Probabilistic Tensor Decomposition of Neural Population Spiking Activity

Hugo Soulat
Gatsby Unit
University College London
London, W1T 4JG
hugos@gatsby.ucl.ac.uk

Sepiedeh Keshavarzi
Sainsbury Wellcome Centre
University College London
London, W1T 4JG
s.keshavarzi@ucl.ac.uk

Troy W. Margrie
Sainsbury Wellcome Centre
University College London
London, W1T 4JG
t.margrie@ucl.ac.uk

Maneesh Sahani
Gatsby Unit
University College London
London, W1T 4JG
maneesh@gatsby.ucl.ac.uk

Abstract

The firing of neural populations is coordinated across cells, in time, and across experimental conditions or repeated experimental trials, and so a full understanding of the computational significance of neural responses must be based on a separation of these different contributions to structured activity. Tensor decomposition is an approach to untangling the influence of multiple factors in data that is common in many fields. However, despite some recent interest in neuroscience, wider applicability of the approach is hampered by the lack of a full probabilistic treatment allowing principled inference of a decomposition from non-Gaussian spike-count data. Here, we extend the Pólya-Gamma (PG) augmentation, previously used in sampling-based Bayesian inference, to implement scalable variational inference in non-conjugate spike-count models. Using this new approach, we develop techniques related to automatic relevance determination to infer the most appropriate tensor rank, as well as to incorporate priors based on known brain anatomy such as the segregation of cell response properties by brain area. We apply the model to neural recordings taken under conditions of visual-vestibular sensory integration, revealing how the encoding of self- and visual-motion signals is modulated by the sensory information available to the animal.

1 Introduction

Large-scale neural population recording offers a unique window through which to study brain computations that involve the coordinated action of many neurons. Although such rich data sets are increasingly common, the information they contain can only be put to full use once we are able to separate the underlying factors that contribute to the neural activity in a simple and interpretable way. Rich data sets often fit naturally within a multidimensional array or tensor, and a variety of tensor decomposition techniques are available to identify factorial contributions that shape such data. But whereas tensor decompositions have long been used in fields such as chemometrics [1] or computer vision [2], until recently their use in neuroscience had been limited to the study of continuous traces like EEG [3], fMRI signals [4] or LFP [5]. Encouragingly, in the last few years, tensor methods have proven useful to segregate the influence of time dynamics, trial to trial variability [6, 7] or

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

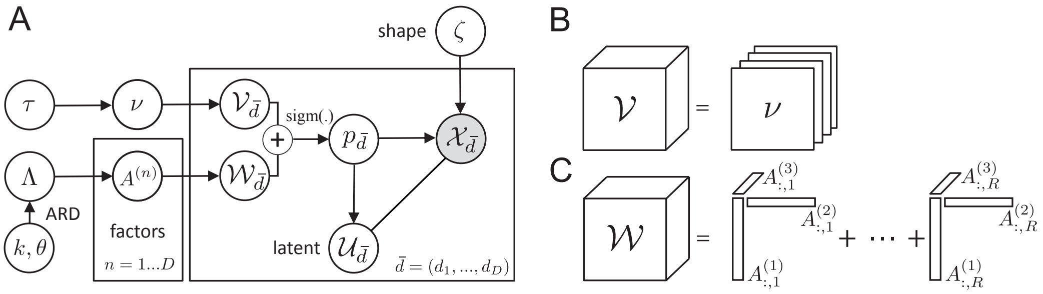

Figure 1: Probabilistic tensor decomposition model. (A) Graphical model. (B) The Offset tensor $\mathcal{V}$ is constrained to vary along a limited set of dimensions. (C) $\mathcal{W}$ is a low rank tensor.

experimental condition [8] in neural spike-count data, even though the models employed have not yet provided a full probabilistic treatment appropriate to counts.

Canonical Polyadic (CP) decomposition [9] is a widely used technique in which a tensor is decomposed into a sum of rank-1 elements. When applied to noisy data, it is usually cast as a least-squares problem and therefore implicitly assumes that noise is normally distributed. Spike counts violate this assumption but alternated-gradient-based optimization methods can be adapted to non-Gaussian likelihood functions. Hong et al. [10] introduced a general framework for Generalized CP (GCP) decomposition that can be applied to Poisson or negative binomial (NB) distributed datasets but it has not been used in neuroscience to our knowledge. Moreover, practical methods have largely focused on point estimates of the decomposition rather than a full probabilistic treatment. As such, they do not provide principled ways to incorporate prior knowledge about the data (such as the (co)location of recorded neurons in the brain), to automatically determine the rank of the observed tensor (although it is a generalization of the matrix rank, tensor rank is not as well behaved and understood [9]), nor to estimate posterior confidence in results.

One approach to the Bayesian treatment of count models is based on augmentation using Pólya-Gamma (PG) variables [11]. This solution enables Gibbs sampling in models which combine Gaussian latent structure with logistic, Poisson or negative binomial observations, including factor models [12] and generalised-linear regression [11], and it has been used successfully in neuroscience settings [13]. While PG augmentation has also been considered for tensor decomposition [14], these authors recognised the computational challenges posed by Gibbs sampling in typical size problems, and so focused primarily on point estimation while incorporating a limited range of priors.

Here, we introduce a variational approach to PG-augmented models with negative-binomial observations, which enables a relatively efficient Variational Bayesian (VB) treatment. Our approach is based on a principled approximation of the PG cross entropy, which is especially well suited for neural datasets in which conditional Fano factors (FF) (once the contribution of population-wide shared influences is discounted) are often close to one [15, 16]. The fully probabilistic formulation is able to handle missing observations, extending to situations where data from multiple animals or recording sessions are to be combined (sometimes referred to as "stitching" [17]). VB also makes it possible to combine Automatic Relevance Determination (ARD) with knowledge about expected group structures based on anatomical or histological information about the recorded neurons (brain area, morphology, protein expression etc.). Finally, we also augment standard CP decomposition models using a constrained offset tensor, which allows us to study modulation of neural activity around baseline values, improving readability and interpretability of the CP-factors.

The paper is organised as follows: In section 2 we review background material on tensor decomposition and PG augmentation schemes. In section 3, we derive a VB algorithm for approximate inference of the tensor factors and model parameters. Last, in section 4, we analyse synthetic data and neural recordings in mice performing passive multisensory integration [18] and show that our method can estimate the population-level effects of temporal dynamics and experimental condition in a fully probabilistic manner. We show improvements in variance explained, deviance and, last but not least, decomposition robustness when compared to standard CP and GCP baselines.

2 Background

2.1 Tensor decomposition

Consider observed spike counts gathered in a D-dimensional tensor

$$\mathcal{X} \in \mathbb{R}^{I_1 \times \cdots \times I_D}\,, \tag{1}$$

with dimensions corresponding to neurons, time in trial, various (factorial) experimental manipulations, and possibly trial number. Our goal is to decompose $\mathcal{X}$ into a pair of simpler objects: a low rank tensor $\mathcal{W}$ and an "offset" tensor $\mathcal{V}$ which varies along a restricted set of dimensions. Here $\mathcal{W}$ is designed to capture low-dimensional structure in the time-course of population activity that varies systematically with experimental condition, while the offset $\mathcal{V}$ models potential changes in baseline firing rates with trial or experimental condition.

More specifically, we seek a rank-R Canonical-Polyadic (CP) [9] decomposition of $\mathcal{W}$, often denoted $\mathcal{W} = [|A^{(1)}, \dots, A^{(D)}|]$, such that for $\bar{d} = (d_1, \dots, d_D) \in \prod_{d=1}^{D} \{1 \dots I_d\}$:

$$\mathcal{W}_{\bar{d}} = \sum_{r=1}^{R} A^{(1)}_{d_1 r} \dots A^{(D)}_{d_D r}\,. \tag{2}$$

For $n = 1 \dots D$, $A^{(n)}$ is an $I_n \times R$ matrix, termed a factor, whose rows are $a^{(n)}_i = [A^{(n)}_{i1}, \dots, A^{(n)}_{iR}]$. For a tensor $\mathcal{T}$, we represent by $\mathcal{T}_{(n)}$ its n-th unfolding and recall that:

$$\mathcal{W}_{(n)} = A^{(n)} B^{(n)\intercal}\,, \tag{3}$$

where $B^{(n)}$ is the Khatri-Rao product $\bigodot_{p \neq n} A^{(p)}$.

In addition, we assume that $\mathcal{V}$ varies only along a subset of $D_* \leq D$ dimensions, remaining fixed along the complementary dimensions. For convenience, in what follows we will write $\bar{d} = \bar{d}^* \cup \bar{d}^\bullet$ where $\bar{d}^* = (d^*_1, \dots, d^*_{D_*})$ contains the dimesions that vary and $\bar{d}^\bullet = (d^\bullet_1, \dots, d^\bullet_{D-D_*})$ those that are fixed, regardless of the dimension ordering. Thus, instead of a $I_1 \times \cdots \times I_D$ tensor, in practice we need only estimate the $I_1 \times \cdots \times I_{D_*}$ tensor ν and define $\mathcal{V}$ such that for all $\bar{d}^*$, $\bar{d}^\bullet$:

$$\mathcal{V}_{\bar{d}* \cup \bar{d}\bullet} = \nu_{\bar{d}*}\,. \tag{4}$$

We then take $\mathcal{X}$ to be generated by a negative-binomial noise process from the decomposition determined by $\mathcal{W}$ and $\mathcal{V}$.

2.2 Negative-binomial distribution

The negative binomial (NB) model can be seen as a doubly stochastic relaxation of the Poisson distribution to overdispersed (FF >1) data. We model each spike count $\mathcal{X}_{\bar{d}}$ as Poisson-distributed with random mean $\mu_{\bar{d}}$ drawn from a Gamma distribution with shape ζ and scale $e^{\mathcal{W}_{\bar{d}} + \mathcal{V}_{\bar{d}}}$:

$$\begin{aligned} (\mathcal{X}_{\bar{d}} | \mu_{\bar{d}}) &\sim \text{Poisson}(\mu_{\bar{d}})\,, \\ (\mu_{\bar{d}} | \mathcal{V}_{\bar{d}}, \mathcal{W}_{\bar{d}}, \zeta) &\sim \text{Gamma}(\zeta, e^{\mathcal{W}_{\bar{d}} + \mathcal{V}_{\bar{d}}})\,. \end{aligned} \tag{5}$$

Marginalizing over the Poisson rates $\mu_{\bar{d}}$ then yields the NB distribution [13] :

$$P(\mathcal{X}_{\bar{d}} | \mathcal{V}_{\bar{d}}, \mathcal{W}_{\bar{d}}, \zeta) = \frac{\Gamma(\zeta + \mathcal{X}_{\bar{d}})}{\mathcal{X}_{\bar{d}}!\,\Gamma(\zeta)} (1 - p_{\bar{d}})^{\zeta} (p_{\bar{d}})^{\mathcal{X}_{\bar{d}}} \quad \Rightarrow \quad (\mathcal{X}_{\bar{d}} | \mathcal{V}_{\bar{d}}, \mathcal{W}_{\bar{d}}, \zeta) \sim \text{NB}(\zeta, p_{\bar{d}})\,, \tag{6}$$

where $p_{\bar{d}} = 1/(1 + e^{-(\mathcal{W}_{\bar{d}} + \mathcal{V}_{\bar{d}})})$ is the success probability. We recall that the mean and variance of the NB distribution are respectively given by $\mathbb{E}(\mathcal{X} | \zeta, \mathcal{W}, \mathcal{V}) = \zeta e^{\mathcal{W} + \mathcal{V}}$ and $\mathbb{V}(\mathcal{X} | \zeta, \mathcal{W}, \mathcal{V}) = \zeta e^{\mathcal{W} + \mathcal{V}} (1 + e^{\mathcal{W} + \mathcal{V}})$, giving a Fano factor $\text{FF} = 1 + e^{\mathcal{W}_{\bar{d}} + \mathcal{V}_{\bar{d}}}$ that is greater than one and approaches the Poisson case when ζ increases to infinity while keeping the mean constant.

2.3 Pólya-Gamma augmentation

Unfortunately, the negative binomial observation probability does not admit a conjugate prior on $\mathcal{W}$ and $\mathcal{V}$ which makes direct Bayesian inference intractable. This challenge is resolved by adopting a PG augmentation [11]. We first recall key properties of PG distributions before deriving the augmented model likelihood.

2.3.1 Definition and properties

For $\xi \in \mathbb{R}^+$ and $\omega \in \mathbb{R}$, a random variable U is PG(ξ, ω) distributed iff:

$$U \overset{D}{=} \frac{1}{2\pi^2} \sum_{k=1}^{\infty} \frac{g_k}{(k-1/2)^2 + \omega^2/(4\pi^2)}\,, \tag{7}$$

where $g_k \sim \text{Gamma}(\xi, 1)$. We recall three major results from [11].

(i) For $\xi > 0$, if $U \sim \text{PG}(\xi, 0)$, then for all ζ, v:

$$\frac{e^{v\zeta}}{(1+e^v)^\xi} = 2^{-\xi} e^{(\zeta - \xi/2)v} \int_0^\infty e^{-\frac{U}{2}v^2} p(U|\xi, 0) dU\,. \tag{8}$$

(ii) The densities of PG(ξ, ω) and PG$(\xi, 0)$ are linked through:

$$p(U|\xi, \omega) = \cosh^\xi(\omega/2) e^{-\frac{U}{2}\omega^2} p(U|\xi, 0)\,. \tag{9}$$

(iii) Although the PG density can be expressed as an alternating-sign sum of inverse gamma distributions, the terms of the sum diverge rapidly as ξ increases. However, the Laplace transform of the density (the sign-reflected moment generating function) has the closed form:

$$\mathbb{E}\left(e^{-tU}\right) = \cosh^\xi(\omega/2) \cosh^{-\xi}\left(\sqrt{\frac{\omega^2/2 + t}{2}}\right)\,. \tag{10}$$

Moments of the distribution are obtained from derivatives of this transform at 0, giving:

$$\mathbb{E}(U) = \frac{\xi}{2\omega} \tanh(\omega/2) \text{ and } \mathbb{V}(U) = \frac{\xi}{4\omega^3 \cosh^2(\omega/2)} \left(\sinh(\omega) - \omega\right)\,, \tag{11}$$

with similar closed-form expressions for higher-order moments.

These properties will, respectively, allow us to (i) augment our model with a latent tensor and obtain a Gaussian likelihood in $\mathcal{W}$ and $\mathcal{V}$, (ii) derive the latent posterior distribution and (iii) introduce and validate a moment-matching approximation to PG cross entropies, thus facilitating variational Bayesian inference.

2.3.2 Model augmentation

Let $x \sim \text{NB}(\zeta, p)$, with $p = 1/(1+e^{-v})$. Introducing a random variable $u \sim \text{PG}(\zeta + x, 0)$ we combine (6) and (8) to obtain

$$P(x|\zeta, v) = \frac{\Gamma(\zeta+x)}{x!\,\Gamma(\zeta)} \frac{(e^{-v})^\zeta}{(1+e^{-v})^{\zeta+x}} = \frac{\Gamma(\zeta+x)}{x\Gamma(\zeta)} 2^{-(\zeta+x)} e^{\left(\frac{x-\zeta}{2}\right)v} \int_0^\infty e^{-\frac{u}{2}v^2} p(u|\zeta+x, 0) du\,. \tag{12}$$

This expression allows us to condition on u to obtain the augmented likelihood:

$$P(x|\zeta, u, v) = F(x, \zeta) e^{-\frac{u}{2}\left(v - \frac{x-\zeta}{2u}\right)^2} \tag{13}$$

(with F gathering terms independent of u and v) which is conjugate to a Gaussian prior on v.

Applying this augmentation to the tensor model, we introduce a tensor of PG variates $\mathcal{U}$ of the same size as $\mathcal{X}$, with each $\mathcal{U}_{\bar{d}} \sim \text{PG}(\zeta + \mathcal{X}_{\bar{d}}, 0)$ to obtain the log-likelihood

$$\log P(\mathcal{X}|\mathcal{U}, \mathcal{V}, \mathcal{W}, \zeta) =_{+C} -\frac{1}{2} \sum_{\bar{d}} \mathcal{U}_{\bar{d}} \left(\mathcal{W}_{\bar{d}} + \mathcal{V}_{\bar{d}} - \frac{\mathcal{X}_{\bar{d}} - \zeta}{2\mathcal{U}_{\bar{d}}}\right)^2\,. \tag{14}$$

3 Structured variational inference

PG-based augmentation was developed to facilitate Gibbs sampling in otherwise non-conjugate models. However, despite the availability of closed-form updates, sampling quickly becomes impractical for large datasets [14]. Thus, in this section, we propose a variational Bayes estimation procedure and report the associated updates (derived in Supplementary Materials A) to fit variational posteriors over the low rank tensor, the constrained offset tensor, and the tensor of PG latents. We incorporate prior knowledge about the neural recordings and perform Automatic Relevance Determination (ARD) using Gamma priors on the factor precision matrices. Last, we treat the shape ζ as a model parameter that we efficiently optimize during the variational M-step by approximating PG cross entropy terms with Gamma moment matching.

3.1 Priors

The prior on $\mathcal{U}$ is defined implicitly by the model augmentation through the conditional $P(\mathcal{U}_{\bar{d}}|\mathcal{X}_{\bar{d}}) = \mathrm{PG}(\zeta + \mathcal{X}_{\bar{d}}, 0)$. We set the prior on the offset matrix to be element-wise normal: $P(\nu_{\bar{d}*}) = \mathcal{N}(\mu_{\bar{d}*}, 1/\tau_{\bar{d}*})$, for $\bar{d}^* = (d_1^*, \dots, d_{D_*}^*)$. Finally, for $n = 1 \dots D$, $i = 1 \dots I_n$ we use diagonal precision matrices $\Lambda_i^{(n)}$ to define the priors on the factor rows within $\mathcal{W}$:

$$P(a_i^{(n)}|\Lambda_i^{(n)}) = \mathcal{N}(0, \left(\Lambda_i^{(n)}\right)^{-1}) \,. \tag{15}$$

The parametrisation of $\Lambda_i^{(n)}$ can encode various forms of structure expected in the data, for example:

Case 1: Rank ARD. We assume that precision matrices are shared across a set of modes and rows and use a gamma prior to perform automatic relevance determination. That is, we take:

$$\Lambda_i^{(n)} = \mathrm{Diag}(\lambda_1, \dots, \lambda_R) \,, \qquad \text{where } P(\lambda_r) = \mathrm{Gamma}(k_{0\lambda}, \theta_{0\lambda}) \,. \tag{16}$$

During the optimization procedure, the shared diagonal elements of $\Lambda_i^{(n)}$ may diverge, effectively reducing the rank of the tensor.

Case 2: Neuron-group constraints. Alternatively, we can impose a mode-specific precision. For example we can group neurons based on their recording site and impose shared precision matrices across neurons in each group. If neuron factors are gathered in $A^{(n)}$, we denote by $g(i) \in \mathcal{G}$ the group of neuron i and define:

$$\Lambda_i^{(n)} = \mathrm{Diag}(\lambda_{g(i),1}, \dots, \lambda_{g(i),R}) \,, \qquad \text{where } P(\lambda_{g(i)r}) = \mathrm{Gamma}(k_{0\lambda}, \theta_{0\lambda}) \,. \tag{17}$$

In this case, divergence of a subset of precisions during optimization may lead components linked to one or more groups of neurons to shrink away. Thus, the model will favour explanations of the activity of each group that use as few components as possible.

3.2 Variational distribution

For the variational distribution on the latents $\mathcal{Z} = \{\mathcal{U}, \mathcal{V}, \mathcal{W}, \Lambda\}$, we adopt the factorisation:

$$q(\mathcal{Z}) = q(\mathcal{U})q(\mathcal{V})q(\mathcal{W})q(\Lambda) = \prod_{\bar{d}} q(\mathcal{U}_{\bar{d}}) \prod_{\bar{d}*} q(\nu_{d*}) \prod_{n,i} q(a_i^{(n)}) q(\Lambda|k_\lambda, \theta_\lambda) \,, \tag{18}$$

and then iteratively maximize the VB free energy (or ELBO) $\mathcal{F}(\zeta) = \langle \log P(\mathcal{X}, \mathcal{Z}|\zeta) \rangle_q + \mathcal{H}[q]$. If $q_{\neg x}$ corresponds to the variational distribution with variable x marginalised out, we recall that the variational E-step update has the form $q(x) \propto \exp\langle \log P(\mathcal{X}, \mathcal{Z}|\zeta) \rangle_{q_{\neg x}}$ [19] (when obvious, we drop the explicit form of the distribution from the angle brackets).

3.3 Variational E-step

Here we describe the updates to infer the variational distributions of $\mathcal{W}$,$\mathcal{V}$ and $\mathcal{U}$. Details are provided in Supplementary A. We write $\langle \mathcal{Y}_{\bar{d}} \rangle := (\mathcal{X}_{\bar{d}} - \zeta)/(2\langle \mathcal{U}_{\bar{d}} \rangle)$, retaining the expectation brackets in the notation to emphasise the dependence on the complementary variational distributions.

Update of low rank tensor $\mathcal{W}$. For $n = 1 \dots D, \; i = 1 \dots I_n$:

$$q\left(a_i^{(n)\intercal}\right) = \mathcal{N}\left(m_i^{(n)}, \Sigma_i^{(n)}\right) \,, \tag{19}$$

where the mean and variance are given by:

$$\Sigma_i^{(n)} = \left(\left\langle B^{(n)\intercal} \text{Diag}\left(\langle \mathcal{U}_{(n),i:} \rangle \right) B^{(n)} \right\rangle + \left\langle \Lambda_i^{(n)} \right\rangle \right)^{-1} , \tag{20}$$

$$m_i^{(n)} = \Sigma_i^{(n)} \Big(\langle B^{(n)\intercal} \rangle \text{Diag}\left(\langle \mathcal{U}_{(n),i:} \rangle \right) \Big) \Big(\langle \mathcal{Y}_{(n),i:} \rangle - \langle \mathcal{V}_{(n),i:} \rangle \Big) . \tag{21}$$

Update of factor precisions Λ. By conjugacy, the precision update takes the form:

$$q(\Lambda | k, \theta) \sim \text{Gamma}(k_\lambda, \theta_\lambda) , \tag{22}$$

where the parameters k_λ and θ_λ depend on the form of the prior (Rank ARD or Neuron-group constraint).

Update of offset tensor $\mathcal{V}$. For $\bar{d}^* = (d_1^*, \dots, d_{D_1}^*)$, we need to sum over constrained dimension to update $q(\nu_{\bar{d}*})$:

$$q(\nu_{\bar{d}*}) = \mathcal{N}\left(m_{\bar{d}*}, \Sigma_{\bar{d}*} \right) , \tag{23}$$

with

$$\Sigma_{\bar{d}*} = \left(\tilde{\mathcal{U}}_{\bar{d}*} + \tau_{\bar{d}*} \right)^{-1} , \quad m_{\bar{d}*} = \Sigma_{\bar{d}*} \left(\tilde{\mathcal{U}}_{\bar{d}*} \tilde{\mathcal{Y}}_{\bar{d}*} + \tau_{\bar{d}*} \mu_{\bar{d}*} \right) , \tag{24}$$

and

$$\tilde{\mathcal{U}}_{\bar{d}*} = \sum_{\bar{d}\bullet} \langle \mathcal{U}_{\bar{d}*\cup\bar{d}\bullet} \rangle , \quad \tilde{\mathcal{Y}}_{\bar{d}*} = \frac{\sum_{\bar{d}\bullet} \langle \mathcal{U}_{\bar{d}*\cup\bar{d}\bullet} \rangle \left(\langle \mathcal{Y}_{\bar{d}*\cup\bar{d}\bullet} \rangle - \langle \mathcal{W}_{\bar{d}*\cup\bar{d}\bullet} \rangle \right)}{\sum_{\bar{d}\bullet} \langle \mathcal{U}_{\bar{d}*\cup\bar{d}\bullet} \rangle} . \tag{25}$$

Update of PG latent $\mathcal{U}$. For $\bar{d} = (d_1, \dots, d_D)$, using (12) and (9), we deduce that the PG variational distribution takes the form

$$q(\mathcal{U}_{\bar{d}}) = \text{PG}\left(\zeta + \mathcal{X}_{\bar{d}},\ \Omega_{\bar{d}} \right) , \text{ where } \Omega_{\bar{d}} = \sqrt{\left\langle \left(\mathcal{W}_{\bar{d}} + \mathcal{V}_{\bar{d}} \right)^2 \right\rangle} . \tag{26}$$

3.4 M-step

Finally, we fix the variational distribution and maximize the free energy with respect to the shape parameter. The main computational challenge in this step is to estimate the values of PG cross entropies. As we need one such estimate for each entry in $\mathcal{X}$, both sampling methods and numerical approximations of PG densities prove too expensive. We therefore developed a moment-matching procedure (see Supplementary Materials B) to obtain an efficient approximation for the regime where the conditional Fano factor is near 1.

Denote by m_{PG}^n and m_G^n the n^{th} moments respectively of PG(ξ, ω) and of the Gamma distribution with matched mean and variance. As we show in Supplementary Materials B, the learnt model introduces coupling between the parameters ξ and ω and the inferred conditional Fano Factor, with the result that both parameters diverge as FF $\to 1$ but the ratio $\xi e^{-\omega}$ remains finite. Furthermore, provided that $\Omega = \mathcal{O}(|\mathcal{W} + \mathcal{V}|)$ (ie. the variational posterior variance of the estimated tensor is of the order of its squared mean), the leading order terms in the polynomial expansions of m_{PG}^n and m_G^n are both equal to $\left(\frac{\xi}{2\omega}\right)^n$ and so their higher-order moments are asymptotically equivalent in the limit FF $\to 1$. This motivates an approximation of the PG KL divergence by the divergence between corresponding moment-matched gamma distributions.

Empirically, we find that the approximation is most accurate for small values of KL (see Supplementary Fig. 7), but in practice posterior and prior PG distributions may be very different. Fortunately, we can exploit the analytic properties of the PG distribution (Eq. 9) to re-write the free energy in terms of KL divergences between PG distributions which differ only in their first parameter:

$$\begin{aligned} \mathcal{F}(\zeta) =_{+C} & \sum_{\bar{d}} \log \Gamma(\mathcal{X}_{\bar{d}} + \zeta) / \Gamma(\zeta) - \zeta \left(\log 2 + \langle \mathcal{W}_{\bar{d}}/2 \rangle + \langle \mathcal{V}_{\bar{d}}/2 \rangle + \log\cosh\left(\Omega_{\bar{d}}/2 \right) \right) \\ & - \sum_{\bar{d}} \text{KL}\left(PG(\mathcal{X}_{\bar{d}} + \hat{\zeta}, \Omega_{\bar{d}}) || PG(\mathcal{X}_{\bar{d}} + \zeta, \Omega_{\bar{d}}) \right) , \end{aligned} \tag{27}$$

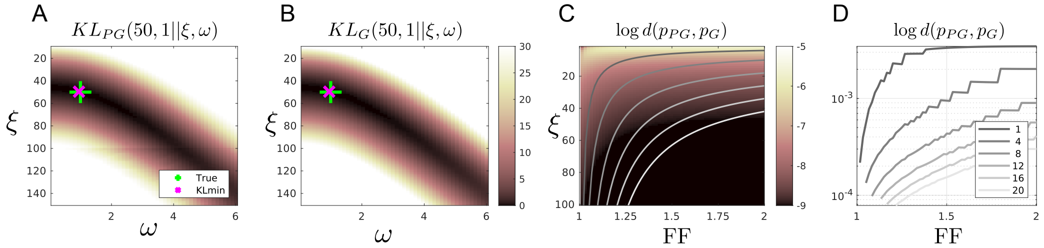

Figure 2: Moment-Matching Pólya-Gamma (PG) and Gamma (G) distribution in the unity Fano factor (FF) limit. (A) Numerical estimates of KL divergences between PG$(50, 1)$ and PG(ξ,ω). (B) Moment matched Gamma KL values. $KL_G(50, 1||\xi, \omega)$ indicates that we estimate $KL_{PG}(50, 1||\xi, \omega)$ using the closed formed KL divergence between two Gamma distribution moment matched to $PG(50, 1)$ and $PG(\xi, \omega)$. Color crosses point to the minimum of the KL divergences. (C) Normalized L_2 norm between the densities p_{PG} and p_G as a function of an effective Fano factor FF $= 1 + e^{-\omega}$, $d(f_{PG}, f_G) = 2||p_{PG} - p_G||^2_{L_2} / (||p_{PG}||^2_{L_2} + ||p_G||^2_{L_2})$). The plain traces indicate relationship between ξ and FF for different number of observed spikes. (D) Values of $d(f_{PG}, f_G)$ along such traces.

where we have omitted terms independent of ζ, and $\hat{\zeta}$ is the current estimate of the shape parameter. The remaining KL divergence in (27) is comparatively small, and can be well approximated by the moment-matched gamma divergence.

We verified the validity of this approximation in numerical experiments (Supplementary B.5). The moment-matched gamma divergence replicated that between the PG distributions over a wide range (Figure 2 A-B) with accuracy increasing monotonically as conditional $FF \to 1$ (Figure 2 C-E).

It is important to stress that the quality of the approximation depends on the conditional rather than marginal Fano factor; that is the FF derived from residual variance after shared trial-to-trial variability captured by common factors in the population has been discounted. While marginal Fano factors that include all sources of variance in neural data are often found to exceed 1.5, population models tend to identify conditional Fano factors much closer to 1 [20]. Thus the moment-matched gamma approximation makes it possible to scale PG augmentation methods to large tensor decompositions in a regime appropriate to neural data analysis.

3.5 Dealing with missing data

Finally, we note that neural data analysis often faces missing data (failing electrode, confounded experiments, etc.) or multi-subject experiments, where a given neuron is only recorded in some trials. Although alternating gradient type updates, where the likelihood is maximized only over the observed entries, can accommodate such cases, they can be highly inefficient as they deal with a full size tensor. Instead, our variational scheme reduces the effective size of the observed dataset as the variational updates previously described are adjusted simply by considering observed entries only.

4 Experiments

In this section, we first explore the accuracy of tensor factorisations obtained by our proposed algorithm on simulated datasets. We confirm the robustness of the method to missing data, and compare the accuracy of ARD-based estimates of tensor rank to alternatives. We then benchmark the method against standard CP [9] and GCP [10] decompositions on neural spike data. We report performance in terms of variance explained, deviance explained, and a decomposition similarity metric. Finally, we look at the decomposition factors themselves, how they can be interpreted, and the consequences for neural data analysis.[1]

[1] All experiment were performed using an Intel(R) Xeon(R) CPU E5-2620 v3 @ 2.40GHz with 65GB of RAM. Benchmark analysis took approximately 12 hours.

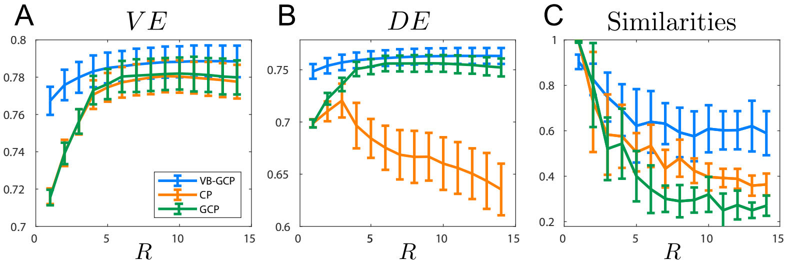

Figure 3: Held out (trial based) performances as a function of the tensor rank R between our probabilistic model (VB-GCP), CP and Poisson GCP. (A) Variance Explained. (B) Poisson Deviance explained. (C) Similarity metric. Error bars indicate 1 standard deviation.

4.1 Simulations

We simulated two 5 dimensional data tensors (size $100 \times 70 \times 3 \times 5 \times 4$) using the generative NB model described in (6). In both cases, the generative shape parameter was fixed to $\zeta = 80$ and the generative tensor rank was $R = 4$. Assuming that the first dimension accounts for neuron loadings, we simulated "neuron groups" by restricting each CP-component to load only on a subset of neurons.

The first simulation included an offset tensor that was constrained to vary across the first and third dimension, and incorporated a multi-experiment stitching setting in which only 1/4 of the full tensor was observed. Supplementary Figures 8 and 9 show that the algorithm, initialized with $R = 6$ components, infers posteriors concentrated around the correct tensor decomposition, while identifying the correct shape parameter ζ. Automatic relevance determination eliminates two of the components thus effectively estimating the rank of the data tensor. For comparison, using the exact same initialization but without ARD, two spurious components are identified (See Supplementary Figure 9).

The second simulation was designed to compare the quality of ARD-based rank estimates to other empirical rank selection metrics. To ensure standard CP and GCP methods [7, 9] were applicable we did not include an offset tensor. Results in Supplementary Figure 10 illustrate the fragility of alternative approaches.

4.2 Spike recordings

We then applied our method to neural spiking data [18] recorded in mice during a multisensory integration paradigm that aimed to elucidate the contribution of vestibular and visual signals to self-motion representation in the cortex. In brief, high-density single-unit recordings (Neuropixels and Neuronexus) were made from the retrosplenial cortex (RSP) of adult mice while they were presented with three types of motion stimuli: (i) passive horizontal rotation of the mouse in the dark (vestibular), (ii) passive horizontal rotation of the mouse in presence of a stationary surround vertical grating (vestibular + visual = both), and (iii) horizontal rotation of the surround vertical grating while the mouse was stationary, thus simulating optic flow alone (visual). (Figure 4-A). Single units were isolated in both granular and dysgranular divisions of the RSP (RSPg, RSPd) and across all cortical layers (Figure 4-C, left). The motion profile of the rotation is shown in Figure 4-B and was adapted such that the visual motion component under 'both' and 'visual' conditions were identical. The dataset consists of 676 neurons recorded under all experimental conditions over 10 trials. Each trial lasted 7 seconds (arranged in 0.1 second non overlapping time bins) during which the mouse was rotated back and forth from a reference 0 degree position to an angle of π. For each of the 24 cross validation folds, we split the dataset in half across trials. The algorithms were trained on the summed spikes of one half, and performance evaluated on the other half as a function of the tensor rank.

We benchmark our algorithm against both CP and Poisson GCP[2]. The latter is defined using an exponential link function for easier model comparison as well as more natural interpretation of the neural data. In plots, we refer to our method as VB-GCP (Variational Bayes GCP). All algorithms were trained with a maximum of 10000 iterations. For VB-GCP, we used $k_{0\lambda} = 100$, $\theta_{0\lambda} = 1$

[2] https://github.com/hugosou/vbgcp and https://www.tensortoolbox.org

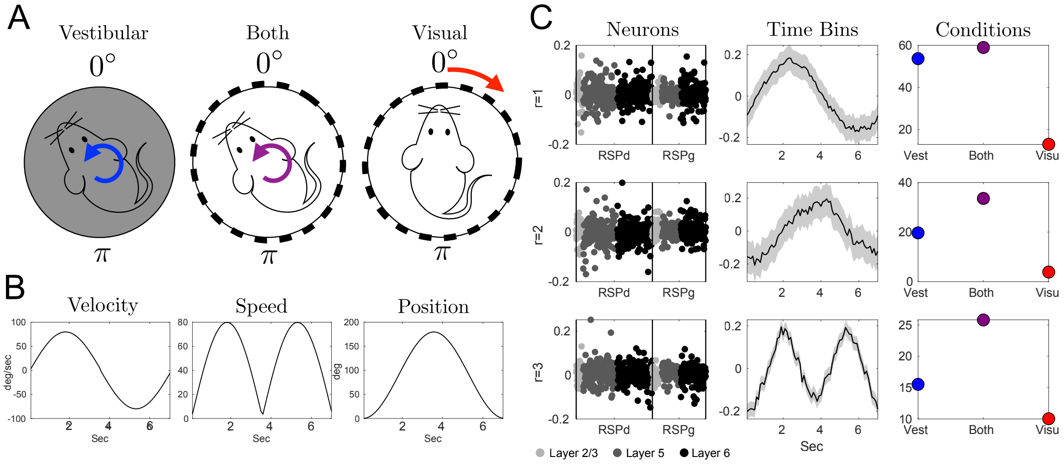

Figure 4: Neural data analysis. (A) Experimental conditions: the mouse is rotated in the dark (Vestibular), rotated in presence of a surround static visual scene (Both), or is stationary while the surrounding visual scene is being rotated (Visual). (B) Temporal profiles of the rotation stimuli under all conditions. (C) First three discovered factors using our decomposition. Grey patches account for 1 standard deviation around the mean based on variational posteriors estimates.

(see (16)). Neuron groups (see (17)) were based on the neuron recording sites (Layer and RSP division), and the offset tensor was only allowed to vary across the neuron and experimental condition dimensions.

We report testing variance explained (VE) and Poisson deviance explained (DE) as well as a similarity metric (all defined in Supplementary Section C). The latter is essentially a normalised scalar product between different the CP factors (on a scale of 0 to 1) that accounts for possible permutations of the components. Similarity is measured for factors obtained with different initializations and from different cross-validation folds. Thus it reflects both algorithmic stability and stability of the extracted structure across different trials under consistent experimental conditions.

We find that VB-GCP outperforms CP and GCP, even though VE and DE explicitly assess performance under the Gaussian and Poisson likelihood losses assumed by those methods (Figure 3 A-B). At low tensor rank, some of the improvement might come from the inclusion of the constrained offset tensor in the model. In particular, for a fixed rank R, our model includes more parameters than standard (G)CP. For a dataset of size $N \times T \times K$ (neurons × times × conditions/trials), a standard rank-R CP decomposition has $R \times (N + T + K)$ parameters, to which the offset adds $N \times K$. However, VB-GCP test performance continues to climb with increased R, while the maximum-likelihood approaches appear to overfit even though they are applied to a model with fewer parameters (both VE and DE on training data exceed that of the VB model, but test performance falls; Figure 11 Supplementary Materials). Thus, the robustness of VB estimation plays an important role in facilitating the offset expansion of the model. While the expanded model appears valuable in the context of our experimental data, it may be omitted in other contexts without disrupting the remainder of the algorithm.

We also note that VB-GCP generates more reliable decompositions across cross-validation folds (Figure 3 C). This reliability reinforces the interpretability of the specific decompositions found, making them of particular practical value in a neuroscience context.

The Poisson distribution is a standard model of stochastic neuronal firing [21] and so we present Poisson-GCP as a relatively familiar baseline model. Our dataset displayed only moderate overdispersion (see Supplementary Materials), and so was close to the Poisson noise assumption. Furthermore, we compared model performance in terms of Poisson Deviance Explained (DE), which is the loss directly optimized by Poisson GCP. We also fit NB-GCP models but did not find notable performance improvements compared to the Poisson case (see Supplementary Section E).

Finally, we considered the interpretability of the factors found by the different models. In Supplementary Figures 12,13 and 14, for each method, we report the rank $R = 6$ decomposition which was the most similar to others across folds. Contrary to CP and GCP, our method is able to report posterior estimates of the factors, and, thanks to the offset tensor model, it directly renders patterns of modulation around condition-dependent baseline values. The discovered temporal dynamics are smoother and centered, hence easier to compare between each other and to relate to experimental variables. In Figure 4-C we show the first three inferred components (sorted by amplitude). Strikingly, and even though they are extracted in a fully unsupervised manner, they are highly reminiscent of the experimental motion variables (namely the angular velocity, position, and absolute angular speed, 4-B). Particularly, one can notice that component $r = 3$, which correlates almost perfectly with mouse absolute rotation velocity, corroborates the decoding analysis performed in [18]. In the latter, combined vestibular and visual information led to higher decoding accuracy. Here, it led to bigger modulatory effects on population activity as shown in the "conditions" panel. Both analyses thus suggest vestibulo-visual integration of rotation speed. In summary, our analysis revealed the encoding of motion variables while disentangling the contribution of different sensory modalities and brain region or layers.

5 Discussion

We have introduced a variational Bayesian framework for probabilistic tensor decomposition of count data, amenable and scalable to neural datasets. Our algorithm provides a principled way to estimate the rank of the data and can incorporate prior knowledge about the neural recordings (such as the neuron shape or the recording site). In practice, and although this comes at a significant computational cost compared to finely optimized blackbox tensor toolboxes [9], our algorithm leads to improved reconstruction performance, and perhaps more importantly to greater robustness compared to baseline (G)CP decompositions. That is, our method was both able to better capture variance in the dataset, and recovered decompositions that were more stable across experimental trials. Such stability is key in order to draw accurate conclusions from the multi-way analysis of neural datasets. One additional feature of our method is the introduction of the constrained offset tensor to the decomposition. In the context of the multisensory integration experiment [18] presently analyzed, it factors out the baseline firing rates of neurons and centers temporal factors, which facilitates the interpretation of the modulatory impact of experimental parameters on neural activity.

Our method builds upon previous work using PG augmentation schemes [11–13] and extends it to a Bayesian treatment of tensor decomposition. Although the PG approach was also adopted in [14], the authors used it to develop an EM algorithm for maximum a posterior point-estimates of the tensor factors, with no posterior variance. Moreover, they proposed a different way of inferring the tensor rank, relying on a multiplicative Gamma prior on the factor amplitudes. Although this scheme also allows probabilistic treatment of the rank, the focus on the amplitudes prevents the model from incorporating knowledge about specific modes of the dataset (such as neuron groups in the present paper). In contrast, we made a variational treatment of the factors possible by approximating PG KL divergences with their Gamma moment-matched counterparts in regions where we expect them to be well behaved. Nevertheless, as with many mean-field or factored variational inference procedures, it is possible that our method underestimate the true uncertainty in the model variables. Possible extensions could therefore involve low-rank approximations to the posterior covariance between factors, or the extension of linear-response methods [22] to PG-augmented models that lie outside the exponential family.

Finally, our model can cope with Poisson ($FF = 1$) and overdispersed datasets ($FF > 1$) but underdispersion ($FF < 1$) has been reported in several neural circuits like the early visual system [23], the somatosensory cortex [24], the Dorsal Premotor Cortex (after presentation of a stimulus) [20] or the auditory cortex [25]. Interestingly, in this last example, DeWeese and Zador found that single-unit spiking activity was underdispersed to the point that the majority of neurons exhibited binary behavior with few multi-spike responses. In this particular case, the Pólya-Gamma augmentation based tensor decomposition could easily be adapted by switching from a negative binomial to binomial observation model using the original model from [11]. In the other cases, neuron responses seem more complex and would require a more sophisticated extension of our model. One solution to low variability settings is to model spiking history and neuron coupling (see for example [26]), but its incorporation to the tensor factorisation framework would require further developments.

Acknowledgments and Disclosure of Funding

We would like to thank Joaquin Rapela for early contributions to this project and Marc Deisenroth and Céline Marié for helpful comments on the manuscript. This work was funded by the Simons Foundation and the Gatsby Charitable Foundation.

References

[1] C. J. Appellof and E. R. Davidson. Strategies for Analyzing Data from Video Fluorometric Monitoring of Liquid Chromatographic Effluents. *Analytical Chemistry*, 53(13):2053–2056, 1981.

[2] M. Alex O. Vasilescu and Demetri Terzopoulos. Multilinear analysis of image ensembles: Tensorfaces. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 2350, pages 447–460. Springer Verlag, 2002.

[3] Fumikazu Miwakeichi, Eduardo Martínez-Montes, Pedro A. Valdés-Sosa, Nobuaki Nishiyama, Hiroaki Mizuhara, and Yoko Yamaguchi. Decomposing EEG data into space-time-frequency components using Parallel Factor Analysis. *NeuroImage*, 22(3):1035–1045, jul 2004.

[4] C. F. Beckmann and S. M. Smith. Tensorial extensions of independent component analysis for multisubject FMRI analysis. *NeuroImage*, 25(1):294–311, mar 2005.

[5] Justen Geddes, Gaute T. Einevoll, Evrim Acar, and Alexander J. Stasik. Multi-Linear Population Analysis (MLPA) of LFP Data Using Tensor Decompositions. *Frontiers in Applied Mathematics and Statistics*, 6:41, sep 2020.

[6] Arno Onken, Jian K. Liu, P. P.Chamanthi R. Karunasekara, Ioannis Delis, Tim Gollisch, and Stefano Panzeri. Using Matrix and Tensor Factorizations for the Single-Trial Analysis of Population Spike Trains. *PLoS Computational Biology*, 12(11), nov 2016.

[7] Alex H. Williams, Tony Hyun Kim, Forea Wang, Saurabh Vyas, Stephen I. Ryu, Krishna V. Shenoy, Mark Schnitzer, Tamara G. Kolda, and Surya Ganguli. Unsupervised Discovery of Demixed, Low-Dimensional Neural Dynamics across Multiple Timescales through Tensor Component Analysis. *Neuron*, 98(6):1099–1115.e8, jun 2018.

[8] Jeffrey S Seely, Matthew T Kaufman, Stephen I Ryu, Krishna V Shenoy, John P Cunningham, and Mark M Churchland. Tensor Analysis Reveals Distinct Population Structure that Parallels the Different Computational Roles of Areas M1 and V1. *PLoS Comput Biol*, 12(11):1005164, 2016.

[9] Tamara G Kolda and Brett W Bader. Tensor Decompositions and Applications *. *Society for Industrial and Applied Mathematics*, 51(3):455–500, 2009.

[10] David Hong, Tamara G. Kolda, and Jed A. Duersch. Generalized Canonical Polyadic Tensor Decomposition. *SIAM Review*, 62(1):133–163, 2020.

[11] Nicholas G. Polson, James G. Scott, and Jesse Windle. Bayesian inference for logistic models using Polya-Gamma latent variables. *Journal of the American Statistical Association*, 108(504):1339–1349, may 2012.

[12] Arto Klami, Dinh Phung, and Hang Li. Polya-Gamma augmentations for factor models. *PMLR*, 39:112–128, feb 2014.

[13] Jonathan W Pillow and James G Scott. Fully Bayesian inference for neural models with negative-binomial spiking. *Advances in Neural Information Processing Systems*, 25, 2012.

[14] Piyush Rai, Changwei Hu, Matthew Harding, and Lawrence Carin. Scalable probabilistic tensor factorization for binary and count data. *IJCAI International Joint Conference on Artificial Intelligence*, 2015-Janua:3770–3776, 2015.

[15] Giedrius T. Buračas, Anthony M. Zador, Michael R. DeWeese, and Thomas D. Albright. Efficient discrimination of temporal patterns by motion-sensitive neurons in primate visual cortex. *Neuron*, 20(5):959–969, may 1998.

[16] Adam S Charles, Gregory D Horwitz, and Jonathan W Pillow. Dethroning the Fano Factor: A Flexible, Model-Based Approach to Partitioning Neural Variability. *Neural Computation*, 30:1012–1045, 2018.

[17] Srinivas C Turaga, Lars Buesing, Adam M Packer, Henry Dalgleish, Noah Pettit, Michael Häusser, and Jakob H Macke. Inferring neural population dynamics from multiple partial recordings of the same neural circuit. *Advances in Neural Information Processing Systems*, 26, 2013.

[18] Sepiedeh Keshavarzi, Edward F Bracey, Richard A Faville, Dario Campagner, Adam L Tyson, Stephen C Lenzi, Tiago Branco, and Troy W Margrie. The retrosplenial cortex combines internal and external cues to encode head velocity during navigation. *bioRxiv*, page 2021.01.22.427789, jan 2021.

[19] Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. Introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233, nov 1999.

[20] Mark M. Churchland, Byron M. Yu, John P. Cunningham, Leo P. Sugrue, Marlene R. Cohen, Greg S. Corrado, William T. Newsome, Andrew M. Clark, Paymon Hosseini, Benjamin B. Scott, David C. Bradley, Matthew A. Smith, Adam Kohn, J. Anthony Movshon, Katherine M. Armstrong, Tirin Moore, Steve W. Chang, Lawrence H. Snyder, Stephen G. Lisberger, Nicholas J. Priebe, Ian M. Finn, David Ferster, Stephen I. Ryu, Gopal Santhanam, Maneesh Sahani, and Krishna V. Shenoy. Stimulus onset quenches neural variability: A widespread cortical phenomenon. *Nature Neuroscience*, 13(3):369–378, mar 2010.

[21] Peter Dayan, Laurence F Abbott, et al. Theoretical neuroscience: computational and mathematical modeling of neural systems. *Journal of Cognitive Neuroscience*, 15(1):154–155, 2003.

[22] Ryan Giordano, Tamara Broderick, and Michael Jordan. Linear Response Methods for Accurate Covariance Estimates from Mean Field Variational Bayes. *Advances in Neural Information Processing Systems*, 2015-Janua:1441–1449, jun 2015.

[23] Prakash Kara, Pamela Reinagel, and R Clay Reid. Low response variability in simultaneously recorded retinal, thalamic, and cortical neurons. *Neuron*, 27(3):635–646, 2000.

[24] Samuel Andrew Hires, Diego A Gutnisky, Jianing Yu, Daniel H O'Connor, and Karel Svoboda. Low-noise encoding of active touch by layer 4 in the somatosensory cortex. *Elife*, 4:e06619, 2015.

[25] Michael R DeWeese, Anthony M Zador, et al. Binary coding in auditory cortex. *Advances in neural information processing systems*, pages 117–124, 2003.

[26] Wilson Truccolo, Uri T Eden, Matthew R Fellows, John P Donoghue, and Emery N Brown. A point process framework for relating neural spiking activity to spiking history, neural ensemble, and extrinsic covariate effects. *Journal of neurophysiology*, 93(2):1074–1089, 2005.

[27] Marlene R. Cohen and Adam Kohn. Measuring and interpreting neuronal correlations. *Nature Neuroscience*, 14(7):811–819, jul 2011.

[28] Enes Makalic and Daniel F. Schmidt. High-Dimensional Bayesian Regularised Regression with the BayesReg Package. *arXiv.org*, nov 2016.

Supplementary Materials: Probabilistic Tensor Decomposition of Neural Population Spiking Activity

Hugo Soulat
Gatsby Unit
University College London
London, W1T 4JG
hugos@gatsby.ucl.ac.uk

Sepiedeh Keshavarzi
Sainsbury Wellcome Centre
University College London
London, W1T 4JG
s.keshavarzi@ucl.ac.uk

Troy W. Margrie
Sainsbury Wellcome Centre
University College London
London, W1T 4JG
t.margrie@ucl.ac.uk

Maneesh Sahani
Gatsby Unit
University College London
London, W1T 4JG
maneesh@gatsby.ucl.ac.uk

A Details of variational updates

In this section, we derive the variational updates. The complete log joint likelihood of the augmented model can be written

$$\log L = \log P(\mathcal{X},\mathcal{U},\mathcal{V},\mathcal{W},\Lambda|\zeta,\mu,\tau,k,\theta) = \sum_{\bar{d}} \log \mathcal{P}(\mathcal{X}_{\bar{d}},\mathcal{U}_{\bar{d}}|\mathcal{V}_{\bar{d}},\mathcal{W}_{\bar{d}},\zeta)$$

$$+ \sum_{\bar{d}*} \log \mathcal{N}\left(\nu_{\bar{d}*}|\mu_{\bar{d}*}, 1/\tau_{\bar{d}*}\right) + \sum_{n,i} \log \mathcal{N}\left(a_i^{(n)}|0, \left(\Lambda_i^{(n)}\right)^{-1}\right) + \log P(\Lambda|k_\lambda,\theta_\lambda)\,, \quad (28)$$

where the first term represents the Pólya-Gamma augmented factor defined in (12) and (13),

$$\log \mathcal{P}(\mathcal{X}_{\bar{d}},\mathcal{U}_{\bar{d}}|\mathcal{V}_{\bar{d}},\mathcal{W}_{\bar{d}},\zeta) = \log p_{\mathrm{PG}}(\mathcal{U}_{\bar{d}}|\zeta + \mathcal{X}_{\bar{d}}, 0)$$

$$- \tfrac{1}{2}\mathcal{U}_{\bar{d}}\left(\mathcal{W}_{\bar{d}} + \mathcal{V}_{\bar{d}}\right)^2 - \tfrac{1}{2}(\mathcal{X}_{\bar{d}} - \zeta)(\mathcal{W}_{\bar{d}} + \mathcal{V}_{\bar{d}}) + \log F(\mathcal{X}_{\bar{d}},\zeta)\,, \quad (29)$$

and we recall that $\mathcal{V}$ is determined by ν, and $\mathcal{W} = [|A^{(1)}, \dots, A^{(D)}|]$ with $a_i^{(n)}$ the rows of $A^{(n)}$.

The VB factored posterior on $\mathcal{Z} = \{\mathcal{U},\mathcal{V},\mathcal{W},\Lambda\}$ is given by,

$$q(\mathcal{Z}) = q(\mathcal{U})q(\mathcal{V})q(\mathcal{W})q(\Lambda) = \prod_{\bar{d}} q(\mathcal{U}_{\bar{d}}) \prod_{\bar{d}*} q(\nu_{d*}) \prod_{n,i} q(a_i^{(n)}) q(\Lambda|k_\lambda,\theta_\lambda)\,. \quad (30)$$

For each $X \in \{\mathcal{U},\mathcal{V},\mathcal{W},\Lambda\}$, we represent by $q_{\neg X}$ the variational distribution marginalized over X. Then each VB inference update takes the form

$$q\left(X\right) \propto \exp\langle \log P(\mathcal{X},\mathcal{Z}|\zeta,\mu,\tau,k,\theta)\rangle_{q_{\neg X}} \quad (31)$$

We note that the expectation of the PG augmentation term (29) with respect to $q(\mathcal{U})$ can be expressed as

$$\langle \log \mathcal{P}(\mathcal{X},\mathcal{U}|\mathcal{V},\mathcal{W},\zeta\rangle_{q(\mathcal{U})} =_{+C} -\frac{1}{2}\sum_{\bar{d}} \langle \mathcal{U}_{\bar{d}}\rangle \left(\mathcal{W}_{\bar{d}} + \mathcal{V}_{\bar{d}} - \frac{\mathcal{X}_{\bar{d}} - \zeta}{2\langle \mathcal{U}_{\bar{d}}\rangle}\right)^2\,, \quad (32)$$

and introduce the notation

$$\langle \mathcal{Y}\rangle = \frac{\mathcal{X} - \zeta}{2\langle \mathcal{U}\rangle}\,, \quad (33)$$

for compactness, retaining the expectation brackets to emphasise the dependence on $q(\mathcal{U})$.

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

A.1 Updates for low rank tensor $\mathcal{W}$

For $n = 1 \dots D$, $i = 1 \dots I_n$:

$$
\begin{aligned}
&\langle \log L \rangle_{q_{\neg a_i^{(n)}}} =_{+C} \\
&- \frac{1}{2} \left\langle \sum_{i=1}^{I_n} \left(a_i^{(n)} B^{(n)\intercal} - \langle \mathcal{Y}^{\mathcal{V}}_{(n),i:} \rangle \right) \operatorname{Diag}\left(\mathcal{U}_{(n),i:} \right) \left(a_i^{(n)} B^{(n)\intercal} - \langle \mathcal{Y}^{\mathcal{V}}_{(n),i:} \rangle \right)^{\intercal} \right\rangle_{q_{\neg a_i^{(n)}}} ,
\end{aligned}
\tag{34}
$$

where $\mathcal{Y}^{\mathcal{V}} = \mathcal{V} - \langle \mathcal{V} \rangle$. Keeping only the terms in $a_i^{(n)}$, we can write the update:

$$
q\left(a_i^{(n)\intercal} \right) = \mathcal{N}\left(m_i^{(n)}, \Sigma_i^{(n)} \right) ,
\tag{35}
$$

where

$$
\begin{aligned}
\Sigma_i^{(n)} &= \left(\left\langle B^{(n)\intercal} \operatorname{Diag}\left(\langle \mathcal{U}_{(n),i:} \rangle \right) B^{(n)} \right\rangle + \left\langle \Lambda_i^{(n)} \right\rangle \right)^{-1} \\
m_i^{(n)} &= \Sigma_i^{(n)} \left(\langle B^{(n)\intercal} \rangle \operatorname{Diag}\left(\langle \mathcal{U}_{(n),i:} \rangle \right) \right) \left(\langle \mathcal{Y}_{(n),i:} \rangle - \langle \mathcal{V}_{(n),i:} \rangle \right) .
\end{aligned}
\tag{36}
$$

In this expression:

- $\langle \mathcal{U} \rangle$ is given in closed form using the mean of a PG distribution and (47).
- $\langle B^{(n)} \rangle$ is obtained by replacing $a_l^{(m)}$ by $m_l^{(m)}$ (for $m \neq n$) in the Khatri-Rao product.

To simplify this experession further, we write $\tilde{A}^{(n)}$ for the I_n by R^2 matrix whose rows are: $\tilde{a}_i^{(n)} = \operatorname{Vec}\left(m_i^{(n)\intercal} m_i^{(n)} + \Sigma_i^{(n)} \right)^{\intercal}$ and $\tilde{B}^{(n)} = \bigodot_{p \neq n} \tilde{A}^{(p)}$. If j_p is the index associated to the p-th factor in the j-th row $B_{j:}^{(n)}$ of the Khatri-Rao product $B^{(n)}$, then for $r_1, r_2 = 1 \dots R$:

$$
\langle B^{(n)\intercal} \operatorname{Diag}\left(\mathcal{U}_{(n),i:} \right) B^{(n)} \rangle_{r_1 r_2} = \sum_j \langle \mathcal{U}_{(n)ij} \rangle \prod_{p \neq n} \left(m_{j_p}^{(p)\intercal} m_{j_p}^{(p)} + \Sigma_{j_p}^{(p)} \right)_{r_1, r_2} ,
\tag{37}
$$

or, using "$\circ$" to denote the Hadamard product and $\mathbb{1}_{1 \times k}$ a 1 by k vector of ones,

$$
\operatorname{Vec}\left[\left\langle B^{(n)\intercal} \operatorname{Diag}\left(\mathcal{U}_{(n),i:} \right) B^{(n)} \right\rangle \right]^{\intercal} = \mathbb{1}_{1 \times (\prod I_p)} \left(\tilde{B}^{(n)} \circ \left(\mathbb{1}_{1 \times R} \otimes \langle \mathcal{U}_{(n),i:} \rangle \right) \right) \in \mathbb{R}^{1 \times R^2} .
$$

The various components can then be reassembled to give the moments of $q(\mathcal{W})$:

$$
\langle \mathcal{W}_{\bar{d}} \rangle = [| \langle A^{(1)} \rangle, \dots, \langle A^{(D)} \rangle |]_{\bar{d}} \quad \text{and} \quad \langle \mathcal{W}_{\bar{d}}^2 \rangle = [| \tilde{A}^{(1)}, \dots, \tilde{A}^{(D)} |]_{\bar{d}} .
$$

A.1.1 Updates for factor precisions Λ

By conjugacy, the precision update takes the form

$$
q(\Lambda | k, \theta) \sim \operatorname{Gamma}(k_\lambda, \theta_\lambda) .
\tag{38}
$$

Case 1: Rank ARD. For $r = 1 \dots R$, if $\mathcal{D}$ indicates the dimensions sharing a precision matrix, we have

$$
k_{\lambda,r} = k_{0\lambda} + \frac{1}{2} \sum_{n \in \mathcal{D}} I_n \quad \text{and} \quad \theta_{\lambda,r} = \left(\theta_{0\lambda}^{-1} + \frac{1}{2} \sum_{n \in \mathcal{D}} \sum_{i=1}^{I_n} \left\langle a_{ir}^{(n)2} \right\rangle \right)^{-1} .
\tag{39}
$$

Case 2: Neuron-Group constraints. For $r = 1 \dots R$ and a neuron group $g \in \mathcal{G}$

$$
k_{\lambda,gr} = k_{0\lambda} + \frac{1}{2} |g| \quad \text{and} \quad \theta_{\lambda,gr} = \left(\theta_{0\lambda}^{-1} + \frac{1}{2} \sum_{i \in g} \left\langle a_{ir}^{(n)2} \right\rangle \right)^{-1} .
\tag{40}
$$

A.1.2 Updates for offset tensor $\mathcal{V}$

For $\bar{d}^* = (d_1^*, \ldots, d_{D_1}^*)$, we want to update $\nu_{\bar{d}*}$. Gathering the terms repeated across constrained dimensions, we get:

$$\langle \log L \rangle_{q_{\neg(\nu_{\bar{d}*})}} =_{+C} \left(-\frac{1}{2} \sum_{\bar{d}\bullet} \langle \mathcal{U}_{\bar{d}*\cup\bar{d}\bullet} \rangle \right) \nu_{\bar{d}*}^2 + \left(\sum_{\bar{d}\bullet} \langle \mathcal{U}_{\bar{d}*\cup\bar{d}\bullet} \rangle \langle \mathcal{Z}_{\bar{d}*\cup\bar{d}\bullet} \rangle \right) \nu_{\bar{d}*} \,. \tag{41}$$

Therefore

$$q(\nu_{\bar{d}*}) = \mathcal{N}\left(m_{\bar{d}*}, \Sigma_{\bar{d}*}, \right) \,, \tag{42}$$

with

$$\Sigma_{\bar{d}*} = \left(\tilde{\mathcal{U}}_{\bar{d}*} + \tau_{\bar{d}*} \right)^{-1} \quad \text{and} \quad m_{\bar{d}*} = \Sigma_{\bar{d}*} \left(\tilde{\mathcal{U}}_{\bar{d}*} \tilde{\mathcal{Y}}_{\bar{d}*} + \tau_{\bar{d}*} \mu_{\bar{d}*} \right) \,, \tag{43}$$

where

$$\tilde{\mathcal{U}}_{\bar{d}*} = \sum_{\bar{d}\bullet} \langle \mathcal{U}_{\bar{d}*\cup\bar{d}\bullet} \rangle \quad \text{and} \quad \tilde{\mathcal{Y}}_{\bar{d}*} = \frac{\sum_{\bar{d}\bullet} \langle \mathcal{U}_{\bar{d}*\cup\bar{d}\bullet} \rangle \left(\langle \mathcal{Y}_{\bar{d}*\cup\bar{d}\bullet} \rangle - \langle \mathcal{W}_{\bar{d}*\cup\bar{d}\bullet} \rangle \right)}{\sum_{\bar{d}\bullet} \langle \mathcal{U}_{\bar{d}*\cup\bar{d}\bullet} \rangle} \,. \tag{44}$$

A.1.3 Updates for Pólya-Gamma latents $\mathcal{U}$

For $\bar{d} = (d_1, \ldots, d_D)$, we leverage (9) to rewrite:

$$\begin{aligned} \left\langle \log \left(e^{-\frac{\mathcal{U}_{\bar{d}}}{2} (\mathcal{W}_{\bar{d}} + \mathcal{V}_{\bar{d}})^2} p(\mathcal{U}_{\bar{d}} | \zeta + \mathcal{X}_{\bar{d}}, 0) \right) \right\rangle_{q_{\neg\mathcal{U}_{\bar{d}}}} &= \log \left(e^{-\frac{\mathcal{U}_{\bar{d}}}{2} \Omega_{\bar{d}}} p(\mathcal{U}_{\bar{d}} | \zeta + \mathcal{X}_{\bar{d}}, 0) \right) \\ &=_{+C} \log p(\mathcal{U}_{\bar{d}} | \zeta + \mathcal{X}_{\bar{d}}, \Omega_{\bar{d}}) \,, \end{aligned} \tag{45}$$

where

$$\Omega_{\bar{d}} = \sqrt{\left\langle \left(\mathcal{W}_{\bar{d}} + \mathcal{V}_{\bar{d}} \right)^2 \right\rangle} \,. \tag{46}$$

Thus, we find that

$$q(\mathcal{U}_{\bar{d}}) = \text{PG}\left(\zeta + \mathcal{X}_{\bar{d}},\ \Omega_{\bar{d}} \right) \,. \tag{47}$$

A.2 Rewriting the free energy

The free energy (or ELBO) is given by:

$$\mathcal{F}(\zeta) = \langle \log P(\mathcal{X}, \mathcal{Z} | \zeta) \rangle_q + \mathcal{H}[q] \,. \tag{48}$$

Dropping terms in the log joint probability which do not depend on ζ:

$$\begin{aligned} \mathcal{F}(\zeta) &=_{+C} \langle \log P(\mathcal{X} | \mathcal{Z}, \zeta) \rangle_q + \langle \log P(\mathcal{U}; \zeta + \mathcal{X}, 0) \rangle_q + \mathcal{H}[q(\mathcal{U})] \\ &=_{+C} \sum_{\bar{d}} \log \Gamma(\mathcal{X}_{\bar{d}} + \zeta) / \Gamma(\zeta) - \zeta \left(\log 2 + \langle \mathcal{W}_{\bar{d}}/2 \rangle + \langle \mathcal{V}_{\bar{d}}/2 \rangle \right) \\ &\quad - \sum_{\bar{d}} \langle \log q(\mathcal{U}_{\bar{d}}) - \log P(\mathcal{U}_{\bar{d}}; \zeta + \mathcal{X}_{\bar{d}}, 0) \rangle_{q(\mathcal{U}_{\bar{d}})} \,. \end{aligned} \tag{49}$$

The last term contains cross entropies between prior and variational PG distributions. We then use the fact that after a variational E-step, $q(\mathcal{U}_{\bar{d}})$ is fixed, and to avoid confusion, we denote by $\hat{\zeta}$ the current estimate of the shape parameter. Recalling that we have:

$$P(\mathcal{U}_{\bar{d}}; \zeta + \mathcal{X}_{\bar{d}}, 0) \propto \cosh^{-(\zeta + \mathcal{X}_{\bar{d}})}(\Omega_{\bar{d}}/2) P(\mathcal{U}_{\bar{d}}; \zeta + \mathcal{X}_{\bar{d}}, \Omega_{\bar{d}}) \,, \tag{50}$$

we obtain:

$$\begin{aligned} \mathcal{F}(\zeta) =_{+C} & \sum_{\bar{d}} \log \Gamma(\mathcal{X}_{\bar{d}} + \zeta) / \Gamma(\zeta) - \zeta \left(\log 2 + \langle \mathcal{W}_{\bar{d}}/2 \rangle + \langle \mathcal{V}_{\bar{d}}/2 \rangle + \log \cosh(\Omega_{\bar{d}}/2) \right) \\ & - \sum_{\bar{d}} \text{KL}\left(PG(\mathcal{X}_{\bar{d}} + \hat{\zeta}, \Omega_{\bar{d}}) || PG(\mathcal{X}_{\bar{d}} + \zeta, \Omega_{\bar{d}}) \right) \,. \end{aligned} \tag{51}$$

B Asymptotic behaviour of the moment matching approximation

In this section we examine the behaviour of the moment-matched Gamma(α,β)-based approximation to the PG cross entropy terms in (27). We first explore the properties of the VB approximation to a PG-augmented negative binomial model, arguing that in realistic neural data settings solutions it will often converge to a parameter regime associated with low conditional Fano factors. We then leverage the Laplace transforms of the PG and gamma densities to show that expansions of the higher-order moments of both distributions are dominated by identical leading terms in this regime. This provides justification for the approximation, the empirical quality of which we then evaluate in numerical experiments.

B.1 The behaviour of the Pólya-Gamma augmented model

Recall from (12) that PG augmentation replaces the factors associated with the observation likelihoods by terms derived from the identity

$$\frac{(e^{-v})^{\zeta}}{(1+e^{-v})^{\zeta+x}} \equiv 2^{-(\zeta+x)} e^{\left(\frac{x-\zeta}{2}\right)v} \int_0^\infty e^{-\frac{u}{2}v^2} p(u|\zeta+x,0)du\,, \tag{52}$$

where $p(u|\zeta+x,0)$ is a PG density. This augmented form would not be useful if our goal was to generate samples of x, as the implied distribution $p(x|u,v)$ is considerably more complex than the original negative binomial $p(x|v)$. Its value arises solely in inference, where the conditional likelihood on v, that is $p(x|v,u)$ viewed as a function of v, *is* simplified to a Gaussian-conjugate form.

This view also emphasises the fact that despite the appearance of the density $p(u|\zeta+x,0)$ in the definition, the implied prior on u is more complex. Nonetheless, as we saw in the previous section, the joint likelihood is tractable and yields a PG variational posterior on u.

To understand the behaviour of the model when fit to neural data, it is helpful to consider the shape of the likelihood landscape generated by the underlying NB observation model

$$p(x|v,\zeta) = \frac{\Gamma(\zeta+x)}{x!\,\Gamma(\zeta)} \frac{e^{-\zeta v}}{(1+e^{-v})^{\zeta+x}}\,. \tag{53}$$

The likelihood is highest when the NB mean (given by ζe^{v}) is close to the observation x and its dispersion, controlled by the Fano factor $\text{FF} = 1+e^{v}$, is small. Thus, for a single observation, optimisation of the NB likehood will drive ζ to diverge, while v remains close to $\log(x/\zeta)$.

The full VB model affects this limit in two ways. First, the likelihoods associated with all the observations $\mathcal{X}$ are linked by the restricted structure of the tensor $\mathcal{W}+\mathcal{V}$, and so cannot be optimised individually. Second, the incorporation of posterior uncertainty on the tensor $\mathcal{W}+\mathcal{V}$ means that rather than optimising a single-point likelihood, we optimise the expectation of its logarithm under $q(\mathcal{W},\mathcal{V})$. However, a similar limit applies if the model provides a good fit to the data with low dispersion. That is, roughly, if the posterior is concentrated such that the $\mathcal{X}$ are dispersed around $\zeta e^{\langle\mathcal{W}+\mathcal{V}\rangle}$ with an effective FF approaching 1. Now, again, optimisation of the model will drive ζ to grow, while the posterior on $\mathcal{W}+\mathcal{V}$ concentrates to keep $\zeta e^{\langle\mathcal{W}+\mathcal{V}\rangle}$ close to $\mathcal{X}$. However, in this case the added constraints of the model introduce a natural limit to this divergence.

To see the impact on the inferred PG parameters, recall that the variational posterior on $\mathcal{U}$ is given by

$$q(\mathcal{U}_{\bar{d}}) = \text{PG}(\zeta+\mathcal{X}_{\bar{d}},\Omega_{\bar{d}}) \quad \text{where} \quad \Omega_{\bar{d}} = \sqrt{\langle(\mathcal{W}_{\bar{d}}+\mathcal{V}_{\bar{d}})^2\rangle}\,. \tag{54}$$

Thus, in the "good fit" limit, the growth of ζ directly increases the first parameter of each PG posterior. Its impact on the second parameter is indirect. When $\zeta > \mathcal{X}_{\bar{d}}$, the corresponding tensor entry will be negative, and grow more negative with increasing ζ (thus driving FF → 1). As $\Omega_{\bar{d}}$ depends on its magnitude, we see that it too grows, but at a rate given by $\log\zeta$.[3]

How plausible is this "good fit" regime? Single neuron spike-counts over repeated trials of an experiment are often observed to be substantially overdispersed relative to Poisson [16]. However,

[3] Parenthetically, we note that this behaviour would be more evident in a different but broadly equivalent parametrisation of the likelihood model as $\mathcal{X}_{\bar{d}} \sim \text{NB}\big(\zeta, 1/(1+e^{(\mathcal{W}_{\bar{d}}+\mathcal{V}_{\bar{d}})-\log\zeta})\big)$.

much of this variability appears to be correlated across neurons [27] and population studies often report conditional Fano factors closer to 1 once such common variance has been captured by a latent variable model such as ours [20]. Thus, this "good fit" setting is one that we expect to be encountered in many neural data sets. For the neural data explored in this study we found a overall conditional FF of approximately 1.17.

B.2 Moment-matched gamma distribution

Matching the first two central moments of Gamma(α, β) to those of PG(ξ, ω) (11) we have

$$\frac{\alpha}{\beta} = \xi\frac{\tanh(\omega/2)}{2\omega} \quad \text{and} \quad \frac{\alpha}{\beta^2} = \frac{\xi}{4\omega^3\cosh^2(\omega/2)}\left(\sinh(\omega) - \omega\right) . \tag{55}$$

As we will see below, these forms determine the leading terms in the higher-order (non-central) moments.

B.3 Higher-order moments

We write f_{PG} and f_G for the Laplace transforms of PG(ξ, ω) and Gamma(α, β) respectively, and recall that

$$f_{PG}(t) = \cosh^{\xi}(\omega/2)\cosh^{-\xi}\left(\sqrt{\frac{\omega^2/2+t}{2}}\right) \quad \text{and} \quad f_G(t) = \left(1+\frac{t}{\beta}\right)^{-\alpha} . \tag{56}$$

The nth moment of each distribution is given by the nth derivative with respect to $-t$ at 0 of the corresponding Laplace transform, that is $(-1)^n f^{(n)}(0)$.

For Gamma(α, β), we have immediately that

$$f_G^{(n)}(0) = (-1)^n\frac{\alpha(\alpha+1)\ldots(\alpha+n-1)}{\beta^n} . \tag{57}$$

The corresponding form for the Pólya-Gamma is more complex. However, we show by induction that for all $n \in \mathbb{N}$, the derivative can be written as a sum of at most 3^n terms in polynomials $\{P_i^n\}$ and $\{Q_i^n\}$, indexed by tuples $i \in \{1, 2, 3\}^n$.

$$f_{PG}^{(n)}(t) = \frac{\cosh^{\xi}(\omega/2)}{4^n}\left(\sum_i \frac{P_i^n[\xi]\,Q_i^n[\cosh g_\omega(t), \sinh g_\omega(t)]}{\cosh^{\xi+p_i} g_\omega(t)\; g_\omega(t)^{d_i}}\right) , \tag{58}$$

where

$$g_\omega(t) = \sqrt{\frac{\omega^2/2+t}{2}} \tag{59}$$

and the degrees of the polynomials satisfy

$$0 \le d(Q_i^n) \le d(P_i^n) = p_i \le n \le d_i . \tag{60}$$

The result follows immediately for $n = 0$ by inspection of the Laplace transform (56). Assume it holds for n. Then differentiating (58) with respect to t and noting that $g'_\omega(t) = 1/(4g_\omega(t))$ yields:

$$\begin{aligned} f_{PG}^{(n+1)}(t) = \frac{\cosh^{\xi}(\omega/2)}{4^n}\sum_i \Bigg(& \frac{P_{i,1}^{n+1}[\xi]\,Q_{i,1}^{n+1}[\cosh g_\omega(t), \sinh g_\omega(t)]}{\cosh^{\xi+p_i+1} g_\omega(t)\; g_\omega(t)^{d_i}} \\ & + \frac{P_{i,2}^{n+1}[\xi]\,Q_{i,2}^{n+1}[\cosh g_\omega(t), \sinh g_\omega(t)]}{\cosh^{\xi+p_i} g_\omega(t)\; g_\omega(t)^{d_i}} \\ & + \frac{P_{i,3}^{n+1}[\xi]\,Q_{i,3}^{n+1}[\cosh g_\omega(t), \sinh g_\omega(t)]}{\cosh^{\xi+p_i} g_\omega(t)\; g_\omega(t)^{d_i+1}}\Bigg)\left(\frac{1}{4g_\omega(t)}\right) , \end{aligned} \tag{61}$$

where,

$$\begin{aligned} P_{i,1}^{n+1}[\xi] &= -P_i^n[\xi](\xi+p_i) & Q_{i,1}^{n+1} &= Q_i^n[\cosh g_\omega(t), \sinh g_\omega(t)]\sinh g_\omega(t) \\ P_{i,2}^{n+1}[\xi] &= P_i^n[\xi] & Q_{i,2}^{n+1} &= \frac{d}{dg_\omega}Q_i^n[\cosh g_\omega(t), \sinh g_\omega(t)] \\ P_{i,3}^{n+1}[\xi] &= -P_i^n[\xi]d_i & Q_{i,3}^{n+1} &= Q_i^n . \end{aligned} \tag{62}$$

Note that as $\cosh'(x) = \sinh(x)$ and vice-versa, $Q^{n+1}_{i,2}$ is of the same degree as Q^n_i. Thus, all the degree conditions (60) are preserved for $n+1$, completing the induction. Note in particular that induction on the first term of the expansion in (61) leads to the only term that satisfies the conditions with equality: $d(Q^n_{1,1,\ldots}) = d(P^n_{1,1,\ldots}) = p_1 = n = d_{1,1,\ldots}$. This term has the form:

$$\cosh^{\xi}(\omega/2)\frac{(-1)^n(\xi(\xi+1)(\xi+2)\ldots(\xi+n-1))\sinh^n g_\omega(t)}{4^n\cosh^{\xi+n} g_\omega(t)\; g_\omega(t)^n} \tag{63}$$

Evaluating (58) at $t=0$ yields:

$$\begin{aligned} f^{(n)}_{PG}(0) = & \frac{(-1)^n(\xi(\xi+1)(\xi+2)\ldots(\xi+n-1))\sinh^n(\omega/2)}{\cosh^n(\omega/2)\;(2\omega)^n} \\ & + \sum_i \frac{P^n_i[\xi]Q^n_i[\cosh(\omega/2),\sinh(\omega/2)]}{4^n\text{cosh}^{p_i}(\omega/2)\;(\omega/2)^{d_i}}\,. \end{aligned} \tag{64}$$

where we have separated the leading term from the remainder of the sum.

B.4 Asymptotic behaviour

We recall that the mean and variance of the NB model are given by:

$$\mathbb{E}(\mathcal{X}|\zeta,\mathcal{W},\mathcal{V}) = \zeta e^{\mathcal{W}+\mathcal{V}} \quad \text{and} \quad \mathbb{V}(\mathcal{X}|\zeta,\mathcal{W},\mathcal{V}) = \zeta e^{\mathcal{W}+\mathcal{V}}(1+e^{\mathcal{W}+\mathcal{V}})\,, \tag{65}$$

We can therefore define:

$$\begin{aligned} FF &= \langle 1+e^{\mathcal{W}+\mathcal{V}}\rangle_q \\ &= 1+e^{\langle\mathcal{W}+\mathcal{V}\rangle+\frac{1}{2}\left(\langle(\mathcal{W}+\mathcal{V})^2\rangle-\langle\mathcal{W}+\mathcal{V}\rangle^2\right)}\,. \end{aligned} \tag{66}$$

Now consider a single $q(\mathcal{U}_{\bar{d}}) = \text{PG}(\xi,\omega)$ with $\xi = \zeta + \mathcal{X}_{\bar{d}}$ and $\omega = \Omega_{\bar{d}}$. Based on the argument of Appendix B.1 in the "good fit" limit, $\zeta \approx \mathcal{X}_{\bar{d}}/(FF-1)$ and $\langle(\mathcal{W}_{\bar{d}}+\mathcal{V}_{\bar{d}})^2\rangle = \mathcal{O}\left(\langle\mathcal{W}_{\bar{d}}+\mathcal{V}_{\bar{d}}\rangle^2\right)$, so:

$$\begin{aligned} \omega &= -\langle\mathcal{W}_{\bar{d}}+\mathcal{V}_{\bar{d}}\rangle\sqrt{1+\frac{\langle(\mathcal{W}_{\bar{d}}+\mathcal{V}_{\bar{d}})^2\rangle-\langle\mathcal{W}_{\bar{d}}+\mathcal{V}_{\bar{d}}\rangle^2}{\langle\mathcal{W}_{\bar{d}}+\mathcal{V}_{\bar{d}}\rangle^2}} \\ &= -\langle\mathcal{W}_{\bar{d}}+\mathcal{V}_{\bar{d}}\rangle+\frac{\langle(\mathcal{W}_{\bar{d}}+\mathcal{V}_{\bar{d}})^2\rangle-\langle\mathcal{W}_{\bar{d}}+\mathcal{V}_{\bar{d}}\rangle^2}{\langle\mathcal{W}_{\bar{d}}+\mathcal{V}_{\bar{d}}\rangle}+\mathcal{O}\left(\frac{(\langle(\mathcal{W}_{\bar{d}}+\mathcal{V}_{\bar{d}})^2\rangle-\langle\mathcal{W}_{\bar{d}}+\mathcal{V}_{\bar{d}}\rangle^2)^2}{\langle\mathcal{W}_{\bar{d}}+\mathcal{V}_{\bar{d}}\rangle^3}\right) \\ &= -\log(FF-1)+\frac{1}{2}\Big(\langle(\mathcal{W}_{\bar{d}}+\mathcal{V}_{\bar{d}})^2\rangle-\langle\mathcal{W}_{\bar{d}}+\mathcal{V}_{\bar{d}}\rangle^2\Big)+\mathcal{O}\left(\frac{\langle(\mathcal{W}_{\bar{d}}+\mathcal{V}_{\bar{d}})^2\rangle-\langle\mathcal{W}_{\bar{d}}+\mathcal{V}_{\bar{d}}\rangle^2}{\langle\mathcal{W}_{\bar{d}}+\mathcal{V}_{\bar{d}}\rangle}\right) \\ &= -\log(FF-1)\left(1+\underset{FF\to 1}{\mathcal{O}}(1)\right) \end{aligned} \tag{67}$$

As a conclusion, although both ω and ξ will grow as $FF \to 1$, we have:

$$\omega = \log\xi\left(1+\underset{FF\to 1}{\mathcal{O}}(1)\right)\,, \tag{68}$$

which we will now use to look at the higher moments of $PG(\omega,\xi)$ and Gamma(α,β). For the latter, from (57), we have:

$$\begin{aligned} m^n_G = (-1)^n f^{(n)}_G &= \frac{\alpha(\alpha+1)\ldots(\alpha+n-1)}{\beta^n} \\ &= \left(\frac{\alpha}{\beta}\right)^n\left(1+\mathcal{O}\left(n^2\cdot\frac{\alpha}{\beta^2}\cdot\left(\frac{\alpha}{\beta}\right)^{-2}\right)\right) \\ &= \left(\frac{\xi}{2\omega}\right)^n\left(\tanh^n(\omega/2)+\mathcal{O}\left(\frac{n^2}{\xi\omega}\frac{\sinh(\omega)-\omega}{\sinh^2(\omega/2)}\right)\right) \\ &= \left(\frac{\xi}{2\omega}\right)^n+\underset{FF\to 1}{\mathcal{O}}\left(\frac{\xi^{n-1}}{\omega^n}(1+\xi e^{-\omega})\right)\,, \end{aligned} \tag{69}$$

giving a leading term of $(\xi/2\omega)^n$ as $FF \to 1$.

For the PG distribution (64) gives (recalling that $d(Q^n_i) \le d(P^n_i) = p_i \le n \le d_i$ with all equalities holding only for the leading term)

$$\begin{aligned} m^n_{PG} = (-1)^n f^{(n)}_{PG} &= (-1)^n \sum_i \frac{P^n_i[\xi]Q^n_i[\cosh(\omega/2), \sinh(\omega/2)]}{4^n \cosh^{p_i}(\omega/2)\ (\omega/2)^{d_i}} \\ &= \left(\frac{\xi}{2\omega}\right)^n (\tanh^n(\omega/2)) + \mathcal{O}\left(\frac{n^2 3^n}{\omega}\left(\frac{\xi}{2\omega}\right)^{n-1}\right) \\ &= \left(\frac{\xi}{2\omega}\right)^n + \underset{FF\to 1}{\mathcal{O}}\left(\frac{\xi^{n-1}}{\omega^n}(1+\xi e^{-\omega})\right), \end{aligned} \tag{70}$$

Thus, we see that both sets of higher-order moments are equivalent in the leading terms of their expansions.

B.5 Numerical evaluation of the moment matching approach.

In this section, we report a range of numerical examples and experiments to evaluate the moment matching approximation and its asymptotic behavior.

We first compared the accuracy with which the Pólya-Gamma density could be approximated using moment-matched gamma, generalised gamma, inverse gamma and normal distributions. We exploited the convolutional property of the PG distribution to obtain numerical density estimates. Each sample-based density was estimated with at least 100000 draws using the bayesreg toolbox [28]. Visually, the gamma distribution provides the best approximation, following the sampled PG density closely for $\xi \ge 10$ (Fig. 5).

We quantified the match using the normalised L_2 discrepancy, obtained analytically from the PG, gamma and normal characteristic functions; that is, for $X \in \{G, IG, N\}$,

$$d(f_{PG}, f_X) = \frac{2||p_{PG} - p_X||^2_{L_2}}{||p_{PG}||^2_{L_2} + ||p_X||^2_{L_2}}\,. \tag{71}$$

This L_2 discrepancy was consistently smallest for the gamma-based approximation (Fig. 6).

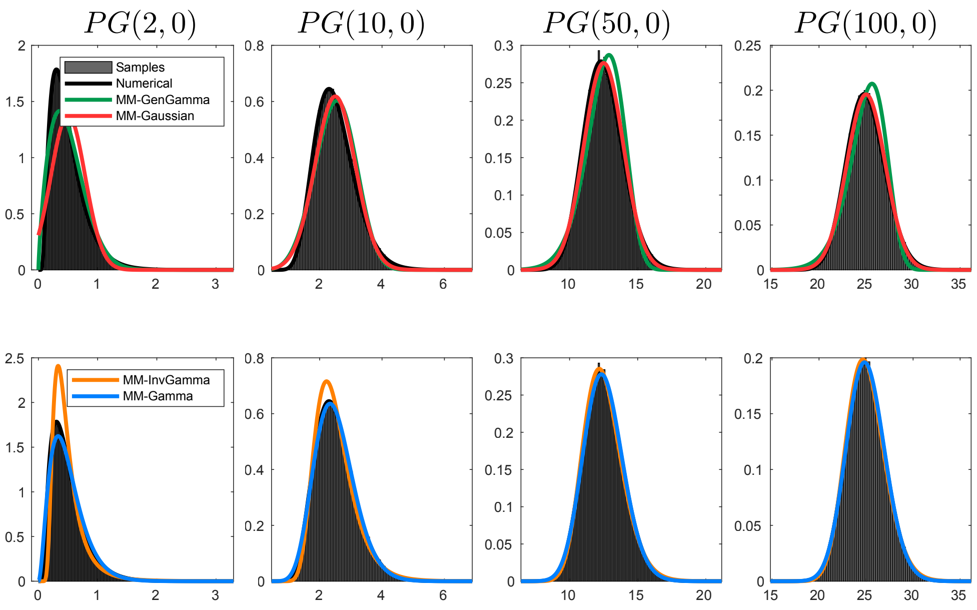

Figure 5: Densities of gamma, inverse gamma, normal and generalised gamma distributions moment-matched to PG(ξ, 0) for different values of ξ (In the generalized gamma case, the first three moments were matched).

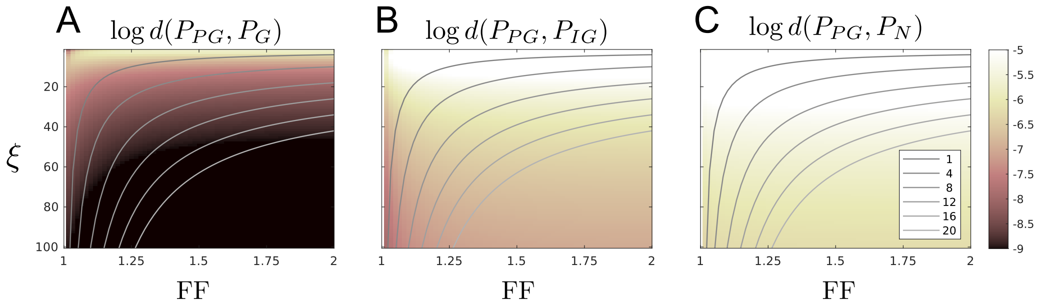

Figure 6: Normalized L_2 norm between p_{PG} and p_X for $X \in \{G, IG, N\}$ as a function of the Fano factor $\text{FF} = 1 + e^{-\omega}$ and ξ. The lines indicate the expected relationship between ξ and FF for different numbers of observed spikes.

Finally, we compared numerical estimates of the KL divergences between PG distributions in non diverging regions, to the closed form divergence between the moment-matched gamma approximations (Fig. 7; comparisons are anchored on $PG(50, 1)$). We found the match to be good, particularly when the divergence was relatively small.

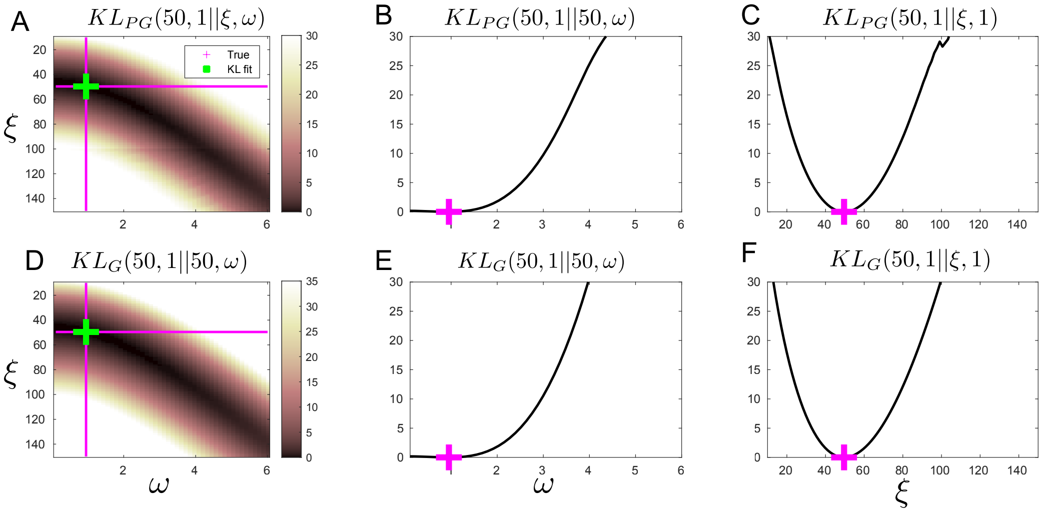

Figure 7: KL divergence between PG distributions $PG(50, 1)$ and $PG(\xi, \omega)$ obtained by numerical estimation (A-C) and moment matching mean and variance of Gamma distribution (D-E). Crosses indicate the minimum of the KL.

C Metrics

Performances are reported in terms of Variance Explained (VE), deviance explained (DE) and a similarity metric, which we define here. If $\mathcal{X}_0$ is the average spike count of the dataset across all conditions, neurons and time points, and $\hat{\mathcal{X}}$ a reconstructed tensor from the decompositions estimates, the VE and Poisson DE are given by

$$VE = 1 - \frac{\sum_{\bar{d}} (\hat{\mathcal{X}}_{\bar{d}} - \mathcal{X}_{\bar{d}})^2}{\sum_{\bar{d}} (\mathcal{X}_0 - \mathcal{X}_{\bar{d}})^2} \quad \text{and} \quad DE = 1 - \frac{\sum_{\bar{d}} (\mathcal{X}_{\bar{d}} \log \mathcal{X}_{\bar{d}} / \hat{\mathcal{X}}_{\bar{d}} + \hat{\mathcal{X}}_{\bar{d}} - \mathcal{X}_{\bar{d}})}{\sum_{\bar{d}} (\mathcal{X}_{\bar{d}} \log \mathcal{X}_{\bar{d}} / \mathcal{X}_0 + \mathcal{X}_0 - \mathcal{X}_{\bar{d}})} \,. \tag{72}$$

Although we assessed similarities across both cross-validation folds and initializations, we used the same metric as employed in Ref. [7]. Given two CP [9] decompositions $A = [|A^{(1)}, \dots, A^{(D)}|]$ and $V = [|V^{(1)}, \dots, V^{(D)}|]$, we normalize each column $A^{(n)}_{0,:r} = A^{(n)}_{:r} / ||A^{(n)}_{:r}||_{L_2}$ and $V^{(n)}_{0,:r} = V^{(n)}_{:r} / ||V^{(n)}_{:r}||_{L_2}$ and gather the normalization constants across modes in γ^A_r and γ^V_r. In essence, the similarity metric is a generalized scalar product ($\in [0, 1]$) between factors which accounts for possible permutations ($\sigma \in \sigma_R$) of the components:

$$S = \max_{\sigma \in \sigma_R} \sum_{r=1}^{R} \left(\frac{|\gamma^A_r + \gamma^V_{\sigma(r)}| - |\gamma^A - \gamma^V_{\sigma(r)}|}{|\gamma^A + \gamma^V_{\sigma(r)}| + |\gamma^A_r - \gamma^V_{\sigma(r)}|} \right) \prod_{n=1}^{D} A^{(n)\intercal}_{0,:r} V^{(n)}_{0,:\sigma(r)} \,, \tag{73}$$

We used an efficient search over permutation by leveraging Munkres (Hungarian) algorithm for the linear assignment problem.

D Qualitative Experiments

Figures 8 to 10 show the results of two qualitative experiments described in Section 4.1. Briefly, in both cases, we generated a 5 dimensional dataset (size $100 \times 70 \times 3 \times 5 \times 4$) using a generative NB model with shape $\zeta = 80$ and $R = 4$. Each CP-component loads only on a subsets of "neurons". We assumed that the subsets were known and used them to define neuron-group priors (section 3.1).

In the first experiment, we added an offset tensor that varies across the first and third dimensions and treated only 1/4 of the full tensor $\mathcal{X}$ as observed. In this stitching setting, each "neuron" is observed in a single "experimental session", which is accounted by the last dimension (not plotted Figure 8) of the observed tensor.

The algorithm converged over about 4000 iterations (Fig. 8) and was able to retrieve the dataset shape parameters and infer both the tensor factors and rank in a fully probabilistic manner (Fig. 9 left). When the algorithm was run from the same initial values, but without ARD or neuron group constraints it inaccurately identified additional components with similar amplitudes to the true factors (Fig. 9 right).

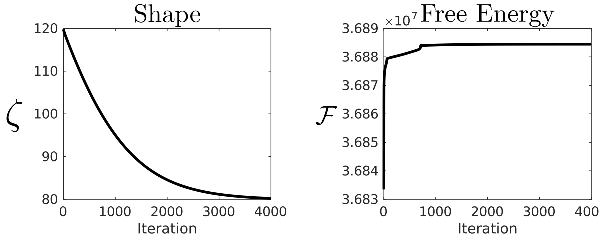

Figure 8: Evolution of the shape parameter (Left) and approximate free energy (Right) across variational EM iterations for probabilistic decomposition of the count tensor from Section 4.1. Inference used ARD and knowledge about the "neuron" groups.

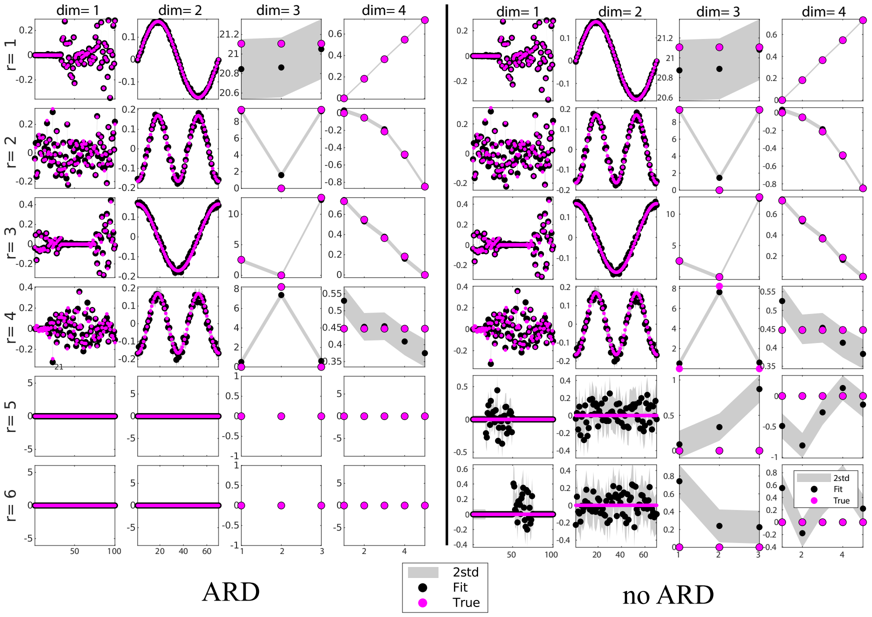

Figure 9: Probabilistic decomposition of the count tensor from Section 4.1. Ground truth generating values are in pink, estimates in black; grey shading represents 2 standard deviations around the mean based on variational posteriors. Each row represents an estimated CP component, with each column reflecting one dimension. We compare inference results using ARD and knowledge about the "neuron" groups (left) or not (right).

In the second experiment, we illustrate how standard GCP decompositions are prone to misidentify the dataset rank and how this can affect the decomposition itself. We removed the offset and tested our method against GCP with NB observation model [10] for a range of shape parameter ζ (GCP-NB required an additional hyperparameter selection step). We report the error (defined as the squared Euclidean distance between the reconstructed tensor $\hat{\mathcal{X}}$ and the true $\mathbb{E}(\mathcal{X}|\zeta, \mathcal{W})$) and the similarities between decompositions obtained with 20 random initializations for each algorithm and each putative rank. Results are reported Figure 10. The decompositions that are plotted maximize the similarity metric across the initialization sets. With standard GCP method, it is not straightforward to balance goodness of fit with robustness. Although the results suggest that all models require at least 4 components to accurately reconstruct the simulated dataset, one might need to fix an arbitrary thresholds to select the tensor rank, which can lead to incorrect rank selection or bad decomposition (Figure 10 top and bottom right).

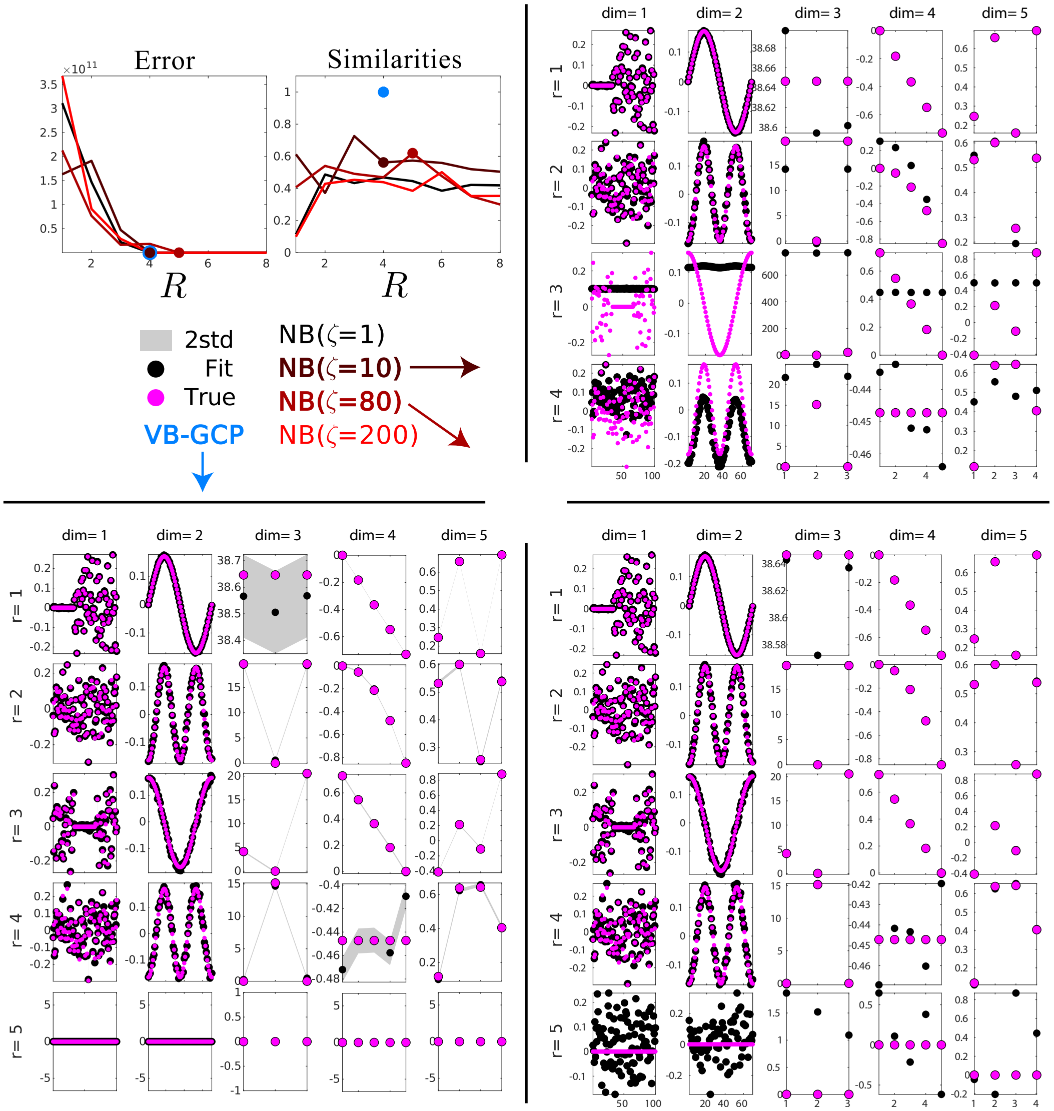

Figure 10: Goodness of Fit and Similarities (Top Left) are often balanced to select the rank of an empirical tensor decomposition. As a consequence, standard GCP (Right: top and bottom) decomposition might suffer from rank (or other hyperparameter) misidentification. Our probabilistic method (VB-GCP) provides a principled way to select the tensor rank, incorporate prior knowledge about the data and it estimates factors posterior. The decompositions plotted here maximize the similarity metric across 20 random initializations for a given rank.

E Spike Recordings

In this section, we detail the results of the benchmark analysis described in Section 4.2. Training and testing results are plotted in Figure 11 and Figures 12 to 14 show the full VB-GCP (ours), CP and GCP decompositions described in Section 4.2. As mentioned in the main text, we also tested NB observation models with exponential link functions. Hong et al. [10] worked with a fixed shape parameter for the NB likelihood. As the most appropriate parameter value for our data was unknown, we ran a gridsearch using 50 shape parameters ranging from 1 to 100 and looked at the best models in terms of DE, Robustness (as assessed by the similarity metric) or NB likelihood. NB-GCP turned out to perform little or no better than Poisson GCP.

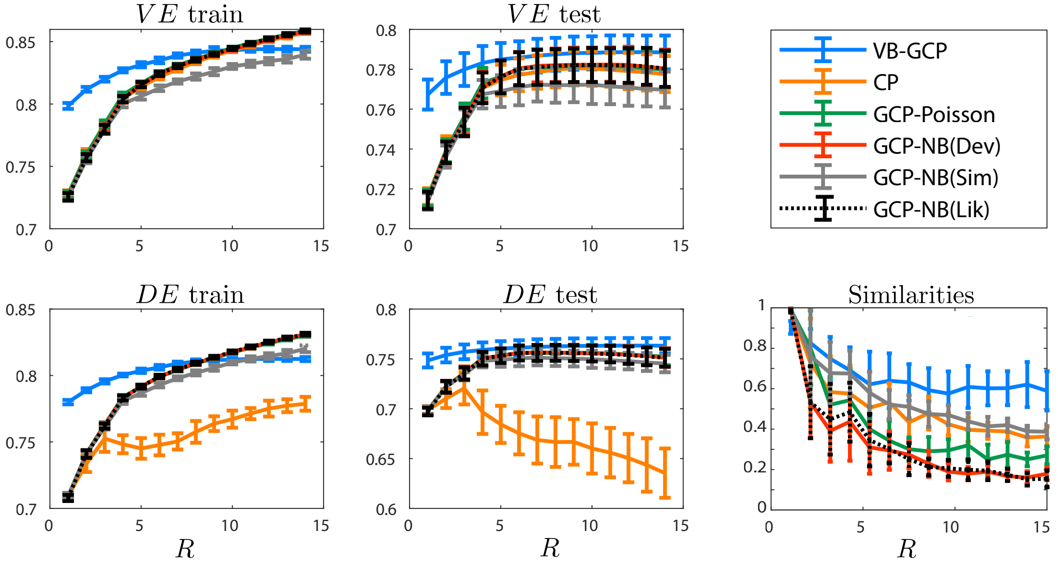

Figure 11: Train and Test variance and deviance explained for VB-GCP (ours), CP, Poisson and NB GCP as a function of the model Tensor rank. NB models plotted here were selected by maximizing similarities (grey), test DE (red) or test likelihood (black).

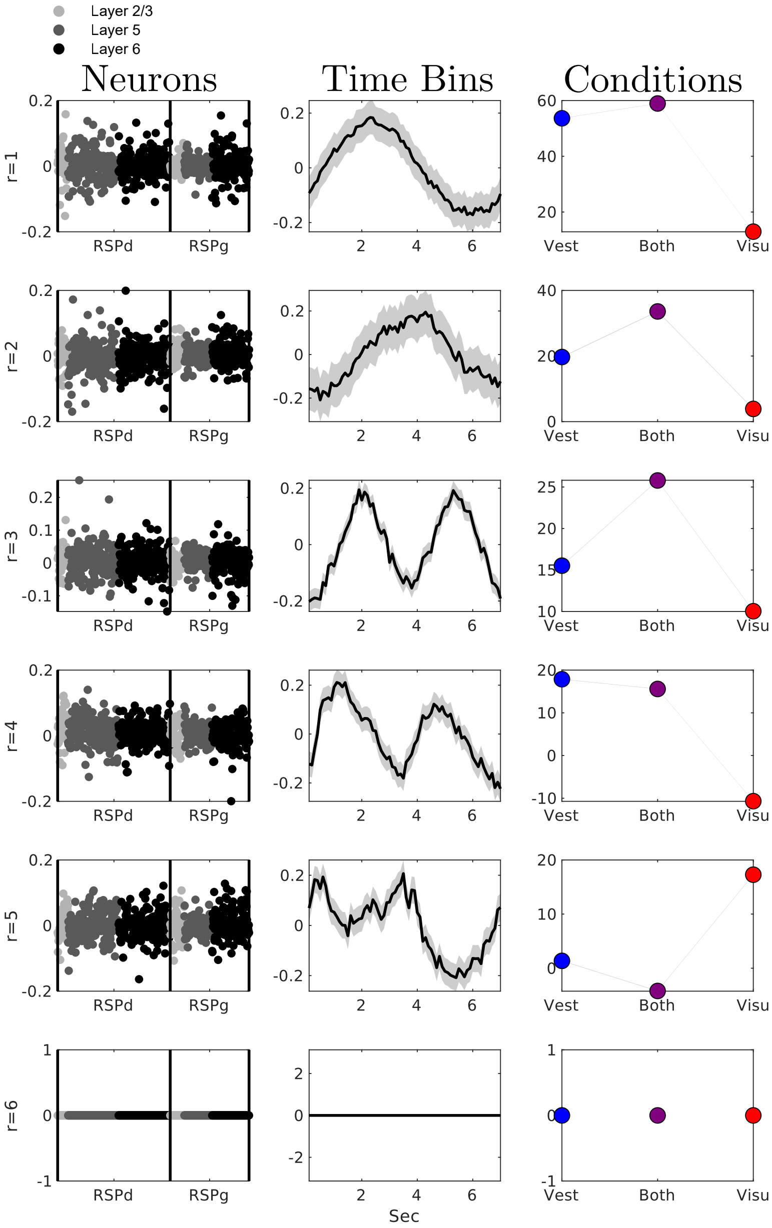

Figure 12: Full inferred factors obtained with our probabilistic decomposition. Grey patches represent 1 standard deviation around the mean based on variational posteriors.

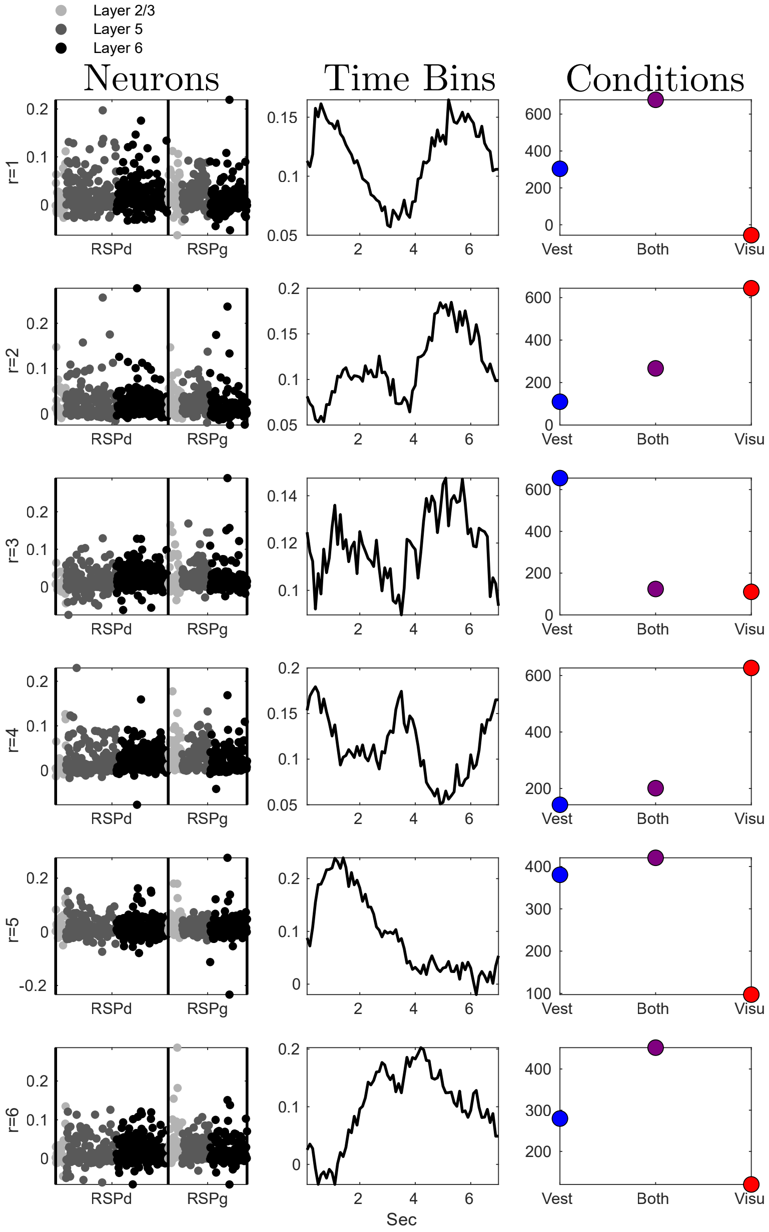

Figure 13: Full inferred factors obtained with standard CP decomposition.

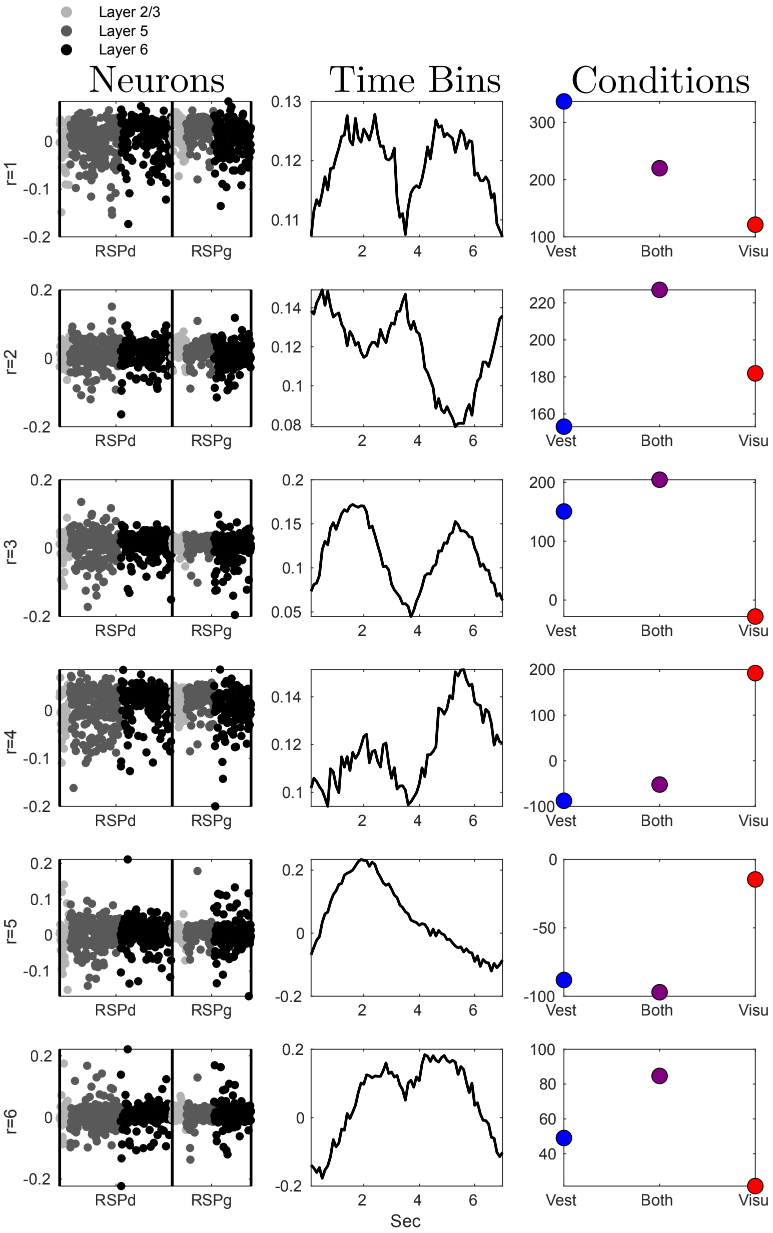

Figure 14: Full inferred factors obtained with standard GCP decomposition.