

# *Supplementary material to:* Clustered factor analysis of multineuronal spike data

Lars Buesing<sup>1</sup>, Timothy A. Machado<sup>1,2</sup>, John P. Cunningham<sup>1</sup> and Liam Paninski<sup>1</sup>

<sup>1</sup> Department of Statistics, Center for Theoretical Neuroscience  
& Grossman Center for the Statistics of Mind

<sup>2</sup> 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.

$$\mathbf{x} := \begin{pmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_T \end{pmatrix}, \quad \mathbf{y} := \begin{pmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_T \end{pmatrix}, \quad \bar{\mathbf{b}} := \begin{pmatrix} \mathbf{b} \\ \vdots \\ \mathbf{b} \end{pmatrix} \Bigg\} T\text{-times} \quad (1)$$

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

$$\boldsymbol{\eta} := W\mathbf{x} + \bar{\mathbf{b}} \quad (3)$$

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

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

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

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

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

The prior mean is given by:

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

## 1.2 Gaussian variational inference

We make the following Gaussian approximation to the posterior :

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

The variational lower bound reads:

$$\mathcal{L}(\mathbf{m}, V) \leq \log p(\mathbf{y}) \quad (10)$$

$$\mathcal{L}(\mathbf{m}, V) = \frac{1}{2} (\log |V| - \text{tr}[\Sigma^{-1}V] - (\mathbf{m} - \mu)^\top \Sigma^{-1}(\mathbf{m} - \mu)) + \sum_n \mathbb{E}_{q(\mathbf{x})}[\log p(y_n|\eta_n)] \quad (11)$$

$$\underbrace{- \sum_n \log(y_n!) - \frac{1}{2} \log |\Sigma| + \frac{dT}{2}}_{\text{constant in } \mathbf{m}, V}. \quad (12)$$

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

$$\mathbb{E}_{q(\mathbf{x})}[\log p(y_n|\eta_n)] =: -f_n(h_n, \rho_n) \quad (13)$$

$$f_n(h_n, \rho_n) = -y_n h_n + \exp(h_n + \rho_n/2) \quad (14)$$

$$\mathbf{h} := W\mathbf{m} + \bar{\mathbf{b}} \quad (15)$$

$$\rho := \text{diag}(WVW^\top). \quad (16)$$

The bound then reads (ignoring additive constants):

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

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

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

## 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:

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

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:

$$D(\lambda) := \frac{1}{2}(\lambda - \mathbf{y})^\top W \Sigma W^\top (\lambda - \mathbf{y}) - (W\mu + \bar{\mathbf{b}})^\top (\lambda - \mathbf{y}) - \frac{1}{2} \log |A_\lambda| + \sum_n f^*(\lambda_n) \quad (20)$$

$$f^*(\lambda_n) := \lambda_n (\log \lambda_n - 1) \quad (21)$$

$$A_\lambda := \Sigma^{-1} + W^\top \text{diag}(\lambda) W. \quad (22)$$

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

$$\mathbf{m}^* = \mu - \Sigma W^\top (\lambda^* - \mathbf{y}) \quad (23)$$

$$V^* = (\Sigma^{-1} + W^\top \text{diag}(\lambda^*) W)^{-1} = A_{\lambda^*}^{-1}. \quad (24)$$

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

$$\mathcal{L}^* = D(\lambda^*) - \sum_n \log y_n! - \frac{1}{2} \log |\Sigma| \quad (25)$$

$$= -\frac{1}{2} \log |A_\lambda| + \frac{1}{2} \lambda^{*\top} \text{diag}(W A_\lambda^{-1} W^\top) - \frac{1}{2} (\lambda - \mathbf{y})^\top W \Sigma W^\top (\lambda - \mathbf{y}) - \sum_n f_n(h_n^*, \rho_n^*) \quad (26)$$

$$- \sum_n \log y_n! - \frac{1}{2} \log |\Sigma|, \quad (27)$$

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

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

Evaluating the dual function  $D$  and its gradient  $\nabla_\lambda$  requires computing all  $T$  blocks of size  $d \times d$  on the diagonal of  $A_\lambda$ . 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:

$$\log p(y_{kt}|\mathbf{x}_t, s_k) = \sum_{m=1}^M \delta(s_k, m) (y_{kt}(C_{k:}^m \mathbf{x}_t^m + b_k) - \exp(C_{k:}^m \mathbf{x}_t^m + b_k)) + \text{const}, \quad (28)$$

where  $\delta$  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:

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

The variational lower bound for the mixPLDS reads:

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

$$- \sum_{m=1}^M \sum_{k=1}^K \sum_{t=1}^T \pi_k^m f_{kt}(h_{kt}^m, \rho_{kt}^m) + \sum_{k=1}^K D_{KL}[q(s_k)||p(s_k)] \quad (31)$$

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

$$C := \begin{pmatrix} \tilde{C}^1 \\ \vdots \\ \tilde{C}^M \end{pmatrix} \quad (32)$$

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

$$\mathbf{h}_t^m := \tilde{C}^m \mathbf{m}_t + \mathbf{b} \quad (34)$$

$$\rho_t^m := \text{diag}(\tilde{C}^m V_t (\tilde{C}^m)^\top) \quad (35)$$

$$\pi_k^m := \mathbb{E}_{q(s_k)}[\delta(s_k, m)] \propto \exp(\phi_k^m). \quad (36)$$

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

$$\begin{aligned} \min_{\lambda} \quad & D(\lambda) \\ \text{subject to} \quad & \lambda > 0, \end{aligned} \quad (37)$$

where

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

$$+ \sum_{m,k,t} \pi_k^m f^* \left( \frac{\lambda_{kt}^m}{\pi_k^m} \right) \quad (39)$$

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

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

$$\bar{\mathbf{b}} = \underbrace{(\mathbf{b}^\top, \dots, \mathbf{b}^\top)^\top}_{MT\text{-times}}. \quad (42)$$

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  $s_1, \dots, s_K$  under the prior:

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

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

The updates for the variational parameters are given by:

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

where  $\phi_0^m$  are the parameters of the prior  $p(s_k)$ :

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

## 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.