Probabilistic Tensor Decomposition of Neural Population Spiking Activity

Anonymous Author(s)
Affiliation
Address
email

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 recordings offer 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 fully leveraged if we are able to disentangle the underlying factors that contribute to neural activity and extract them in a simple and interpretable way. One way to approach underlying structure within multifactorial data is based on the decomposition of measurements structured in a multidimensional array or tensor. Yet, despite the prevalence of tensor decomposition techniques in many fields like chemometrics [1] or computer vision [2], their use in neuroscience has historically been limited to the study of continuous traces like EEG [3], fMRI signals [4] or LFP [5]. In recent years, attention has turned to neuronal spike, and tensor methods have proven useful to segregate the influence of time dynamics, trial to trial variability [6] [7] or experimental condition [8], even though the models employed have not provided a full probabilistic treatment of spike counts.

Canonical Polyadic (CP) decomposition [9] is an increasingly wide-spread technique in which a tensor is decomposed into a sum of rank-1 elements. CP decomposition has been generalized to non-Gaussian loss functions [10] [11], but 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

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

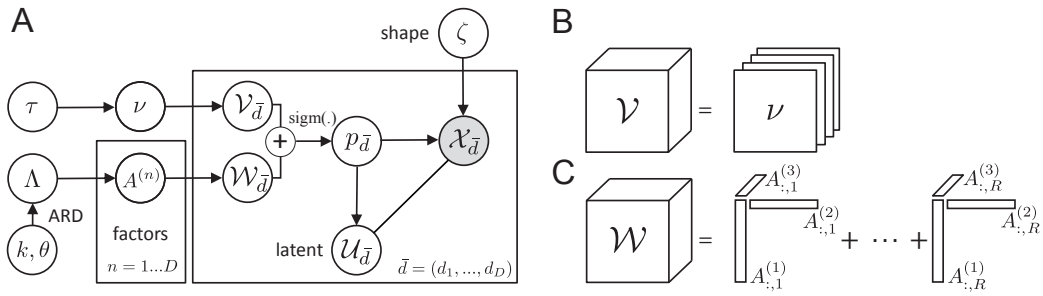

Figure 1: Probabilistic tensor decomposition model. (A) Graphical model. (B) The Offset tensor $\mathcal{V}$ is constrained to vary along a limited set of dimensions. (C) $\mathcal{W}$ is a low rank tensor.

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 [12]. This solution enables Gibbs sampling in models which combine Gaussian latent structure with logistic, Poisson or negative binomial observations, including factor models [13] and generalised-linear regression [12], including within neuroscience [14]. While PG-augmentation has also been proposed for tensor decomposition [11], the typical scale of the problems renders Gibbs sampling impractical and so the authors focused primarily on point estimation while incorporating a limited range of priors.

Here, we introduce a variational approach to Pólya-Gamma augmented models with negative-binomial observations, which enables a relatively efficient Variational Bayesian (VB) treatment. Our approach is based on a principled approximation of Pólya-Gamma cross entropy, which is especially well suited for neural datasets in which conditional Fano factors (FF) (once the contribution of common stochastic factors is discounted) are often close to one [15] [16]. The fully probabilistic formulation is able to handle missing observations, including in 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 Pólya-Gamma 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 use synthetic data and neural recordings [18] in mice performing passive multisensory integration and show that our method can probabilistically disentangle the contribution of neurons, temporal dynamics and experimental condition on spiking activity. We show improved performances in terms of variance explained, deviance and, last but not least, decomposition robustness compared to standard CP and GCP baselines.

2 Background

2.1 Tensor Decomposition

We observe spike counts gathered in a D dimensional tensor:

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

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 dynamical structure in the 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 $\mathcal{W} = [|A^{(1)}, \dots, A^{(D)}|]$ and 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 $a_i^{(n)} = [A^{(n)}_{i1}, \dots, A^{(n)}_{iR}]$. 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 $\bar{d}^* = (d^*_1, \dots, d^*_{D_*})$ and $\bar{d}^\bullet = (d^\bullet_1, \dots, d^\bullet_{D-D_*})$ regardless of the dimension ordering. Instead of a $I_1 \times \dots \times I_D$ tensor, in practice we hence need to estimate the smaller 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 link $\mathcal{W}$ and $\mathcal{V}$ to the observed data $\mathcal{X}$ through a negative binomial model.

2.2 Negative-binomial distribution

The negative binomial model can be seen as 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 firing rates $\mu_{\bar{d}}$ then yields a negative binomial (NB) distribution [14] :

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

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 Pólya-Gamma (PG) augmentation [12]. We first recall key properties of PG distributions before deriving the augmented model likelihood.

2.3.1 Definition and properties

For $\xi > 0$ and $\omega \in \mathbb{R}$, a random variable U is PG(ξ, ω) distributed i.i.f:

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

where $g_k \sim \text{Gamma}(\xi, 1)$. Three major results from [12] 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.

(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 $\text{PG}(\xi, \omega)$ and $\text{PG}(\xi, 0)$ 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 those densities can be expressed as alternating-sign sums of inverse gamma distributions, the terms within the sums diverge rapidly as ξ increases. In what follows, we will instead work with the Laplace transform of the PG distribution, given by:

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

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

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

with similar closed-form expressions for higher-order moments, which we will use to justify our inference procedure.

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

The Pólya-Gamma augmentation approach was developed to facilitate Gibbs sampling in otherwise non-conjugate models. However, despite the availability of closed-formed updates, sampling quickly becomes impractical for large datasets [11]. Thus, in this section, we propose a variational Bayes estimation procedure and report the associated variational 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 $P(\mathcal{U}_{\bar{d}}|\mathcal{X}_{\bar{d}}) = \text{PG}(\zeta + \mathcal{X}_{\bar{d}}, 0)$. 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 $\bar{d}^* = (d_1^*, \ldots, d_{D_*}^*)$. Finally, for $n = 1 \ldots D$, $i = 1 \ldots I_n$ 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.

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. In such a case:

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

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 of a same group. If neuron factors are gathered in $A^{(n)}$, we denote by $g(i) \in \mathcal{G}$ the group of neuron i and define:

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

In this case, divergence of a subset of precisions during optimization will lead to components linked to separate neuron groups to shrink away. Thus, the model will tend to converge on an explanation of the activity of each group using as few components as possible.

3.2 Variational Distribution

We are now equipped to write the full variational distribution. For $\mathcal{Z} = \{\mathcal{U}, \mathcal{V}, \mathcal{W}, \Lambda\}$, it is given by:

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

We 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 marginalized over x, 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 omit explicit statement of the marginal to simplify notation).

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 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:

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

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

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

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

where the parameters k_λ and θ_λ depends on prior (Rank ARD / Neuron Group constraint) type.

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

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

where we defined:

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

and:

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

Update PG latent $\mathcal{U}$. For $\bar{d} = (d_1, \ldots, d_D)$, 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 $|\mathcal{X}|$ such estimates, 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 that applies as the conditional Fano factor approaches 1.

Denote by m^n_{PG} and m^n_G the nth moments respectively of PG(ξ, ω) and of the Gamma distribution with matched mean and variance. In the learnt model, the parameters ξ and ω are linked by inferred mean of the observed count. As ω is also explicitly linked to the inferred conditional Fano factor, both parameters diverge as FF $\to 1$ with $\xi e^{-\omega}$ remaining finite. As shown in the Supplementary Materials B, 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 polynomial expansions of m^n_{PG} and m^n_G are both equal to $\left(\frac{\xi}{2\omega}\right)^n$ and so the higher-order moments 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. Thus, we further 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:

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

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.4). 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 Pólya-Gamma 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.

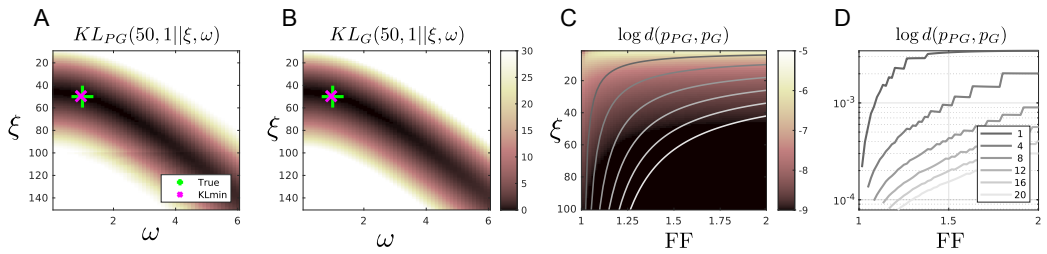

Figure 2: Moment-Matching Pólya-Gamma (PG) and Gamma distribution in the unity fano-factor (FF) limit. (A) Numerical estimates of KL divergences between PG$(50, 1)$ and PG(ξ,ω). (B) Moment matched Gamma KL values. $KL_G(50, 1||\xi, \omega)$ indicates that the we estimate $KL_{PG}(50, 1||\xi, \omega)$ 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 $= 1 + e^{-\omega}$, $d(f_{PG}, f_{IG}) = 2||p_{PG} - p_{IG}||^2_{L_2}/(||p_{PG}||^2_{L_2} + ||p_{IG}||^2_{L_2}))$. The plain traces indicate relationship between ξ and FF for different number of observed spikes. (D) Values of $d(f_{PG}, f_{IG})$ along such traces.

4 Experiments

In this section, we first describe a qualitative results on a simulated toy dataset demonstrating the ability of our algorithm to infer tensor rank and factor posteriors in a missing data situation. We then benchmark it against standard CP [9] and GCP [10] decomposition on neural spike data. We report performances 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]

4.1 Simulations

We generated a 5 dimensional dataset (size $100 \times 70 \times 3 \times 4 \times 5$) using the generative NB model described in (6). The shape parameter was fixed to $\zeta = 80$, the generative tensor rank was $R = 4$ and the offset was constrained to vary only across the first and third dimension. Assuming that the first dimension accounts for neuron loadings, we simulated "neuron groups" by having each CP-component load only on some subsets of the neurons. The subsets were assumed to be known, but not the pattern of the loadings. Only 1/4 of the full tensor was observed. Supplementary Figures 8 and 9 show that the algorithm initialized with $R = 6$ components probabilistically infers the correct tensor decomposition as well as the shape parameter ζ. Automatic relevance determination shrinks extra components which effectively estimate the rank of the data tensor. For comparison, using the exact same initialization but without ARD, two extra components are identified (See Supplementary Figure 10).

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 and sensory 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

[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.

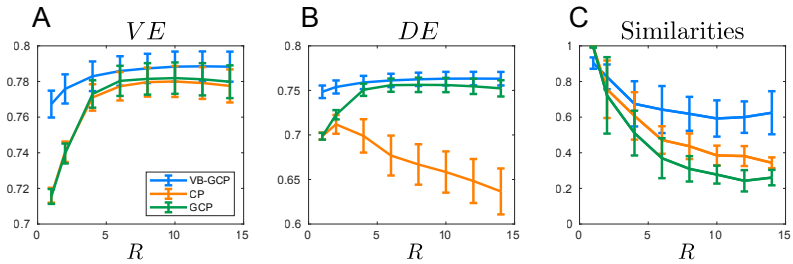

Figure 3: Held out (trial based) performances as a function of the tensor rank R between our probabilistic model (VB-GCP), CP and GCP. (A) Variance Explained. (B) Poisson Deviance explained. (C) Similarity metric. Error bars indicate one standard deviation.

was adapted such that the visual motion component under 'both' and 'visual' conditions are 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 across trials. The algorithms are then trained on the summed spikes over half of the trials, and their performances are 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. All algorithms where trained with a maximum of 10000 iterations. For VB-GCP, we used $k_\lambda = 100$, $\theta = 1$ (see equation 16), neuron groups (see equation (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 Supplementary Section D). The latter is essentially a scalar product between CP factors, which estimates how close two decomposition are (from 0 to 1) and accounts for possible permutations of the components. Importantly, here, we not only estimate similarities across initializations, but also across across validation fold, which is a much more stringent test that allows us to compare how robust is the decomposition from one experimental trial to another.

We find that our algorithm outperforms (G)CP even though VE and DE explicitly assess the methods reconstructing performances under the Gaussian and Poisson likelihood losses assumed by those methods (Figure 3 A-B). Perhaps more importantly, the similarity metric (Figure 3 C) does not collapse when increasing the rank of the fitted tensor, which demonstrates that our probabilistic treatment of the factor precision matrices increases the reliability of the fitting procedure. We then compare models in terms of interpretability. 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. 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). Our analysis therefore 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

[2]https://www.tensortoolbox.org

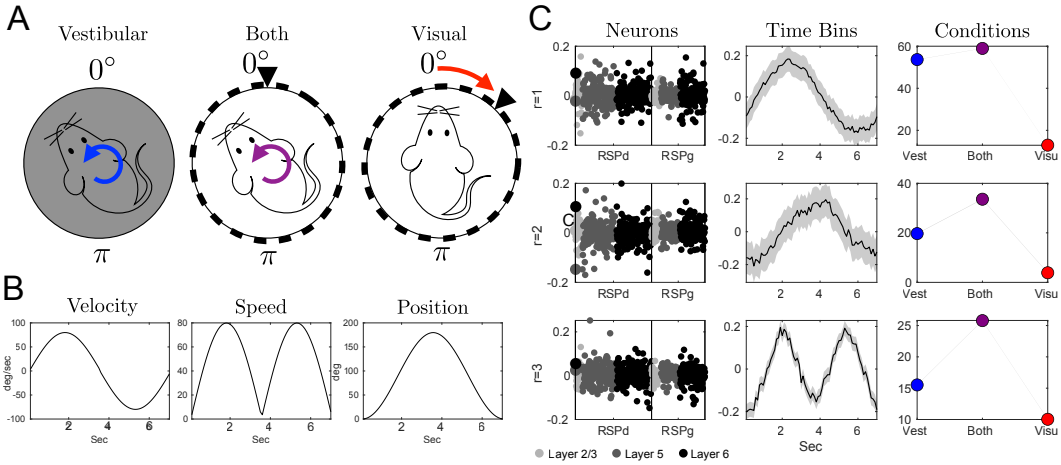

Figure 4: Neural data analysis. (A) Experimental conditions: the mouse is rotated in the dark (Vestibular), rotated in presence of a surround static visual scene (Both), or is stationary while the surrounding visual scene is being rotated (Visual). (B) Temporal profiles of the rotation stimuli all conditions.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

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 possible source of these advantages 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 Pólya-Gamma augmentation schemes [12–14] and extends it to a Bayesian treatment of tensor decomposition. Although the Pólya-Gamma approach was also adopted in [11], 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 Pólya-Gamma KL divergences with their Gamma moment-matched counterparts in regions where we expect them to be well behaved. Nevertheless, although our method provides approximate posteriors over factors, as with many mean-field or factored variational inference procedures, it is expected to underestimate the true uncertainty in the model variables. Possible extensions could involve low-rank approximations to the posterior covariance between factors, or the extension of linear-response methods [21] to PG-augmented models that lie outside the exponential family.

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

[12] Nicholas G Polson, James G Scott, and Jesse Windle. Bayesian inference for logistic models using Pólya-Gamma latent variables. Technical report.

[13] Arto Klami, Dinh Phung, and Hang Li. Polya-Gamma augmentations for factor models. Technical report, feb 2014.

[14] Jonathan W Pillow and James G Scott. Fully Bayesian inference for neural models with negative-binomial spiking. Technical report, 2012.

[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. Technical report, 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. 13(3):369–378, 2010.

[21] 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.

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

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] Yes, see section 5
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Supplementary and section 3
   (b) Did you include complete proofs of all theoretical results? [Yes] See Supplementary
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Both code and data are in Supplementary Materials. Code will be made available upon publication but the dataset is not intended to be openly shared.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See section 4
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See section 4
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See section 4.2.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See section 1
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A] No new asset
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]