SUPPLEMENTARY MATERIAL FOR *JUMP-MEANS: SMALL-VARIANCE ASYMPTOTICS FOR MARKOV JUMP PROCESSES*

JONATHAN H. HUGGINS, KARTHIK NARASIMHAN, ARDAVAN SAEEDI, AND VIKASH K. MANSINGHKA

Appendix A. Parametric MJPs for SVA

To obtain the SVA objective from the parametric MJP model, we begin by scaling the exponential distribution $f(t;\lambda) = \lambda\exp(-\lambda t)$, which is an exponential family distribution with natural parameter $\eta = -\lambda$, log-partition function $\psi(\eta) = -\ln(-\eta)$, and base measure $\nu(\mathrm{d}t) = 1$ [1]. To scale the distribution, introduce the new natural parameter $\tilde{\eta} = \beta\eta$ and log-partition function $\tilde{\psi}(\tilde{\eta}) = \beta\psi(\tilde{\eta}/\beta)$. The new base measure $\tilde{\nu}(\mathrm{d}t)$ is uniquely defined by the integral equation [see 1, Theorem 5]

$$\int \exp(\tilde{\eta}t)\tilde{\nu}(\mathrm{d}t) = \exp(\tilde{\psi}(\tilde{\eta})) = \exp(-\beta\ln(\tilde{\eta}/\beta)) = \frac{\beta^\beta}{\tilde{\eta}^\beta}.$$

Choosing $\tilde{\nu}(\mathrm{d}t) = \frac{t^{\beta-1}\beta^\beta}{\Gamma(\beta)}\mathrm{d}t$ satisfies the condition, so we have

$$\begin{aligned} f(t;\lambda,\beta) = \frac{(\beta\lambda)^\beta}{\Gamma(\beta)}t^{\beta-1}e^{-\beta\lambda t} &= \exp(-\beta\lambda t + (\beta-1)\ln t + \beta\ln\lambda\beta - \ln\Gamma(\beta)) \\ &= \exp\left\{-\beta\left(\lambda t - \ln t - \ln\lambda - \frac{\beta\ln\beta - \ln\Gamma(\beta)}{\beta} + \frac{\ln t}{\beta}\right)\right\}. \end{aligned}$$

It can now be seen that $f(t;\lambda,\beta)$ is the density of a gamma distribution with shape parameter β and rate parameter $\beta\lambda$. Hence, the mean of the scaled distribution is $\frac{1}{\lambda}$ and its variance is $\frac{1}{\lambda\beta}$. Letting $F(t;\lambda,\beta)$ denote the CDF corresponding to $f(t;\lambda,\beta)$, we have $1 - F(t;\lambda,\beta) = \frac{\Gamma(\beta,\beta\lambda t)}{\Gamma(\beta)}$, where $\Gamma(\cdot,\cdot)$ is the upper incomplete gamma function.

For the state at the k-th jump we use a 1-of-M representation; that is, s_k is an M-dimensional binary random variable which satisfies $s_{km} \in \{0,1\}$ and $\sum_{m=1}^M s_{km} = 1$. Hence, we have:

$$p(s_k|s_{k-1,j} = 1) = \prod_{m=1}^M p_{jm}^{s_{km}}. \tag{A.1}$$

Given the Bregman divergence for a multinomial distribution, $d_\phi(s_k, \boldsymbol{p}_j) = \mathrm{KL}(s_k||\boldsymbol{p}_j)$ where $\boldsymbol{p}_j \triangleq (p_{j1},\dots,p_{jM})$, this can be written in terms of exponential family notation in the following form [1]:

$$p(s_k|s_{k-1,j} = 1) = b_\phi(s_k)\exp(-d_\phi(s_k,\boldsymbol{p}_j)) \tag{A.2}$$

where $b_\phi(s_k) = 1$. For a scaled multinomial distribution we have $b_{\hat{\beta}\phi}(s_k)\exp(-\hat{\beta}d_\phi(s_k,\boldsymbol{p}_j))$, where $\hat{\beta} = \xi\beta$ is the scaling parameter for the multinomial distribution. Writing the trajectory probility with the scaled exponential families yields:

$$\begin{aligned} p(\mathcal{U}|s_0, s_K, P, \boldsymbol{\lambda}) \propto \exp\Bigg\{-\beta\Bigg(&\frac{\ln\Gamma(\beta) - \ln\Gamma(\beta,\beta\lambda_{s_K}t_.)}{\beta} + \xi\sum_{k=0}^{K-1}\mathrm{KL}(s_{k+1}||\boldsymbol{p}_{s_k}) \\ &+ \sum_{k=0}^{K-1}\left(\lambda_{s_k}t_k - \ln\lambda_{s_k}t_k - \frac{\beta\ln\beta - \ln\Gamma(\beta)}{\beta} + \frac{\ln t_k}{\beta}\right)\Bigg)\Bigg\}, \end{aligned} \tag{A.3}$$

Since $\beta \to \infty$, we can apply the asymptotic expansions for $\Gamma(\cdot)$ and $\Gamma(\cdot,\cdot)$. In particular, applying Stirling's formula and the facts in [2] we have:

$$\frac{\beta \ln \beta - \ln \Gamma(\beta)}{\beta} = \frac{\beta \ln \beta - \beta \ln \beta + \beta + o(\beta)}{\beta} \to 1$$
$$\frac{\ln \Gamma(\beta) - \ln \Gamma(\beta, \beta\lambda t)}{\beta} = \begin{cases} \frac{-\beta - o(\beta) - \beta \ln \lambda t + \beta \lambda t}{\beta} \to \lambda t - \ln \lambda t - 1 & \text{if } t \geq \frac{1}{\lambda} \\ \frac{\beta \ln \beta - \beta - \beta \ln \beta + \beta + o(\beta)}{\beta} \to 0 & \text{if } t < \frac{1}{\lambda} \end{cases}$$

We also place a $\mathsf{Gam}(\alpha_\lambda, \alpha_\lambda \mu_\lambda)$ prior on each λ_i. With $\alpha_\lambda = \xi_\lambda \beta$, we obtain

$$\begin{aligned} \ln p(\lambda_s \,|\, \alpha_\lambda, \alpha_\lambda \mu_\lambda) &= \alpha_\lambda \ln(\alpha_\lambda \mu_\lambda) + (\alpha_\lambda - 1) \ln \lambda_s - \ln \Gamma(\alpha_\lambda) - \alpha_\lambda \mu_\lambda \lambda_s \\ &= \xi_\lambda \beta \ln \lambda_s - \xi_\lambda \mu_\lambda \beta \lambda_s + \xi_\lambda \beta + o(\beta) \\ &= -\beta(\xi_\lambda \mu_\lambda \lambda_s - \xi_\lambda \ln \lambda_s - 1) + o(\beta). \end{aligned}$$

Hence, when $\beta \to \infty$, obtain

$$\begin{aligned} \min_{\mathcal{U}, \boldsymbol{\lambda}, P} \Bigg\{ & \xi \sum_{k=0}^{K-1} \mathrm{KL}(s_{k+1} \| \boldsymbol{p}_{s_k}) + \sum_{k=0}^{K-1} (\lambda_{s_k} t_k - \ln \lambda_{s_k} t_k - 1) \\ & + \mathbb{1}[\lambda_{s_K} t_\cdot \geq 1](\lambda_{s_K} t_\cdot - \ln \lambda_{s_K} t_\cdot - 1) + \xi_\lambda \sum_{s=1}^{M} (\mu_\lambda \lambda_s - \ln \lambda_s - 1) \Bigg\} \end{aligned} \tag{A.4}$$

Appendix B. Bayesian Nonparametric MJPs for SVA

First we recall that the Moran gamma process is a distribution over measures. If $\mu \sim \Gamma\mathrm{P}(H, \gamma)$ is a random measure distributed according to a Moran gamma process with base measure H on the probability space $(\Omega, \mathcal{F})$ and rate parameter γ, then for all measurable partitions of Ω, $(A_1, \dots, A_\ell)$, μ satisfies

$$(\mu(A_1), \dots, \mu(A_\ell)) \sim \mathsf{Gam}(H(A_1), \gamma) \times \cdots \times \mathsf{Gam}(H(A_\ell), \gamma). \tag{B.1}$$

The hierarchical gamma-gamma process (HΓΓP) is defined to be:

$$\mu_0 \sim \Gamma\mathrm{P}(\alpha_0 H_0, \gamma_0) \tag{B.2}$$
$$\mu_i \,|\, \mu_0 \overset{\text{i.i.d.}}{\sim} \Gamma\mathrm{P}(\beta \mu_0, \gamma) \qquad i = 1, 2, \dots \tag{B.3}$$
$$s_k \,|\, \{\mu_i\}_{i=0}^{\infty}, \mathcal{U}_{k-1} \sim \bar{\mu}_{s_{k-1}} \tag{B.4}$$
$$t_k \,|\, \{\mu_i\}_{i=0}^{\infty}, \mathcal{U}_{k-1} \sim \mathsf{Gam}(\beta, \|\mu_{s_{k-1}}\|). \tag{B.5}$$

Consider the gamma-gamma process (ΓΓP), defined by (B.3)-(B.5) (with μ_0 treated as an arbitrary fixed measure). We now show that the ΓΓP retains the key properties of the ΓEP: conjugacy and exchangeability. Let $T_i \triangleq \sum_{j=1}^{k} \mathbb{1}[s_{j-1} = i] t_j$ and $F_i \triangleq \sum_{j=1}^{k} \mathbb{1}[s_{j-1} = i] \delta_{s_j}$ be the sufficient statistics of the observations.

Proposition B.1. *The* ΓΓP *is a conjugate family:* $\mu_i \,|\, \mathcal{U}_k \sim \Gamma\mathrm{P}(\beta \mu_i', \gamma_i')$, *where* $\mu_i' = \mu_0 + F_i$ *and* $\gamma_i' = \gamma + T_i$.

Proof sketch. The proof is analogous to that for Proposition 2 in [4]. The key additional insight is that $X \sim \mathsf{Gam}(\beta a, b)$ and $Y \,|\, X \sim \mathsf{Gam}(\beta, X)$ are conjugate: $X \,|\, Y \sim \mathsf{Gam}(\beta(a+1), b + Y)$. □

In order to give the joint distribution of the times $\mathcal{T} \triangleq \mathcal{T}_K \triangleq (t_1, \dots, t_K)$, we first derive the predictive distribution for the ΓΓP, $(s_{k+1}, t_{k+1}) \,|\, \mathcal{U}_k$. We make use of the following family of densities.

Definition B.2 (Shaped Translated Pareto)**.** Let $\beta > 0, \alpha > 0, \gamma > 0$. A random variable S is *shaped translated Pareto*, denoted $S \sim \mathsf{STP}(\beta, \alpha, \gamma)$, if it has density

$$f(t) = \frac{\gamma^{\alpha\beta}}{B(\beta, \alpha\beta)} \frac{t^{\beta-1}}{(t+\gamma)^{(1+\alpha)\beta}},$$

where $B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ is the beta function.

Proposition B.3. *The predictive distribution of the* ΓΓP *is*

$$(s_{k+1}, t_{k+1}) \,|\, \mathcal{U}_k \sim \bar{\mu}_{s_k}' \times \mathsf{STP}(\beta, \|\bar{\mu}_{s_k}'\|, \gamma_{s_k}'). \tag{B.6}$$

Proof. By Proposition B.1, it suffices to show that if $\mu \sim \Gamma\mathrm{P}(\beta\mu_0, \gamma)$, $s \mid \mu \sim \bar{\mu}$, and $t \mid \mu \sim \mathsf{Gam}(\beta, \|\mu\|)$, then $(s, t) \sim \bar{\mu} \times \mathsf{STP}(\beta, \kappa_0, \gamma)$, where $\kappa_0 \triangleq \|\mu_0\|$. Letting $x = \|\mu\|$, the distribution of t is

$$p(t) = \int_0^\infty p(t \mid x)p(x)\mathrm{d}x = \int_0^\infty \frac{x^\beta t^{\beta-1}e^{-xt}}{\Gamma(\beta)} \frac{\gamma^{\beta\kappa_0}x^{\beta\kappa_0-1}e^{-\gamma x}}{\Gamma(\beta\kappa_0)}\mathrm{d}x$$
$$= \frac{\gamma^{\beta\kappa_0}t^{\beta-1}}{\Gamma(\beta)\Gamma(\beta\kappa_0)} \int_0^\infty x^{\beta(1+\kappa_0)-1}e^{-(\gamma+t)x}\mathrm{d}x = \frac{\gamma^{\beta\kappa_0}t^{\beta-1}}{\Gamma(\beta)\Gamma(\beta\kappa_0)} \frac{\Gamma(\beta(1+\kappa_0))}{(\gamma+t)^{\beta(1+\kappa_0)}}.$$

□

We can now show that the process is exchangeable by exhibiting the joint distribution of waiting times:

Proposition B.4. *Let $\boldsymbol{t}^*_m = (t^*_{m1}, \ldots, t^*_{mK_m})$ be the waiting times following state m. Then $\boldsymbol{t}^*_m$ is an exchangeable sequence with joint distribution*

$$p(\boldsymbol{t}^*_m) = \frac{\Gamma(\beta(\kappa_0 + K_m))}{\Gamma(\beta)^{K_m}} \frac{(\prod_{j=1}^{K_m} t^*_{mj})^{\beta-1}}{(\gamma + \sum_{j=1}^{K_m} \tau_{mj})^{\beta(\kappa_0+K_m)}} \tag{B.7}$$

Proof sketch. Take the product of the predictive distributions of $\tau_{m1}, \ldots, \tau_{mK_m}$. □

The measures $\{\mu_i\}_{i=0}^\infty$ and H_0 can be integrated out of the HΓΓP generative model in a manner analogous to the way in the the Chinese restaurant franchise in obtained from the hierarchical Dirichlet process [5]. However the mass of the measure μ_0 cannot be integrated out. We omit details as they are essentially identical to those in case of the HΓEP [4].

First, we consider the case of integrating out $\{\mu_i\}_{i\geq 0}$. Let M denote the number of used states, K_m the number of transitions out of state m, and r_m the number of states that can be reached from state m in one step. The contribution to the likelihood from the HΓΓP prior is

$$p(\mathcal{U}, \kappa_0 \mid \beta, \gamma_0, \gamma, \alpha_0) = p(\kappa_0 \mid \alpha_0, \gamma_0)p(\mathcal{S} \mid \beta, \alpha_0, \kappa_0)p(\mathcal{T} \mid \beta, \gamma, \kappa_0)$$
$$\propto \kappa_0^{\alpha_0-1}e^{-\gamma_0\kappa_0}\alpha_0^{M-1}\frac{\Gamma(\alpha_0+1)}{\Gamma(\alpha_0+r_\cdot)}\prod_{m=1}^{M}(\beta\kappa_0)^{r_m-1}\frac{\Gamma(\beta\kappa_0+1)}{\Gamma(\beta\kappa_0+K_m)}$$
$$\times \prod_{m=1}^{M}\frac{\Gamma(\beta(\kappa_0+K_m))}{\Gamma(\beta)^{K_m}}\frac{(\prod_{j=1}^{K_m} t^*_{mj})^{\beta-1}}{(\gamma+\sum_{j=1}^{K_m} t^*_{mj})^{\beta(\kappa_0+K_m)}},$$

where $r_\cdot \triangleq \sum_m r_m$. Taking the logarithm, using asymptotic expansions for the Gamma terms, and ignoring $o(\beta)$ terms yields

$$(\alpha_0 - 1)\ln\kappa_0 - \gamma_0\kappa_0 + (M-1)\ln\alpha_0 + \sum_{m=1}^{M}\{(r_m - 1)\ln\kappa_0 + \beta(\kappa_0 + K_m)\ln[\beta(\kappa_0+K_m)]\}$$
$$\sum_{m=1}^{M}\left\{-\beta(\kappa_0+K_m) - K_m[\beta\ln\beta - \beta] + \beta\sum_{j=1}^{K_m}\ln t^*_{mj} - \beta(\kappa_0+K_m)\ln(\gamma + t^*_{m\cdot})\right\},$$

where $t^*_{m\cdot} \triangleq \sum_{j=1}^{K_m} t^*_{mj}$. In order to retain the effects of the hyperparameters in the asymptotics, set $\alpha_0 = \exp(-\xi_1\beta)$ and $\gamma_0 = \exp(\xi_2\beta)$. Thus, $\kappa_0 \to 0$ as $\beta \to \infty$. We require that $\limsup_{\beta\to\infty}\kappa_0\gamma_0 < \infty$, so without loss of generality we can choose $\kappa_0 = \gamma_0^{-1} = \exp(-\xi_2\beta)$ to obtain

$$-\beta\left(\xi_1(M-1) + \sum_{m=1}^{M}\left\{\xi_2(r_m-1) - \sum_{j=1}^{K_m}\ln t^*_{mj} + K_m\ln([\gamma + t^*_{m\cdot}]/K_m)\right\}\right).$$

Thus, the objective function to minimize is

$$\zeta\sum_{\ell=1}^{L}\mathrm{KL}(x_\ell \| \boldsymbol{\rho}_{s_{\tau_\ell}}) + \xi_1 M + \sum_{m=1}^{M}\left\{\xi_2(r_m-1) - \sum_{j=1}^{K_m}\ln t^*_{mj} - K_m\ln([\gamma + t^*_{m\cdot}]/K_m)\right\}. \tag{B.8}$$

Alternatively, the small variance asymptotics can be derived in the case where $\{\mu_i\}_{i\geq 0}$ is not integrated out. To do so, we first rewrite the HΓΓP generative model in an equivalent form, with H_0 integrated out:

$$\pi_0 \sim \mathsf{GEM}(\alpha_0) \tag{B.9}$$

$$\kappa_0 \sim \mathsf{Gam}(\alpha_0, \gamma_0) \tag{B.10}$$

$$\pi_i \,|\, \pi_0 \overset{\text{i.i.d.}}{\sim} \mathcal{DP}(\beta\kappa_0\pi_0), \qquad i = 1, 2, \ldots \tag{B.11}$$

$$\kappa_i \,|\, \pi_0 \overset{\text{i.i.d.}}{\sim} \mathsf{Gam}(\beta, \gamma), \qquad i = 1, 2, \ldots \tag{B.12}$$

$$s_k \,|\, \{\pi_i\}_{i=1}^{\infty}, \mathcal{U}_{k-1} \sim \pi_{s_{k-1}} \tag{B.13}$$

$$t_k \,|\, \{\kappa_i\}_{i=1}^{\infty}, \mathcal{U}_{k-1} \sim \mathsf{Gam}(\beta, \kappa_{s_k}). \tag{B.14}$$

For $0 \leq i \leq M, 1 \leq j \leq M$, let $\bar{\pi}_{i,j} \triangleq \pi_{ij}$ and for $0 \leq i \leq M$, let $\bar{\pi}_{i,M+1} \triangleq 1 - \sum_{j=1}^{M} \pi_{ij}$. Integrating out $\{\kappa_i\}_{i\geq 1}$, the contribution to the likelihood from the HΓΓP prior is now

$$p(\mathcal{U}_K, \kappa_0, \bar{\pi} \,|\, \beta, \gamma_0, \gamma, \alpha_0) \tag{B.15}$$

$$= p(\kappa_0 \,|\, \alpha_0, \gamma_0)p(\bar{\pi}_0 \,|\, \alpha_0)p(\bar{\pi}_{1:M} \,|\, \beta\kappa_0\bar{\pi}_0)p(\mathcal{S}_K \,|\, \bar{\pi}_{1:M})p(\mathcal{T}_K \,|\, \beta, \gamma, \kappa_0) \tag{B.16}$$

$$\propto \kappa_0^{\alpha_0 - 1} e^{-\gamma_0\kappa_0} \prod_{i=1}^{M} \mathsf{Beta}\left(\frac{\bar{\pi}_{0i}}{1 - \sum_{j=1}^{i-1}\pi_{0,j}}\middle| 1, \alpha_0\right) \mathsf{Dir}(\bar{\pi}_i \,|\, \beta\kappa_0\bar{\pi}_0) \left(\prod_{k=1}^{K} \bar{\pi}_{s_{k-1},s_k}\right) p(\mathcal{T}_K \,|\, \beta, \gamma, \kappa_0) \tag{B.17}$$

$$\begin{aligned}
&\propto \kappa_0^{\alpha_0 - 1} e^{-\gamma_0\kappa_0} \prod_{i=1}^{M} \left\{ \frac{\Gamma(1+\alpha_0)}{\Gamma(\alpha_0)} \left(\frac{1 - \sum_{j=1}^{i}\pi_{0,j}}{1 - \sum_{j=1}^{i-1}\pi_{0,j}}\right)^{\alpha_0 - 1} \Gamma(\beta\kappa_0) \prod_{j=1}^{M+1} \frac{\bar{\pi}_{ij}^{\beta\kappa_0\bar{\pi}_{0j} - 1}}{\Gamma(\beta\kappa_0\bar{\pi}_{0j})} \right\} \\
&\quad \times \prod_{k=1}^{K} \bar{\pi}_{s_{k-1},s_k}^{\beta\xi} \times \prod_{m=1}^{M} \frac{\Gamma(\beta(\kappa_0 + K_m))}{\Gamma(\beta)^{K_m}} \frac{(\prod_{j=1}^{K_m} t_{mj}^*)^{\beta - 1}}{(\gamma + \sum_{j=1}^{K_m} t_{mj}^*)^{\beta(\kappa_0 + K_m)}}.
\end{aligned} \tag{B.18}$$

We use a slightly different limiting process, with $\gamma_0 = \kappa_0 = \xi_2$, a positive constant, and scale the multinomial distributions (B.13) by $\beta\xi$. Taking the logarithm and and ignoring $o(\beta)$ terms as before yields

$$\begin{aligned}
&\sum_{i=1}^{M} \left\{ \ln\alpha_0 + \beta\kappa_0 \ln\beta\kappa_0 - \beta + \sum_{j=1}^{M+1} \{-\beta\kappa_0\bar{\pi}_{0,j}\ln(\beta\kappa_0\bar{\pi}_{0,j}) + \beta\kappa_0\bar{\pi}_{0,j} + \beta\kappa_0\bar{\pi}_{0,j}\ln\bar{\pi}_{ij}\} \right\} \\
&\quad + \sum_{k=1}^{K} \beta\xi \ln\bar{\pi}_{s_{k-1},s_k} + \sum_{m=1}^{M} \left\{ \textstyle\sum_{j=1}^{K_m} \beta \ln t_{mj}^* - \beta K_m \ln\left([\gamma + t_{m\cdot}^*]/K_m\right) \right\} \\
&\sim \sum_{i=1}^{M} \left\{ -\beta\xi_1 + \sum_{j=1}^{M+1} \{-\beta\kappa_0\bar{\pi}_{0,j}\ln(\bar{\pi}_{0,j}) + \beta\kappa_0\bar{\pi}_{0,j}\ln\bar{\pi}_{ij}\} \right\} \\
&\quad + \sum_{k=1}^{K} \beta\xi \ln\bar{\pi}_{s_{k-1},s_k} + \sum_{m=1}^{M} \left\{ \textstyle\sum_{j=1}^{K_m} \beta \ln t_{mj}^* - \beta K_m \ln\left([\gamma + t_{m\cdot}^*]/K_m\right) \right\} \\
&\sim -\beta \left\{ \xi_1 M + \xi \sum_{k=1}^{K} \ln\bar{\pi}_{s_{k-1},s_k} + \sum_{m=1}^{M} \left\{ \xi_2 \mathrm{KL}(\bar{\pi}_0 || \bar{\pi}_m) - \textstyle\sum_{j=1}^{K_m} \ln t_{mj}^* - K_m \ln\left([\gamma + t_{m\cdot}^*]/K_m\right) \right\} \right\}.
\end{aligned}$$

Thus, the objective function to minimize is

$$\begin{aligned}
&\zeta \sum_{\ell=1}^{L} \ln\rho_{s_{\tau_\ell} x_\ell} + \xi \sum_{k=1}^{K} \ln\bar{\pi}_{s_{k-1},s_k} + \xi_1 M \\
&+ \sum_{m=1}^{M} \left\{ \xi_2 \mathrm{KL}(\bar{\pi}_0 || \bar{\pi}_m) - \textstyle\sum_{j=1}^{K_m} \ln t_{mj}^* - K_m \ln\left([\gamma + t_{m\cdot}^*]/K_m\right) \right\}.
\end{aligned} \tag{B.19}$$

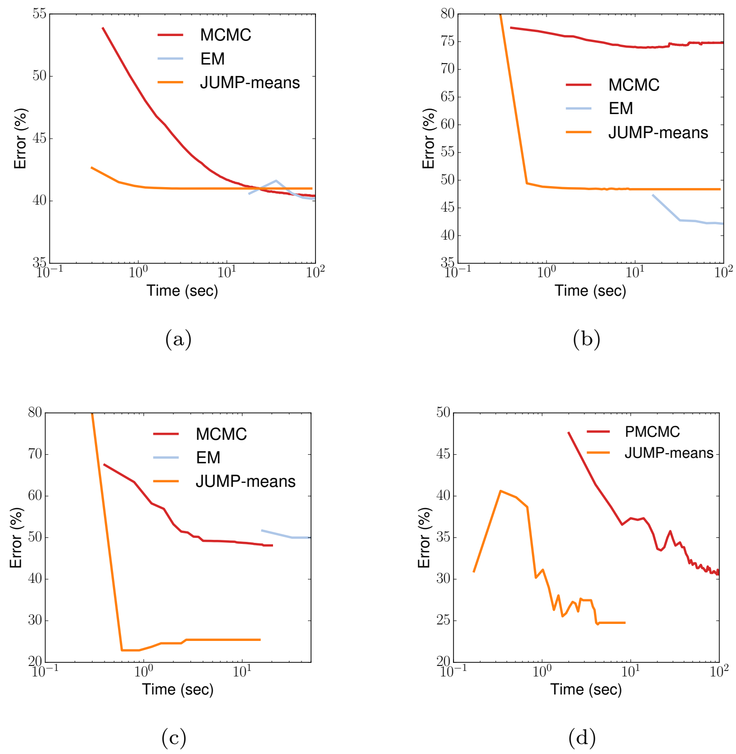

FIGURE C.1. Mean error vs CPU runtime for **(a)** Synthetic 1; **(b)** Synthetic 2; **(c)** MS; and **(d)** MIMIC datasets. In each case the JUMP-means algorithms have better or comparable performance to other standard methods of inference in MJPs.

APPENDIX C. TIME-ACCURACY PLOTS FOR THE EXPERIMENTS

In the main paper we include the error versus iteration as it is more objective than time-accuracy results. In Fig. C.1, we compare the time-accuracy across different methods for different datasets. EM, PMCMC, and JUMP-means are implemented in Java and MCMC is implemented in Python. To plot the MCMC results, we give a speed boost of 100x in the results to compensate for Python's slow interpreter. From our experience with scientific computing applications, we believe this is a generous adjustment. Also we note that the EM implementation used in our experiments is not the most optimized in terms of time per iteration. However, our goal is to show that JUMP-means can achieve comparable performance with a reasonable implementation of MCMC and EM.

APPENDIX D. SCALING EXPERIMENTS

For the scaling experiments we generated 4 datasets consisting of 10^2 to 10^5 sequences. All datasets are sampled from a single hidden state MJP with 5 hidden states and 5 possible observations. For the 20 observations in each sequence a Gaussian likelihood is used. Finally, for the held out results, we categorized the observations in 5 bins, removed 30% of the data points and predicted their category.

References.

[1] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh. "Clustering with Bregman divergences". In: *The Journal of Machine Learning Research* 6 (2005), pp. 1705–1749.
[2] *NIST Digital Library of Mathematical Functions*. http://dlmf.nist.gov/, Release 1.0.8 of 2014-04-25. Online companion to [3]. URL: `http://dlmf.nist.gov/`.
[3] F. W. J. Olver, D. W. Lozier, R. F. Boisvert, and C. W. Clark, eds. *NIST Handbook of Mathematical Functions*. Print companion to [2]. New York, NY: Cambridge University Press, 2010.
[4] A. Saeedi and A. Bouchard-Côté. "Priors over Recurrent Continuous Time Processes." In: *NIPS*. Vol. 24. 2011, pp. 2052–2060.

[5] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. "Hierarchical Dirichlet Processes". In: *Journal of the American Statistical Association* 101.476 (Dec. 2006), pp. 1566–1581.

Computer Science and Artificial Intelligence Laboratory, MIT
E-mail address: `jhuggins@mit.edu`

E-mail address: `karthikn@mit.edu`

E-mail address: `ardavans@mit.edu`

E-mail address: `vkm@mit.edu`