|A)\pi(dq \mid \boldsymbol{\phi}) & \text{if } A = \Omega. \end{cases}$$

Then the forward-summation step is again a bottom-up recursion for computing the $\xi_A(i, \boldsymbol{\phi})$:

$$\xi_A(i, \boldsymbol{\phi}) = \begin{cases} \sum_{i'=1}^{I} \gamma_{i,i'}(A) \cdot \sum_{j=1}^{N(A)} M_A^{i'}(\boldsymbol{\theta}_0, j) \cdot \xi_{A_l^j}(i', \boldsymbol{\phi}) \xi_{A_r^j}(i', \boldsymbol{\phi}) / N(A) & \text{if } n(A) > 1 \\ q_0(\boldsymbol{x}|A) & \text{if } n(A) = 1 \text{ or } A \text{ has no children} \\ 1 & \text{if } n(A) = 0, \end{cases}$$

where again for $A = \Omega$ we simply replace $\gamma_{i,i'}(A)$ with $\gamma_{i'}(\Omega)$.

Then a backward step allows us to compute the posterior transition probability matrix for drawing $C(A)$ given $C(A_p)$ (or a posterior initial state probability vector when $A = \Omega$),

- The initial state probability vector:

$$\tilde{\tilde{\boldsymbol{\gamma}}}(\Omega) = \boldsymbol{\gamma}(\Omega)\boldsymbol{D}''(\Omega)/\xi_\Omega(1, \boldsymbol{\phi})$$

- The state transition probability matrix:

$$\tilde{\tilde{\boldsymbol{\gamma}}}(A) = \boldsymbol{D}'(A)^{-1}\boldsymbol{\gamma}(A)\boldsymbol{D}''(A)$$

where $\boldsymbol{D}'(A)$ is again defined to be the $I \times I$ diagonal matrix with the diagonal elements being $\xi_A(i, \boldsymbol{\phi})$ for $i = 1, 2, \ldots, I$, whereas $\boldsymbol{D}''(A)$ is now redefined to be the $I \times I$ diagonal matrix with the ith diagonal element being $\sum_{j=1}^{N(A)} M_A^i(\boldsymbol{\theta}_0, j)\xi_{A_l^j}(i, \boldsymbol{\phi})\xi_{A_r^j}(i, \boldsymbol{\phi})/N(A)$ if A has children and $q_0(\boldsymbol{x}|A)$ if not for $i = 1, 2, \ldots, I$. For $j = 1, 2, \ldots, N(A)$,

$$M_A^i(\boldsymbol{\theta}_0, j) = \int M_A(\boldsymbol{\theta}_0, \nu, j) dF^i(\nu)$$

where $M_A(\boldsymbol{\theta}_0, \nu, j) = \frac{\Gamma(\theta_0(A)\nu + n(A_l^j))\Gamma((1-\theta_0(A))\nu + n(A_r^j))\Gamma(\nu)}{\Gamma(\nu + n(A))\Gamma(\theta_0(A)\nu)\Gamma((1-\theta_0(A))\nu)}$.

Once $C(A)$ is generated according to the state transition probabilities, say $C(A) = i$, suppose $N(A) > 0$ so A can be divided. Then we randomly draw a way to divide A into children. Specifically, we draw $J(A)$ from $\{1, 2, \ldots, N(A)\}$ based on a multinomial distribution and if $J(A) = j$ then we divide A into A_l^j and A_r^j. Let $\lambda_j(A, i)$ denote the probability for $J(A) = j$. It can also be computed through a backward-sampling step:

$$\lambda_j(A, i) = \begin{cases} \frac{M_A^i(\boldsymbol{\theta}_0, j) \cdot \xi_{A_l^j}(i, \boldsymbol{\phi})\xi_{A_r^j}(i, \boldsymbol{\phi})}{\sum_{j'=1}^{N(A)} M_A^i(\boldsymbol{\theta}_0, j') \cdot \xi_{A_l^{j'}}(i, \boldsymbol{\phi})\xi_{A_r^{j'}}(i, \boldsymbol{\phi})} & \text{if } n(A) > 1 \\ 1/N(A) & \text{if } n(A) \leq 1. \end{cases}$$

After $(C(A), J(A))$ are generated, the procedure then move onto the children A_l^j and A_r^j and repeat itself. This allows us to draw a sample from the marginal posterior $\pi(\mathcal{C}, \mathcal{A}^{(\infty)} \,|\, \boldsymbol{\phi}, \boldsymbol{x})$.

S4. Prior specification of the DPM of normals

We use the `DPpackage` function `DPdensity` to carry out density estimation using the Dirichlet process mixture (DPM) of normals. The model formulation and the hyperparameter values are as follows, which follows an example in the user manual for `DPpackage`

$$\begin{aligned} x_i \,|\, \mu_i, \Sigma_i &\sim \mathrm{N}(\mu_i, \Sigma_i) \quad \text{for } i = 1, 2, \ldots, n \\ (\mu_i, \Sigma_i) \,|\, H &\sim H \\ H \,|\, \alpha, H_0 &\sim \mathrm{DP}(\alpha H_0) \\ H_0 &= \mathrm{N}(\mu | m_1, \Sigma/k_0) \times \mathrm{IW}(\Sigma | \nu_1, \psi_1) \\ \alpha \,|\, a_0, b_0 &\sim \mathrm{Gamma}(a_0, b_0) \\ m_1 \,|\, m_2, s_2 &\sim \mathrm{N}(m_2, s_2) \\ k_0 \,|\, \tau_1, \tau_2 &\sim \mathrm{Gamma}(\tau_1/2, \tau_2/2) \\ \psi_1 \,|\, \nu_2, \psi_2 &\sim \mathrm{IW}(\nu_2, \psi_2) \end{aligned}$$

where $a_0 = 2$, $b_0 = 1$, $m_2 = 0$, $s_2 = 10^5$, $\psi_2 = \mathrm{diagonal}(0.5, 1)$, $\nu_1 = 4$, $\nu_2 = 4$, $\tau_1 = 1$, and $\tau_2 = 100$. We draw 5,000 posterior samples using 1,000 burn-in iterations and a 10-iteration thinning window.

References

Wong, W. H. and Ma, L. (2010). "Optional Pólya tree and Bayesian inference." *Annals of Statistics*, 38(3): 1433–1459.
URL http://projecteuclid.org/euclid.aos/1268056622 2