Supplementary Material to: Clustered factor analysis of multineuronal spike data

Lars Buesing[1], Timothy A. Machado[1,2], John P. Cunningham[1] and Liam Paninski[1]
[1] Department of Statistics, Center for Theoretical Neuroscience
& Grossman Center for the Statistics of Mind
[2] Howard Hughes Medical Institute & Department of Neuroscience
Columbia University, New York, NY
{lars,cunningham,liam}@stat.columbia.edu

Here we provide details of the variational inference method for the mixPLDS model. To this end, we first discuss variational inference for the case of a single mixture component $M = 1$, a model that is equivalent to the Poisson linear dynamical system (PLDS) model defined in Macke et al. (2011).

1 Variational inference for Poisson linear dynamical system

1.1 Notation

We first introduce the "vectorized" notation for the PLDS model. The PLDS is equivalent to the mixPLDS model for $M = 1$. We therefore drop the group index m when focussing on the PLDS.

$$\mathbf{x} := \begin{pmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_T \end{pmatrix}, \qquad \mathbf{y} := \begin{pmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_T \end{pmatrix}, \qquad \overline{\mathbf{b}} := \left.\begin{pmatrix} \mathbf{b} \\ \vdots \\ \mathbf{b} \end{pmatrix}\right\} T-\text{times} \tag{1}$$

$$W = \text{block-diag}(\underbrace{C, \ldots, C}_{T\text{-times}}) \tag{2}$$

$$\eta := W\mathbf{x} + \overline{\mathbf{b}} \tag{3}$$

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\mu, \Sigma) \tag{4}$$

$$p(\mathbf{y}|\mathbf{x}) = \prod_{n=1}^{KT} p(y_n|\eta_n) \tag{5}$$

$$p(y_n|\eta_n) = \text{Poisson}(y_n|\exp(\eta_n)), \tag{6}$$

where the index $n = 1, \ldots, KT$ runs over all observations, i.e. over all observed neurons $k = 1, \ldots, K$ for all time steps $t = 1, \ldots, T$. Slightly overloading the notation, we denote the corresponding observation as y_n for all $n = 1, \ldots, KT$. The precision $\Lambda := \Sigma^{-1}$ of the LDS prior is block-tri-diagonal:

$$\Lambda = \Sigma^{-1} = \begin{pmatrix} Q_0^{-1} + A^\top Q^{-1} A & -A^\top Q^{-1} & & \\ -Q^{-1}A & Q^{-1} + A^\top Q^{-1} A & -A^\top Q^{-1} & \\ & \ddots & \ddots & \ddots \end{pmatrix} \tag{7}$$

The prior mean is given by:

$$\mu = \begin{pmatrix} \mu_1 \\ A\mu_1 \\ \vdots \\ A^{T-1}\mu_1 \end{pmatrix}. \tag{8}$$

1.2 Gaussian variational inference

We make the following Gaussian approximation to the posterior :

$$p(\mathbf{x}|\mathbf{y}) \approx q(\mathbf{x}) \quad = \quad \mathcal{N}(\mathbf{x}|\mathbf{m}, V). \tag{9}$$

The variational lower bound reads:

$$\begin{aligned}
\mathcal{L}(\mathbf{m}, V) &\leq \log p(\mathbf{y}) && (10)\\
\mathcal{L}(\mathbf{m}, V) &= \frac{1}{2}\left(\log|V| - \mathrm{tr}[\Sigma^{-1}V] - (\mathbf{m}-\mu)^\top \Sigma^{-1}(\mathbf{m}-\mu)\right) + \sum_n \mathbb{E}_{q(\mathbf{x})}[\log p(y_n|\eta_n)] && (11)\\
&\quad \underbrace{-\sum_n \log(y_n!) - \frac{1}{2}\log|\Sigma| + \frac{dT}{2}}_{\text{constant in } \mathbf{m}, V}. && (12)
\end{aligned}$$

For Poisson observations with exponential link function we can compute $\mathbb{E}_{q(\mathbf{x})}[\log p(y_n|\eta_n)]$:

$$\begin{aligned}
\mathbb{E}_{q(\mathbf{x})}[\log p(y_n|\eta_n)] &=: -f_n(h_n, \rho_n) && (13)\\
f_n(h_n, \rho_n) &= -y_n h_n + \exp(h_n + \rho_n/2) && (14)\\
\mathbf{h} &:= W\mathbf{m} + \overline{\mathbf{b}} && (15)\\
\rho &:= \mathrm{diag}(WVW^\top). && (16)
\end{aligned}$$

The bound then reads (ignoring additive constants):

$$\mathcal{L}(\mathbf{m}, V) \quad = \quad \frac{1}{2}\left(\log|V| - \mathrm{tr}[\Sigma^{-1}V] - (\mathbf{m}-\mu)^\top \Sigma^{-1}(\mathbf{m}-\mu)\right) - \sum_n f_n(h_n, \rho_n). \tag{17}$$

Variational inference can now be cast as optimizing this lower bound over the variational parameters $\mathbf{m}, V$:

$$\begin{aligned}
\max_{\mathbf{m}, V} \quad & \mathcal{L}(\mathbf{m}, V) && (18)\\
\text{subject to} \quad & V \succeq 0.
\end{aligned}$$

1.3 Variational inference via dual optimization

As shown in Emtiyaz Khan et al. (2013), instead of optimizing the original problem (18), we can solve following dual problem:

$$\begin{aligned}
\min_{\lambda} \quad & D(\lambda) && (19)\\
\text{subject to} \quad & \lambda > 0,
\end{aligned}$$

where $\lambda \in \mathbb{R}^{KT}$ and $\lambda > 0$ denotes the element-wise positivity constraints $\forall n \; \lambda_n > 0$. The dual cost function is given by:

$$\begin{aligned}
D(\lambda) &:= \frac{1}{2}(\lambda - \mathbf{y})^\top W\Sigma W^\top(\lambda - \mathbf{y}) - (W\mu + \overline{\mathbf{b}})^\top(\lambda - \mathbf{y}) - \frac{1}{2}\log|A_\lambda| + \sum_n f^*(\lambda_n) && (20)\\
f^*(\lambda_n) &:= \lambda_n(\log\lambda_n - 1) && (21)\\
A_\lambda &:= \Sigma^{-1} + W^\top \mathrm{diag}(\lambda) W. && (22)
\end{aligned}$$

The dual optimization problem is strictly convex. Given the optimal value λ^*, we can express the optimal variational parameters for $q(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\mathbf{m}^*, V^*)$ as:

$$\begin{aligned}
\mathbf{m}^* &= \mu - \Sigma W^\top(\lambda^* - \mathbf{y}) && (23)\\
V^* &= (\Sigma^{-1} + W^\top \mathrm{diag}(\lambda^*) W)^{-1} = A_{\lambda^*}^{-1}. && (24)
\end{aligned}$$

The variational lower bound at the optimum $\mathbf{m}^*, V^*$ reads:

$$\begin{aligned}
\mathcal{L}^* &= D(\lambda^*) - \sum_n \log y_n! - \frac{1}{2}\log|\Sigma| && (25)\\
&= -\frac{1}{2}\log|A_\lambda| + \frac{1}{2}\lambda^{*\top}\mathrm{diag}(WA_\lambda^{-1}W^\top) - \frac{1}{2}(\lambda - \mathbf{y})^\top W\Sigma W^\top(\lambda - \mathbf{y}) - \sum_n f_n(h_n^*, \rho_n^*) && (26)\\
&\quad - \sum_n \log y_n! - \frac{1}{2}\log|\Sigma|, && (27)
\end{aligned}$$

where $\mathbf{h}^* = W\mathbf{m}^* + \mathbf{b}$ and $\rho^* = \text{diag}(WV^*W^\top)$. The gradient of the dual reads:

$$\nabla_\lambda \;=\; W\Sigma W^\top(\lambda - \mathbf{y}) - W\mu - \overline{\mathbf{b}} + \log\lambda - \frac{1}{2}\,\text{diag}(WA_\lambda^{-1}W^\top).$$

Evaluating the dual function D and its gradient ∇_λ requires computing all T blocks of size $d \times d$ on the diagonal of A_λ. This is equivalent to Kalman smoothing and requires a forward-backward pass through the data which costs $O(Td^3)$ operations.

2 Variational inference for mixPLDS model

The observation model of the mixPLDS is a mixture of Poisson distributions:

$$\log p(y_{kt}|\mathbf{x}_t, s_k) \;=\; \sum_{m=1}^{M} \delta(s_k, m)\left(y_{kt}(C^m_{k:}\mathbf{x}^m_t + b_k) - \exp(C^m_{k:}\mathbf{x}^m_t + b_k)\right) + \text{const}, \tag{28}$$

where δ denotes Kronecker's delta. We do joint inference over the latent variables $\mathbf{x}$ and the cluster assignments $\mathbf{s}$. We make the following factorized variational approximation:

$$p(\mathbf{x}, \mathbf{s}|\mathbf{y}) \;\approx\; q(\mathbf{x})q(\mathbf{s}). \tag{29}$$

The variational lower bound for the mixPLDS reads:

$$\mathcal{L}(\mathbf{m}, V, \phi) \;=\; \frac{1}{2}\left(\log|V| - \text{tr}[\Sigma^{-1}V] - (\mathbf{m}-\mu)^\top\Sigma^{-1}(\mathbf{m}-\mu)\right) \tag{30}$$

$$- \sum_{m=1}^{M}\sum_{k=1}^{K}\sum_{t=1}^{T} \pi^m_k f_{kt}(h^m_{kt}, \rho^m_{kt}) + \sum_{k=1}^{K} D_{KL}[q(s_k)\|p(s_k)] \tag{31}$$

where ϕ are the variational parameters of $q(\mathbf{s})$. Here we used the following notation:

$$C \;:=\; \begin{pmatrix} \tilde{C}^1 \\ \vdots \\ \tilde{C}^M \end{pmatrix} \tag{32}$$

$$W \;:=\; \text{blk-diag}(\underbrace{C, \ldots, C}_{T\text{-times}}) \tag{33}$$

$$\mathbf{h}^m_t \;:=\; \tilde{C}^m\mathbf{m}_t + \mathbf{b} \tag{34}$$

$$\rho^m_t \;:=\; \text{diag}(\tilde{C}^m V_t (\tilde{C}^m)^\top) \tag{35}$$

$$\pi^m_k \;:=\; \mathbb{E}_{q(s_k)}[\delta(s_k, m)] \propto \exp(\phi^m_k). \tag{36}$$

In the equations above, we introduced the matrices $\tilde{C}^m \in \mathbb{R}^{K\times d}$, which are formed by taking the matrices $C^m \in \mathbb{R}^{K\times d^m}$ and adding columns of 0s corresponding to the latent dimensions which are not part system m. Furthermore $V_t \in \mathbb{R}^{d\times d}$ is the t-th $d \times d$ block on the diagonal of V or equivalently $V_t = \text{Cov}_{q(\mathbf{x})}[\mathbf{x}_t]$.

For full variational inference over $\mathbf{x}, \mathbf{s}$ we iterate updates of $q(\mathbf{x})$ and $q(\mathbf{s})$. We observed empirically that this converges very quickly, often in 2-3 iterations to very high precision. Below, we give details for the individual updates.

2.1 Update of $q(\mathbf{x})$

A simple derivation shows that we can do the update of $q(\mathbf{x})$ by solving the following dual problem

$$\min_\lambda \quad D(\lambda) \qquad \text{subject to} \quad \lambda > 0, \tag{37}$$

where

$$D(\lambda) \;:=\; \frac{1}{2}(\lambda-\psi)^\top W\Sigma W^\top(\lambda-\psi) - (W\mu + \overline{\mathbf{b}})^\top(\lambda-\psi) - \frac{1}{2}\log|A_\lambda| \tag{38}$$

$$+ \sum_{m,k,t} \pi^m_k f^*\left(\frac{\lambda^m_{kt}}{\pi^m_k}\right) \tag{39}$$

$$\lambda_t^m := \begin{pmatrix} \lambda_{1t}^m \\ \vdots \\ \lambda_{Kt}^m \end{pmatrix}, \quad \lambda_t := \begin{pmatrix} \lambda_t^1 \\ \vdots \\ \lambda_t^M \end{pmatrix}, \quad \lambda := \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_T \end{pmatrix} \tag{40}$$

$$\psi_{kt}^m := \pi_k^m y_{kt}, \quad \psi_t^m := \begin{pmatrix} \psi_{1t}^m \\ \vdots \\ \psi_{Kt}^m \end{pmatrix}, \quad \psi_t := \begin{pmatrix} \psi_t^1 \\ \vdots \\ \psi_t^M \end{pmatrix}, \quad \psi := \begin{pmatrix} \psi_1 \\ \vdots \\ \psi_T \end{pmatrix} \tag{41}$$

$$\overline{\mathbf{b}} = (\underbrace{\mathbf{b}^\top, \ldots, \mathbf{b}^\top}_{MT\text{-times}})^\top. \tag{42}$$

Hence, the dual variational inference step for a mixPLDS corresponds to the one for a normal PLDS with $M \cdot T \cdot K$ "pseudo-observations" $\psi_{kt}^m = \pi_k^m y_{kt}$.

2.2 Update of $q(\mathbf{s})$

It is straightforward to see that $q(\mathbf{s})$ factorizes further due to the independence assumption of $s_1, \ldots, s_K$ under the prior:

$$q(\mathbf{s}) = \prod_{k=1}^{K} q(s_k) \tag{43}$$

$$\log q(s_k) = \sum_{m=1}^{M} \delta(s_k, m)\phi_k^m + \text{const.} \tag{44}$$

The updates for the variational parameters are given by:

$$\phi_k^m = \phi_0^m - \sum_{t=1}^{T} f_{kt}(h_{kt}^m, \rho_{kt}^m), \tag{45}$$

where ϕ_0^m are the parameters of the prior $p(s_k)$:

$$\log p(s_k) = \sum_{m=1}^{M} \delta(s_k, m)\phi_0^m. \tag{46}$$

References

M. Emtiyaz Khan, A. Aravkin, M. Friedlander, and M. Seeger. Fast dual variational inference for non-conjugate latent gaussian models. In *Proceedings of The 30th International Conference on Machine Learning*, pages 951–959, 2013.

J. H. Macke, L. Buesing, J. P. Cunningham, M. Y. Byron, K. V. Shenoy, and M. Sahani. Empirical models of spiking in neural populations. In *NIPS*, pages 1350–1358, 2011.