The Annals of Applied Statistics
2017, Vol. 11, No. 2, 598–637
DOI: 10.1214/16-AOAS996
© Institute of Mathematical Statistics, 2017

ROBUST AND SCALABLE BAYESIAN ANALYSIS OF SPATIAL NEURAL TUNING FUNCTION DATA

BY KAMIAR RAHNAMA RAD[1],
TIMOTHY A. MACHADO[2] AND LIAM PANINSKI[3]

City University of New York, Cognescent Corporation and Columbia University

A common analytical problem in neuroscience is the interpretation of neural activity with respect to sensory input or behavioral output. This is typically achieved by regressing measured neural activity against known stimuli or behavioral variables to produce a “tuning function” for each neuron. Unfortunately, because this approach handles neurons individually, it cannot take advantage of simultaneous measurements from spatially adjacent neurons that often have similar tuning properties. On the other hand, sharing information between adjacent neurons can errantly degrade estimates of tuning functions across space if there are sharp discontinuities in tuning between nearby neurons. In this paper, we develop a computationally efficient block Gibbs sampler that effectively pools information between neurons to denoise tuning function estimates while simultaneously preserving sharp discontinuities that might exist in the organization of tuning across space. This method is fully Bayesian, and its computational cost per iteration scales subquadratically with total parameter dimensionality. We demonstrate the robustness and scalability of this approach by applying it to both real and synthetic datasets. In particular, an application to data from the spinal cord illustrates that the proposed methods can dramatically decrease the experimental time required to accurately estimate tuning functions.

1. Introduction. Over the past five years, it has become possible to simultaneously record the activity of thousands of neurons at single-cell resolution [Ahrens et al. (2013), Hamel et al. (2015), Portugues et al. (2014), Prevedel et al. (2014)]. The high spatial and temporal resolution permitted by these new methods allows us to examine whether previously unexamined regions of the brain might dynamically map sensory information across space in unappreciated ways. However, the high dimensionality of these data also poses new computational challenges for statistical neuroscientists. Therefore, scalable and efficient methods for extracting as much information as possible from these recordings must be devel-

Received July 2015; revised October 2016.

[1] Supported in part by the PSC-CUNY Award 67377-0045.

[2] Supported in part by the NSF GRFP.

[3] Supported in part by the Gatsby Foundation, NSF CAREER award and Grants ARO MURI W911NF-12-1-0594, NSF CAREER IOS-0641912, ONR N00014-14-1-0243 and DARPA N66001-15-C-4032.

Key words and phrases. Computational neuroscience, scalability and robustness, Bayesian, posterior sampling, spatial statistics.

oped; in turn, improved analytical approaches that can extract information from, for example, shorter experiments, may enable new dynamic closed-loop experimental designs.

In many experimental settings, a key quantity of interest is the tuning function, a filter that relates known information about sensory input or behavioral state to the activity of a neuron. For example, tuning functions permit measurement of orientation selectivity in the visual cortex [Hubel and Wiesel (1968)], allow us to relate movement direction to activity in the primary motor cortex [Georgopoulos, Kettner and Schwartz (1986), Scott (2000)], and let us measure the grid-like spatial sensitivity of neurons within the entorhinal cortex [Hafting et al. (2005)]. This paper focuses on data-efficient methods for tuning function estimation.

To be more concrete, let us first consider example experimental data where the activity of n neurons is measured across d trials of identical lengths, with different stimuli presented during each trial. We can then model the response $\boldsymbol{y}_i \in \mathbb{R}^d$ of neuron i as a function of a stimulus matrix $\boldsymbol{X}_i \in \mathbb{R}^{d\times m}$. Each row of $\boldsymbol{X}_i$ corresponds to the stimulus projected onto neuron i at each of the d trials. In the simplest case, the relationship between the unobserved tuning function $\boldsymbol{\beta}_i \in \mathbb{R}^m$ and the observed activity $\boldsymbol{y}_i$ at neuron i in response to stimulus $\boldsymbol{X}_i$ can be modeled as[2]

$$\text{(1)} \qquad \boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta}_i + \boldsymbol{\epsilon}_i \qquad \text{where } \boldsymbol{\epsilon}_i \sim \mathcal{N}(0, \nu_i^2\sigma^2\boldsymbol{I}).$$

The efficient statistical analysis and estimation of the unobserved tuning functions $\{\boldsymbol{\beta}_i\}$ given the noisy observations $\{\boldsymbol{y}_i\}$ and the stimulus set $\{\boldsymbol{X}_i\}$ is the tuning function estimation problem. In this setting, one standard approach is to use, for example, maximum-likelihood estimation to estimate tuning functions one neuron at a time [e.g., $\boldsymbol{\beta}_{i,\mathrm{ml}} := (\boldsymbol{X}_i'\boldsymbol{X}_i)^{-1}\boldsymbol{X}_i'\boldsymbol{y}_i$].

However, this model neglects a common feature of many neural circuits: the spatial clustering of neurons sharing a similar information processing function. For example, there are maps of tone frequency across the cortical surface in the auditory system [Issa et al. (2014)], visual orientation maps in both cortical [Hubel and Wiesel (1962, 1968), Ohki et al. (2005)] and subcortical brain regions [Feinberg and Meister (2014)], and maps respecting the spatial organization of the body (somatotopy) in the motor system [Bouchard et al. (2013), Leyton and Sherrington (1917), Machado et al. (2015), Penfield and Rasmussen (1950), Romanes (1964)]. As a consequence, neurons in close proximity often have similar tuning functions [see Swindale (2008), Wilson and Moore (2015) for recent reviews]. In each

[2] Empirical findings, to some degree, challenge the linear neural response to the stimulus, the conditionally independent neural activity and the Gaussian noise assumptions. Nevertheless, numerous studies have successfully used these simplifying assumptions to analyze neural data [see Doi et al. (2012), Rieke et al. (1997) and references therein]. In the concluding Section 5, we discuss directions for future work that allow the approach presented here to be extended to more general settings, for example, correlated point process observations.

of these cases, there are typically regions where this rule is violated and largely smooth tuning maps are punctuated by jumps or discontinuities. Therefore, simply smoothing in all cases will erode the precision of any sharp borders that might exist. Ideally, we would use an approach to estimate $\{\boldsymbol{\beta}_i\}$ that would smooth out the tuning map more in areas where there is evidence from the data that nearby tuning functions are similar, while letting the data "speak for itself" and applying minimal smoothing in regions where adjacent neurons have tuning functions that are very dissimilar.

In this paper, we propose a multivariate Bayesian extension of group lasso [Yuan and Lin (2006)], generalized lasso [Tibshirani and Taylor (2011)], network lasso [Hallac, Leskovec and Boyd (2015)], trend filtering on graphs [Wang et al. (2016)] and total-variation (TV) regularization [Rudin, Osher and Fatemi (1992)]. Specifically, we use the following improper prior:

$$\boldsymbol{\beta}|\lambda, \sigma \propto \prod_{i\sim j}\left(\frac{\lambda}{2\sigma}\right)^m \exp\left(-\frac{\lambda}{\sigma}\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|_2\right), \tag{2}$$

where $\|u\|_2 = \sqrt{\sum_{i=1}^m u_i^2}$ and $i \sim j$ if two cells i and j are spatially nearby.[3] This prior allows for a flexible level of similarity between nearby tuning functions. For clarity, we contrast against a $\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|_2^2$ based prior:

$$\prod_{i\sim j}\left(\frac{\lambda^2}{2\pi\sigma^2}\right)^{m/2} \exp\left(-\frac{\lambda^2}{2\sigma^2}\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|_2^2\right),$$

which penalizes large local differences quadratically. The prior defined in (2), on the other hand, penalizes large differences linearly; intuitively, this prior encourages nearby tuning functions to be similar while allowing for large occasional breaks or outliers in the spatial map of the inferred tuning functions. This makes the estimates much more robust to these occasional breaks.

The paper is organized as follows. Section 2 presents the full description of our statistical model, including likelihood, priors and hyperpriors. Section 3 presents an efficient block Gibbs sampler with discussions about its statistical and computational properties. Finally, Section 4 illustrates our robust and scalable Bayesian analysis of simulated data from the visual cortex and real neural data obtained from the spinal cord. We conclude in Section 5 with a discussion of related work and possible extensions to our approach.

2. Bayesian inference.

To complete the model introduced above, we place an inverse Gamma prior on σ and $\{\nu_i\}_{i=1,\dots,n}$, and we place a Gamma prior on λ^2, both of which are fairly common choices in Bayesian inference [Park and Casella

[3] We will clearly define the notion of proximity $i \sim j$ at the end of Section 2.

(2008)]. These choices lead to the likelihood, priors and hyperpriors presented below:

$$\text{likelihood,}\qquad \boldsymbol{y}_i|\boldsymbol{\beta}_i,\sigma,\nu_i \sim \left(\frac{1}{2\pi\nu_i^2\sigma^2}\right)^{d/2}\exp\left(-\frac{1}{2\nu_i^2\sigma^2}\|\boldsymbol{y}_i-\boldsymbol{X}_i\boldsymbol{\beta}_i\|_2^2\right),$$

$$\text{prior,}\qquad \boldsymbol{\beta}|\lambda,\sigma \sim \prod_{i\sim j}\left(\frac{\lambda}{2\sigma}\right)^m \exp\left(-\frac{\lambda}{\sigma}\|\boldsymbol{\beta}_i-\boldsymbol{\beta}_j\|_2\right)$$

and hyperpriors,

$$\begin{aligned}
&\sigma^2 \sim \text{inverse-Gamma}(\kappa,\epsilon) = \frac{\epsilon^\kappa}{\Gamma(\kappa)}(\sigma^2)^{-\kappa-1}e^{-\epsilon/\sigma^2},\\
&\lambda^2 \sim \text{Gamma}(r,\delta) = \frac{\delta^r}{\Gamma(r)}(\lambda^2)^{r-1}e^{-\delta\lambda^2},\\
&\nu_i^2 \sim \text{inverse-Gamma}(\varkappa,\varepsilon) = \frac{\varepsilon^\varkappa}{\Gamma(\varkappa)}(\nu_i^2)^{-\varkappa-1}e^{-\varepsilon/\nu_i^2}.
\end{aligned}\tag{3}$$

The well-known representation [Andrews and Mallows (1974), Casella et al. (2010), Eltoft, Kim and Lee (2006), West (1987)] of the Laplace prior as a scale mixture of Normals,

$$\begin{aligned}
&\left(\frac{\lambda}{2\sigma}\right)^m \exp\left(-\frac{\lambda}{\sigma}\|\boldsymbol{\beta}_i-\boldsymbol{\beta}_j\|_2\right)\\
&\quad = C\int_0^\infty \left(\frac{1}{2\pi\sigma^2\tau_{ij}^2}\right)^{m/2}\\
&\qquad \times \exp\left(-\frac{\|\boldsymbol{\beta}_i-\boldsymbol{\beta}_j\|_2^2}{2\sigma^2\tau_{ij}^2}\right)\underbrace{\frac{(\frac{\lambda^2}{2})^{\frac{m+1}{2}}}{\Gamma(\frac{m+1}{2})}(\tau_{ij}^2)^{\frac{m+1}{2}-1}e^{-\frac{\lambda^2}{2}\tau_{ij}^2}\,d\tau_{ij}^2}_{\tau_{ij}^2\sim\text{Gamma}(\frac{m+1}{2},\frac{\lambda^2}{2})},
\end{aligned}$$

[where $C=\pi^{\frac{m-1}{2}}\Gamma(\frac{m+1}{2})$] allows us to formulate our prior (2) in a hierarchical manner:

$$\tau_{ij}^2|\lambda^2 \sim \frac{(\frac{\lambda^2}{2})^{\frac{m+1}{2}}}{\Gamma(\frac{m+1}{2})}(\tau_{ij}^2)^{\frac{m+1}{2}-1}e^{-\frac{\lambda^2}{2}\tau_{ij}^2}\qquad \text{for all } i\sim j,\tag{4}$$

$$\boldsymbol{\beta}|\{\tau_{ij}^2\},\sigma^2 \sim \exp\left(-\frac{\boldsymbol{\beta}'\boldsymbol{D}'\boldsymbol{\Gamma}\boldsymbol{D}\boldsymbol{\beta}}{2\sigma^2}\right),\tag{5}$$

where (using ⊗ as the Kronecker product)

$$\boldsymbol{D}=\boldsymbol{D}_s\otimes\boldsymbol{I}_m \quad\text{and}\quad \boldsymbol{\Gamma}=\boldsymbol{\Gamma}_s\otimes\boldsymbol{I}_m,$$

$$\boldsymbol{\Gamma}_s=\text{diag}\left(\ldots,\frac{1}{\tau_{ij}^2},\ldots\right)\in\mathbb{R}^{p\times p}$$

and $\boldsymbol{D}_s \in \mathbb{R}^{p\times n}$ is a sparse matrix such that each row accommodates a $+1$ and -1, corresponding to $i \sim j$. We let p denote the number of edges in the proximity network. Note that

$$\boldsymbol{\beta}'\boldsymbol{D}'\boldsymbol{\Gamma}\boldsymbol{D}\boldsymbol{\beta} = \sum_{i\sim j} \frac{\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|_2^2}{\tau_{ij}^2}.$$

In light of the hierarchical representation, illustrated in equations (4), (5), the prior defined in (2) can be viewed as an improper Gaussian mixture model; $\boldsymbol{\beta}$ is Gaussian given $\{\dots, \tau_{ij}^2, \dots\}$, and the τ_{ij}^2s come from a common ensemble. This prior favors spatial smoothness while allowing the amount of smoothness to be variable and adapt to the data. As we will discuss in Section 3, posterior samples of τ_{ij}^2 tend to be smaller in smooth areas than in regions with discontinuities or outliers.

For each edge in the proximity network, and each corresponding row in $\boldsymbol{D}_s$, there is a unique pair of nodes i and j that are spatially "nearby," that is, $i \sim j$. We found that considering the four horizontally and vertically nearby nodes as neighbors, for nodes that lie on a two-dimensional regular lattice, allows us to efficiently estimate tuning functions without contamination from measurement noise or bias from oversmoothing; see Section 4.1.1 for an illustrative example. As for nodes that lie on an irregular grid, we compute the sample mean $\boldsymbol{\mu}$ and sample covariance $\boldsymbol{C}$ of the locations, and then whiten the location vectors $\boldsymbol{v}_i$; that is, $\boldsymbol{v}_{i,\text{whitened}} = \boldsymbol{C}^{-1/2}(\boldsymbol{v}_i - \boldsymbol{\mu})$. We found that connecting each node to its k-nearest-neighbors (within a maximum distance r) in the whitened space works well in practice; see Section 4.2.1 for an illustrative example with $k = 1$ and $r = 5$.

Extending the robust prior presented in equation (2), which is based on the simple local difference $\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|_2$ for $i \sim j$, to a robust prior based on any generic $\|\cdot\|_2$ measure of local roughness is easy; we only need to appropriately modify $\boldsymbol{D}_s$. For example, if $\boldsymbol{y}_1, \dots, \boldsymbol{y}_n$ are equidistant temporal samples, then the following robust prior

$$\boldsymbol{\beta}|\lambda, \sigma \sim \prod_{i=1}^{n-2} \left(\frac{\lambda}{2\sigma}\right)^m \exp\left(-\frac{\lambda}{\sigma}\|\mathbf{2}\boldsymbol{\beta}_i - \boldsymbol{\beta}_{i+1} - \boldsymbol{\beta}_{i-1}\|_2\right)$$

reflects our a priori belief that $\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_n$ are (approximately) piecewise linear [Kim et al. (2009)]. In this case, $\boldsymbol{D}_s$ is a tridiagonal matrix with 2 on the diagonal and -1 on the off-diagonals. As another example, let the matrix $\boldsymbol{D}_s$ be equal to the discrete Laplacian operator; $[\boldsymbol{D}_s]_{ii}$ equals the number of edges attached to node i, and if $i \sim j$, then $[\boldsymbol{D}_s]_{ij} = -1$, otherwise its zero. The discrete Laplacian operator (Laplacian matrix), which is an approximation to the continuous Laplace operator, is commonly used in the spatial smoothing literature to impose a roughness penalty [Wahba (1990)]. Our robust prior based on the discrete Laplacian operator is as follows:

$$\boldsymbol{\beta}|\lambda, \sigma \sim \prod_{i=1}^{n} \left(\frac{\lambda}{2\sigma}\right)^m \exp\left(-\frac{\lambda}{\sigma}\left\|\sum_{j\sim i}(\boldsymbol{\beta}_i - \boldsymbol{\beta}_j)\right\|_2\right),$$

which given the appropriate matrix $\boldsymbol{D}_s$ can easily be formulated in the hierarchical manner of equation (5). On regular grids, this prior is based only on the four (horizontal and vertical) neighbors, but better approximations to the the continuous Laplace operator based on more neighbors is straightforward and within the scope of our scalable block Gibbs sampler presented in Section 3.

Finally, note that the prior defined in (2) is not a proper probability distribution because it can not be normalized to one. However, in most cases the posterior distribution will still be integrable even if we use such an improper prior [Gelman et al. (2003)]. As we see later in Section 3, all the conditional distributions needed for block Gibbs sampling are proper. Furthermore, the joint posterior inherits the unimodality in $\boldsymbol{\beta}$ and σ given $\{\nu_i\}_{i=1,\dots,n}$ and λ from the Bayesian Lasso [Park and Casella (2008)], aiding in the mixing of the Markov chain sampling methods employed here; see the Appendix.

2.1. *Relationship to network lasso.* In related recent independent work, Hallac, Leskovec and Boyd (2015) present an algorithm based on the alternating direction method of multipliers [Boyd et al. (2011)] to solve the network lasso convex optimization problem,

$$\underset{\boldsymbol{\beta}_i \in \mathbb{R}^m \text{ for } i=1,\dots,n}{\text{minimize}} \sum_{i=1}^{n} \|\boldsymbol{y}_i - \boldsymbol{X}_i \boldsymbol{\beta}_i\|_2^2 + \gamma \sum_{i \sim j} \|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|_2, \tag{6}$$

in a distributed and scalable manner. The parameter γ scales the edge penalty relative to the node objectives (and can be tuned using cross-validation). Similar to our formulation (Section 2), the network lasso uses an edge cost that is a sum of norms of differences of the adjacent node variables, leading to a setting that allows for robust smoothing within clusters on graphs. The optimization approach of Hallac, Leskovec and Boyd (2015) leads to fast computation but sacrifices the quantification of posterior uncertainty (which is in turn critical for closed-loop experimental design, for example, deciding which neurons should be sampled more frequently to reduce posterior uncertainty) provided by the method proposed here. A Bayesian version of the network lasso is a special case of our robust Bayesian formulation by setting the variable variance parameters equal to one, that is, $\nu_i^2 = 1$ for $i = 1, \dots, n$. As we will see in the next example, heteroscedastic noise challenges the posterior mean estimate's robustness.

2.2. *Model illustration.* In this section, we show that posterior means based on the prior of equation (3) on $\{\nu_i\}_{i=1,\dots,n}$ are robust to neuron-dependent noise variance. Our numerical experiments for heterogenous noise power show that a model with a homogeneous noise assumption will misinterpret noise as a signal, depicted in Figure 1. Comparisons with the network lasso are presented as well. We postpone the details concerning the block Gibbs sampler presented in this paper to Section 3.

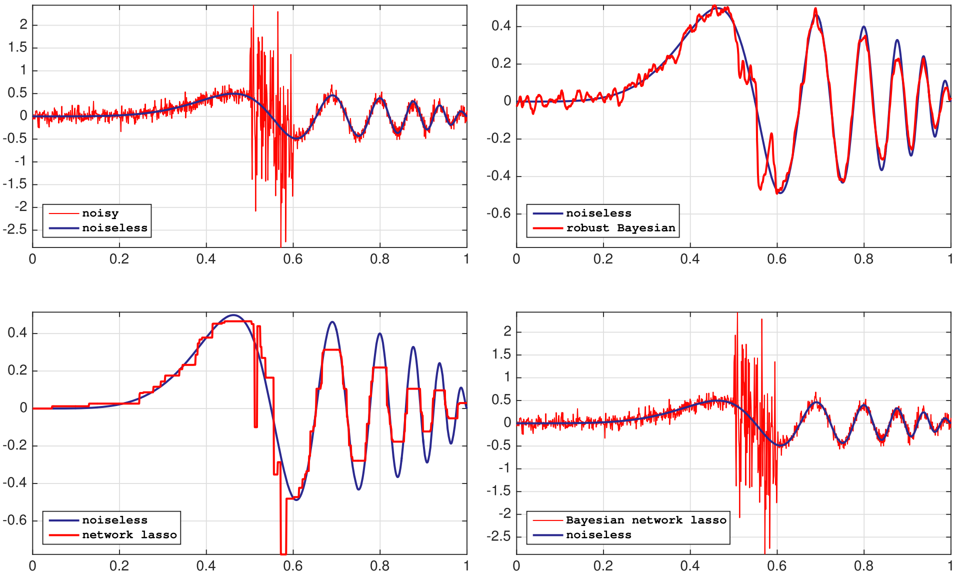


FIG. 1. *Heterogenous noise example. The Bayesian network lasso posterior mean estimate overfits in the region of higher observation noise. The robust Bayesian formulation is less prone to misidentifying heterogenous noise as a signal. The network lasso tends to cluster high frequency variations into piecewise-constant estimates. Note that in the bottom right panel, the Bayesian network lasso looks similar to the noisy observation in the upper left panel; that is, this estimator is overfitting here. The nonrobust Bayesian network lasso formulation does not allow various degrees of noise variability to be estimated from the data, and, hence, in this example it interprets the change in noise variance as a signal.*

The signal and heterogeneous noise models are as follows:

$$y_i = \beta_i + \epsilon_i, \qquad \text{where } \beta_i = \sqrt{\frac{i}{n}\left(1 - \frac{i}{n}\right)} \sin\left(11\pi \frac{i^4}{n^4}\right),$$

$$\epsilon_i \sim \mathcal{N}(0, \sigma_i^2), \qquad \text{with } \sigma_i = \begin{cases} 0.1, & \text{if } \frac{i}{n} \in [0, 0.5) \cup (0.6, 1], \\ 1, & \text{if } \frac{i}{n} \in [0.5, 0.6]. \end{cases}$$

The following hyperpriors were used for the posterior means of the robust Bayesian model:

$$\sigma^2 \sim \text{inverse-Gamma}(\kappa = 0, \epsilon = 0),$$
$$\lambda^2 \sim \text{Gamma}(r = 0.0001, \delta = 0.001),$$
$$\nu_i^2 \sim \text{inverse-Gamma}(\varkappa = 3, \varepsilon = 2).$$

The hyperpriors of λ^2 and σ^2 are relatively flat. For ν_i^2, we set the hyperparameters such that we have the unit prior mean and prior variance. The Bayesian network lasso is only different from the robust Bayesian formulation in that it assumes a constant noise variance, that is, $\nu_i^2 = 1$ for $i = 1, \dots, n$.

The Bayesian network lasso and robust Bayesian posterior mean estimates are based on 10,000 consecutive iterations of the Gibbs sampler (after 5000 burn-in iterations), as discussed in Section 3. The network lasso estimate is the solution to the convex optimization problem equation (6) where the tuning parameter γ is set using 10-fold cross-validation. Note that the network lasso estimate corresponds to the mode of the posterior distribution of the Bayesian network lasso conditioned on σ and λ.

For the sake of comparison, we also present numerical results for a homogeneous noise model. Here, the signal $\boldsymbol{\beta}$ is the same, but the noise variance is $\sigma_i = 0.33$ for $i = 1, \dots, n$. This particular choice of σ_i was made to guarantee that the signal-to-noise ratio is equal to that of the heterogeneous noise model. As for the priors, they remain the same. As expected, the Bayesian network lasso and robust Bayesian posterior means are similar, as depicted in Figure 2.

Figures 1 and 2 illustrate that if the noise power is constant, then the robust Bayesian and Bayesian network lasso posterior means are similar. On the other hand, if the noise power is not constant, then the robust Bayesian posterior mean detects the nonuniform noise power and adapts to it, while the Bayesian network lasso posterior mean misinterprets noise as signal and overfits. Note that in general local differences are sparsened by the network lasso MAP estimate, but are instead

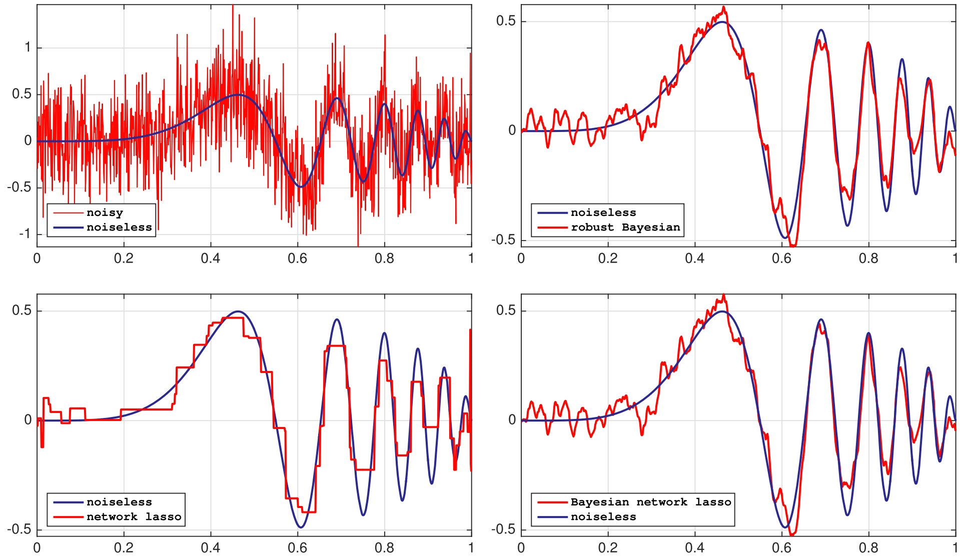


FIG. 2. *Homogeneous noise example. The posterior means of the Bayesian network lasso and our robust Bayesian are very similar. This is expected given the homogeneity of noise power. The network lasso suffers from the staircase effect, that is, the denoised signal is not smooth but piecewise constant.*

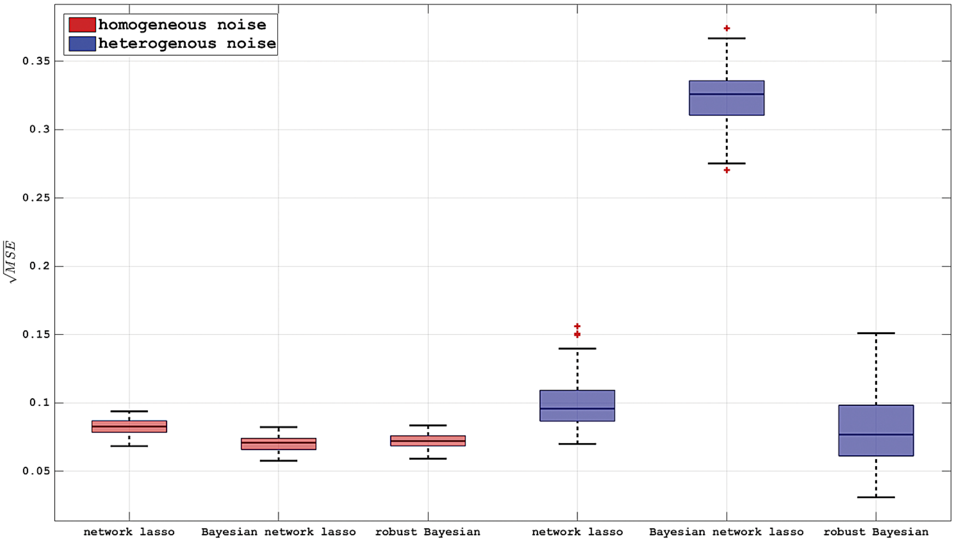


FIG. 3. *Tukey boxplots comparing model* $\sqrt{MSE} := n^{-1/2}\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{\text{model}}\|_2$ *under homogeneous and heterogeneous noise. To make this comparison meaningful, signal-to-noise ratios are the same for both noise models. The boxplots are generated by simulating* 100 *replications of each model. For homogeneous noise, the Bayesian network lasso and robust Bayesian perform similarly. However, when noise is heterogeneous, the Bayesian network lasso tends to overfit, as illustrated in Figure* 1. *In terms of* MSE, *the network lasso is more robust to noise variations than its Bayesian counterpart, but the robust Bayesian performs slightly better.*

shrunk by the posterior mean of the Bayesian network lasso.[4] More importantly, in this example, the Bayesian and non-Bayesian network lasso estimate λ in very different ways (Gibbs sampling and cross-validation, respectively), which lead to different levels of penalization here. Repeated simulations presented in Figure 3 further confirm these observations.

3. Scalable block Gibbs sampling.

We will now introduce some vector and matrix notation before we describe our Gibbs sampling approach to inference. First, we introduce the following variables:

$$\underline{\boldsymbol{y}}_i := \frac{\boldsymbol{y}_i}{\nu_i}, \qquad \underline{\boldsymbol{X}}_i := \frac{\boldsymbol{X}_i}{\nu_i}. \tag{7}$$

We also let $\underline{\boldsymbol{X}} \in \mathbb{R}^{nd\times nm}$ stand for the rectangular blockwise-diagonal matrix $\operatorname{diag}(\dots, \underline{\boldsymbol{X}}_i, \dots)$. Moreover, we let $\boldsymbol{\beta}$, $\underline{\boldsymbol{y}}$ and $\underline{\boldsymbol{X}}'\underline{\boldsymbol{y}}$ stand for the column-wise concatenation (for $i = 1, \dots, n$) of $\boldsymbol{\beta}_i$, $\underline{\boldsymbol{y}}_i$ and $\underline{\boldsymbol{X}}_i'\underline{\boldsymbol{y}}_i$, respectively. $\underline{\boldsymbol{X}}'\underline{\boldsymbol{X}}$ is then the

[4] In the lasso regression problem, likewise differences exist between MAP estimates and posterior estimates [Park and Casella (2008)]. To be concrete, the posterior mean of the Bayesian lasso generically has no nonzero terms, unlike the standard MAP lasso estimate.

blockwise-diagonal matrix $\operatorname{diag}(\ldots, \underline{\boldsymbol{X}}_i' \underline{\boldsymbol{X}}_i, \ldots) \in \mathbb{R}^{nm \times nm}$. Finally, recall that p stands for the number of edges in the proximity network.

Our efficient Gibbs sampler and the full conditional distributions of $\boldsymbol{\beta}$, σ^2, $\{\nu_i^2\}$, λ and $\{\tau_{ij}^2\}$ can then be formulated as follows:

Step 1. The local smoothing parameters $\{\tau_{ij}\}_{i \sim j}$ are conditionally independent, with

$$\tau_{ij}^2 | \beta, \sigma^2, \lambda^2 \sim \left(\frac{1}{\tau_{ij}^2}\right)^{1/2} \exp\left(-\frac{\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|^2}{2\sigma^2 \tau_{ij}^2} - \frac{\lambda^2}{2}\tau_{ij}^2\right).$$

Step 2. The full conditional for $\boldsymbol{\beta}$ is multivariate normal with mean $\boldsymbol{P}^{-1}\underline{\boldsymbol{X}}'\underline{\boldsymbol{y}}$ and covariance $\sigma^2 \boldsymbol{P}^{-1}$, where

$$\boldsymbol{P} = \underline{\boldsymbol{X}}'\underline{\boldsymbol{X}} + \boldsymbol{D}'\boldsymbol{\Gamma}\boldsymbol{D}.$$

Step 3. $\sigma^2 \sim$ inverse-Gamma(κ',ϵ') with

$$\kappa' = \kappa + \frac{(pm + nd)}{2}, \quad \text{and} \quad \epsilon' = \epsilon + \frac{1}{2}\|\underline{\boldsymbol{y}} - \underline{\boldsymbol{X}}\boldsymbol{\beta}\|^2 + \frac{1}{2}\|\boldsymbol{\Gamma}^{1/2}\boldsymbol{D}\boldsymbol{\beta}\|^2.$$

Step 4. $\lambda^2 \sim$ Gamma(r', δ') with

$$r' = r + p(m+1)/2, \quad \text{and} \quad \delta' = \delta + \frac{1}{2}\sum_{i \sim j} \tau_{ij}^2.$$

Step 5. $\nu_i^2 \sim$ inverse-Gamma($\varkappa'$, ε') with

$$\varkappa' = \varkappa + \frac{d}{2}, \quad \text{and} \quad \varepsilon' = \varepsilon + \frac{1}{2\sigma^2}\|\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}_i\|^2.$$

Note that in step 1, the conditional distribution can be rewritten as

$$\frac{1}{\tau_{ij}^2}\bigg|\boldsymbol{\beta}, \sigma^2, \lambda \sim \text{inverse-Gaussian}(\mu', \lambda') \tag{8}$$

with

$$\mu' = \frac{\lambda\sigma}{\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|_2}, \qquad \lambda' = \lambda^2,$$

in the parametrization of the inverse-Gaussian density given by

$$\text{inverse-Gaussian}(\mu', \lambda') \sim f(x) = \sqrt{\frac{\lambda'}{2\pi}} x^{-3/2} \exp\left\{-\frac{\lambda'(x - \mu')^2}{2(\mu')^2 x}\right\}.$$

Moreover, the conditional expectation of $\frac{1}{\tau_{ij}^2}$ [using its inverse-Gaussian density in (8)] is equal to $\frac{\lambda\sigma}{\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|_2}$. This makes the iterative Gibbs sampler above intuitively appealing; if the local difference is significantly larger than typical noise (i.e., $\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|_2 \gg \lambda\sigma$), then there is information in the difference, and therefore minimal smoothing is applied in order to preserve that difference. On the other hand, if the local difference is small, then this difference is likely to be due to noise, and therefore local smoothing will reduce the noise. In other words, the robust Bayesian formulation presented in this paper functions as an adaptive smoother where samples will be less smooth in regions marked with statistically significant local differences, and vice versa.

Furthermore, in step 2, the conditional distribution of $\boldsymbol{\beta}$ depends on the observation $\boldsymbol{y}$ and the local smoothing parameters τ. A large $1/\tau_{ij}^2$ causes the samples of $\boldsymbol{\beta}_i$ and $\boldsymbol{\beta}_j$ to be more similar to each other than their respective ML estimates $\boldsymbol{\beta}_{i,\mathrm{ml}}$ and $\boldsymbol{\beta}_{j,\mathrm{ml}}$ [where $\boldsymbol{\beta}_{i,\mathrm{ml}} := (\boldsymbol{X}_i'\boldsymbol{X}_i)^{-1}\boldsymbol{X}_i'\boldsymbol{y}_i$]. In contrast, if $1/\tau_{ij}^2$ is small, then the conditional samples of $\boldsymbol{\beta}_i$ and $\boldsymbol{\beta}_j$ typically revert to their respective ML estimates, plus block-independent noise. Also, note that we can reduce the variance of the posterior mean's estimate of $\boldsymbol{\beta}$ by Rao–Blackwellization: instead of estimating the posterior mean of $\boldsymbol{\beta}$ from the samples obtained in step 2, the Rao–Blackwellized estimate of the posterior mean is computed by averaging (over all iterations) $\boldsymbol{P}^{-1}\underline{\boldsymbol{X}}'\underline{\boldsymbol{y}}$ (the conditional mean of $\boldsymbol{\beta}$), which can lead to significant improvements over naive Gibbs.

Finally, although unnecessary in our approach, the fully Bayesian sampling of λ in step 4 can be replaced with an empirical Bayes method. The difficulty in computing the marginal likelihood of λ, which requires a high-dimensional integration, can be avoided with the aid of the EM/Gibbs algorithm [Casella (2001)]. Specifically, iteration k of the EM algorithm

$$\lambda^{(k+1)} = \operatorname{argmax}_\lambda \mathrm{E}[\log p(\boldsymbol{\beta}, \tau^2, \lambda|\boldsymbol{y})|\boldsymbol{y}, \lambda^{(k)}]$$

simplifies to

$$\lambda^{(k+1)} = \sqrt{\frac{p(m+1)}{\sum_{i\sim j}\mathrm{E}[\tau_{ij}^2|\boldsymbol{y}, \lambda^{(k)}]}}, \tag{9}$$

which can be approximated by replacing conditional expectations with sample averages from step 1. The empirical Bayes approach gives consistent results with the fully Bayesian setting. The expectation of the conditional Gamma distribution of λ^2 in step 4,

$$\mathrm{E}[\lambda^2|\boldsymbol{y}, \tau] = \frac{2r + p(m+1)}{2\delta + \sum_{i\sim j}\tau_{ij}^2},$$

is similar to the EM/Gibbs update (9). In our experience, both approaches give similar results on high-dimensional data.

3.1. *Computational cost.* The conditional independence of the local smoothing parameters $\{\tau_{ij}\}_{i\sim j}$ given $\boldsymbol{\beta}$ and σ amounts to a computational cost of sampling these variables that scales linearly with their size: $O(pm)$. Similarly, the cost of sampling σ^2 given $\boldsymbol{\beta}$ and$\{\tau_{ij}\}_{i\sim j}$ is due to computing $\sum_{i=1}^{n}\|\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}_i\|^2$, $\sum_{i=1}^{n}\|\underline{\boldsymbol{y}}_i - \underline{\boldsymbol{X}}_i\boldsymbol{\beta}_i\|^2$ and $\|\boldsymbol{\Gamma}^{1/2}\boldsymbol{D}\boldsymbol{\beta}\|^2$ which are, respectively, $O(ndm)$, $O(ndm)$ and $O(pm)$, amounting to a total cost of $O((nd+p)m)$.

The conditional distribution of $\boldsymbol{\beta}$ given $\{\tau_{ij}\}_{i\sim j}$ is multivariate Gaussian with mean $\boldsymbol{P}^{-1}\underline{\boldsymbol{X}}'\underline{\boldsymbol{y}}$ and covariance $\sigma^2\boldsymbol{P}^{-1}$, whose computational feasibility rests primarily on the ability to solve the equation

$$\boldsymbol{P}\boldsymbol{w} = \boldsymbol{b} \tag{10}$$

as a function of the unknown vector $\boldsymbol{w}$ for $\boldsymbol{P} = \underline{\boldsymbol{X}}'\underline{\boldsymbol{X}} + \boldsymbol{D}'\boldsymbol{\Gamma}\boldsymbol{D}$. This is because if $\boldsymbol{\epsilon}_1, \boldsymbol{\epsilon}_2 \sim \mathcal{N}(0, \boldsymbol{I})$, then

$$\boldsymbol{P}^{-1}\underline{\boldsymbol{X}}'\underline{\boldsymbol{y}} + \sigma\boldsymbol{P}^{-1}[\underline{\boldsymbol{X}}'\boldsymbol{\epsilon}_1 + \boldsymbol{D}'\boldsymbol{\Gamma}^{1/2}\boldsymbol{\epsilon}_2] \tag{11}$$

is a Gaussian random vector with mean $\boldsymbol{P}^{-1}\underline{\boldsymbol{X}}'\underline{\boldsymbol{y}}$ and covariance $\sigma^2\boldsymbol{P}^{-1}$. Similar approaches for the efficient realization of Gaussian fields based on optimizing a randomly perturbed cost function (log posterior) were studied in Bardsley et al. (2014), Gilavert, Moussaoui and Idier (2015), Hoffman (2009), Hoffman and Ribak (1991), Papandreou and Yuille (2010). In our case, the randomly perturbed cost function is

$$\begin{aligned} f_{\boldsymbol{\epsilon}_1,\boldsymbol{\epsilon}_2}(\boldsymbol{\theta}) := {} & (\boldsymbol{D}\boldsymbol{\theta} - \sigma\boldsymbol{\Gamma}^{-1/2}\boldsymbol{\epsilon}_2)'\boldsymbol{\Gamma}(\boldsymbol{D}\boldsymbol{\theta} - \sigma\boldsymbol{\Gamma}^{-1/2}\boldsymbol{\epsilon}_2) \\ & + (\underline{\boldsymbol{y}} + \sigma\boldsymbol{\epsilon}_1 - \underline{\boldsymbol{X}}\boldsymbol{\theta})'(\underline{\boldsymbol{y}} + \sigma\boldsymbol{\epsilon}_1 - \underline{\boldsymbol{X}}\boldsymbol{\theta}), \end{aligned}$$

in which case it is easy to see that $\arg\max_{\boldsymbol{\theta}} f_{\boldsymbol{\epsilon}_1,\boldsymbol{\epsilon}_2}(\boldsymbol{\theta})$ is given by equation (11).

Standard methods for computing $\boldsymbol{P}^{\mathbf{-1}}\boldsymbol{b}$ require cubic time and quadratic space, rendering them impractical for high-dimensional applications. A natural idea for reducing the computational burden involves exploiting the fact that $\boldsymbol{P}$ is composed of a block-diagonal matrix $\underline{\boldsymbol{X}}'\underline{\boldsymbol{X}}$ and a sparse matrix $\boldsymbol{D}'\boldsymbol{\Gamma}\boldsymbol{D}$. For instance, matrices based on discrete Laplace operators on regular grids lend themselves well to multigrid algorithms which have linear time complexity [see Brandt (1977), Goodman and Sokal (1989), Papandreou and Yuille (2010) and Section 19.6 of Press et al. (1992)]. Even standard methods for solving linear equations involving sparse matrices (as implemented, for example, in MATLAB's $\boldsymbol{P} \setminus \boldsymbol{b}$ call) are quite efficient here, requiring sub-quadratic time [Rue and Held (2005)]. This sub-quadratic scaling requires that a good ordering is found to minimize fill-in during the forward sweep of the Gaussian elimination algorithm; code to find such a good ordering (via "approximate minimum degree" algorithms [Davis (2006)]) is built into the MATLAB call $\boldsymbol{P} \setminus \boldsymbol{b}$ when $\boldsymbol{P}$ is represented as a sparse matrix. As we will see in Section 4.1.1, exploiting these efficient linear algebra techniques permits sampling from a high-dimensional ($>10^6$) surface defined on a regular lattice in just a few seconds using MATLAB on a 2.53 GHz MacBook Pro.

4. Motivating neuroscience applications.

Here we will discuss the application of our robust Bayesian analysis approach toward the analysis of both synthetic and real neural tuning maps. In both cases, our new algorithm permits the robust estimation of neural tuning with higher fidelity and less data than alternative approaches.

4.1. *Synthetic data.*

4.1.1. *Estimating orientation preference maps.* We will first apply our algorithm to synthetic data modeled after experiments where an animal is presented with a visual stimulus and the neural activity in the primary visual cortex (also known as V1) is simultaneously recorded. V1 is the first stage of cortical visual information processing and includes neurons that selectively respond to sinusoidal grating stimuli that are oriented in specific directions [Hubel and Wiesel (1962)]. Neurons with such response properties are called *simple cells*. See Figure 4 for an illustrative example of the recorded neural activity while a bar of light is moved at different angles [Dayan and Abbott (2001), Henry, Dreher and Bishop (1974),

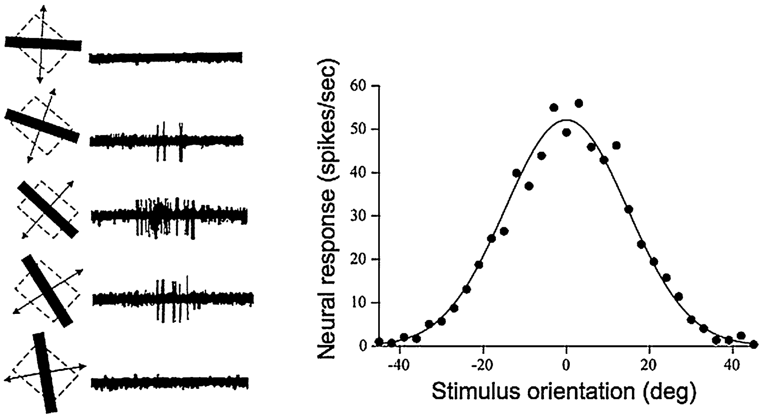


FIG. 4. *Electrophysiological recordings from a single neuron in the primary visual cortex of a monkey. A moving bar of light was projected onto the receptive field of the cell at different angles. In the diagrams on the left, the receptive field is shown as a dashed rectangle and the light source as a superimposed black bar. The angle of the dashed rectangle indicates the preferred orientation. For each bar (stimulus) orientation, the neural response was recorded. The voltage traces in the middle column show the electrophysiological recordings corresponding to the stimulus orientation of that row. Note that the neural response depends on the stimulus orientation; it increases as the bar and the preferred orientation become more aligned. Clearly, the bar orientation of the middle row evoked the largest number of action potentials. The graph on the right shows average number of action potentials per second (neural response) versus the angle of the bar. This graph indicates how the neural response depends on the orientation of the light bar. The data have been fit by a Gaussian function. [Data is from* Henry, Dreher and Bishop (1974), *Hubel and Wiesel* (1968), *and figures are adapted from* Dayan and Abbott (2001), Wandell (1995).]

Hubel and Wiesel (1968), Wandell (1995)]. As can be seen in the figure, action potential firing of the simple cell depends on the angle of orientation of the stimulus.

To capture the essential characteristics of simple cells in the visual cortex, we will use the following model. The response of cell $i \in \{1, 2, \ldots, n\}$ to a grating stimulus with orientation ϕ_ℓ depends on the preferred orientation $\theta_i \in (-90°, +90°]$ and the tuning strength $r_i \in \mathbb{R}^+$ of that cell. The number of cells is n, and the number of trials (with differently oriented stimuli) is d. Formally speaking, in the simplest linear model, during the ℓth trial, the noisy measurement $y_{i,\ell} \in \mathbb{R}$ at neuron i in response to a stimulus with orientation ϕ_ℓ can be written as [Macke et al. (2010, 2011), Swindale (1998)]

$$y_{i,\ell} | \boldsymbol{\beta}_i, \boldsymbol{x}_\ell, \sigma^2 \sim \mathcal{N}(\boldsymbol{\beta}_i' \boldsymbol{x}_\ell, \sigma^2), \qquad i = 1, \ldots, n \text{ and } \ell = 1, \ldots, d,$$

where $\boldsymbol{\beta}_i := r_i [\cos\theta_i \sin\theta_i]'$ is related to θ_i (preferred orientation) and r_i (tuning strength) as follows:

$$\theta_i := \arctan\left[\frac{\beta_{2,i}}{\beta_{1,i}}\right], \qquad r_i := \sqrt{\beta_{2,i}^2 + \beta_{1,i}^2},$$

and $\boldsymbol{x}_\ell = [\cos\phi_\ell \sin\phi_\ell]'$ stands for the grating stimulus with orientation ϕ_ℓ. Writing the stimulus set $\{\boldsymbol{x}_\ell\}_{\ell=1,\ldots,d}$ in matrix notation

$$\boldsymbol{X}_\varnothing := \begin{bmatrix} \vdots \\ \boldsymbol{x}_\ell' \\ \vdots \end{bmatrix}_{d \times 2}$$

allows us to compactly rewrite the neural response $\boldsymbol{y}_i \in \mathbb{R}^d$ as

$$\boldsymbol{y}_i | \boldsymbol{X}_\varnothing, \boldsymbol{\beta}_i, \sigma^2 \sim \mathcal{N}(\boldsymbol{X}_\varnothing, \boldsymbol{\beta}_i, \sigma^2 \boldsymbol{I}) \qquad i = 1, \ldots, n. \tag{12}$$

Note that all neurons respond to the same particular grating stimulus, namely $\boldsymbol{X}_\varnothing$, though, due to different preferred orientations, not all neurons respond similarly.

In this example, the noise variances are set to be equal, that is, $\nu_i = 1$ for $i = 1, \ldots, n$. As for the Gibbs sampler, we skip step 5, and substitute $\nu_i = 1$ in all other steps. In the next section, we present a real data example, where $\{\nu_i\}$ is estimated using step 5 of our Gibbs sampler.

Drawing conclusions regarding the cortical circuitry underlying orientation maps, their formation during visual development, and across evolution, has recently been the subject of numerous studies [Kaschube et al. (2010), Keil et al. (2012), Reichl, Löwel and Wolf (2009), Schnabel et al. (2007)]. For instance, Kaschube et al. (2010) argued that evolutionary history (instead of ecological or developmental constraints) underlies the formation of qualitatively similar pinwheel distributions observed in the visual cortex of disparate mammalian taxa. Consequently, the estimation of orientation maps without contamination from measurement noise or bias from overs-smoothing will help to clarify important questions about evolution and information processing in the visual cortex.

We therefore generated synthetic tuning maps by extracting the phase of superpositions of complex plane waves [see Section 2.4 of the Supplement of Kaschube et al. (2010) for details]. In our simulations, for clarity we assume $\boldsymbol{\beta}_i = (\cos\theta_i, \sin\theta_i)'$, and therefore $r_i = 1$, which means tuning strengths are constant across all neurons. The top left panels of Figure 5 and Figure 6 show the angular components $\{\theta_i\}$ and tuning strengths $r_i = 1$ of the resulting map. It is well known that in some species the preferred orientations $\{\theta_i\}$ are arranged around singularities, called pinwheel centers [Ohki et al. (2005, 2006)]. Around each singularity, the preferred orientations $\{\theta_i\}$ are circularly arranged, resembling a spiral staircase. If we closely examine the top left panel of Figure 5, it is evident that around pinwheel centers the preferred orientations $\{\theta_i\}$ are descending, either clockwise or counterclockwise from $-90°$ to $+90°$. Experimentally measured

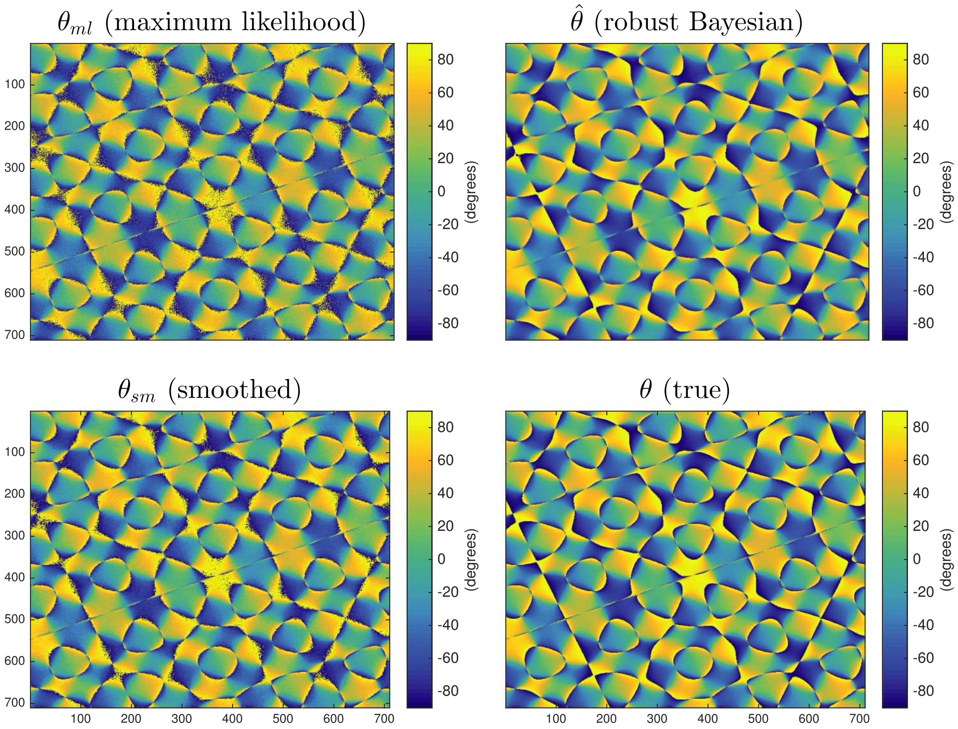


FIG. 5. *Analysis of a synthetic orientation tuning map. θ is a synthetic* 710 × 710 *orientation preference map* [*see Section* 2.4 *of the Supplement of Kaschube et al.* (2010) *for details*]. *Each pixel is a neuron, and $\theta_i \in (-90°, +90°]$* (*the preferred orientation of neuron i*) *is given by* $\arctan(\beta_{2,i}/\beta_{1,i})$. *Likewise, the robust Bayesian $\hat{\theta}$, smoothed θ_{sm} and maximum-likelihood θ_{ml} estimates of preferred orientations are inverse trigonometric functions of $\hat{\boldsymbol{\beta}}$, $\boldsymbol{\beta}_{\mathrm{sm}}$ and $\boldsymbol{\beta}_{\mathrm{ml}}$, respectively. The Bayesian estimate $\hat{\theta}$ of preferred orientations is less noisy than θ_{ml} and more robust than θ_{sm}*; *see also Figure 7 for a zoomed-in view. The $\hat{\boldsymbol{\beta}}$ estimate of posterior expectations is based on* 10,000 *consecutive iterations of the Gibbs sampler* (*after* 500 *burn-in iterations*).

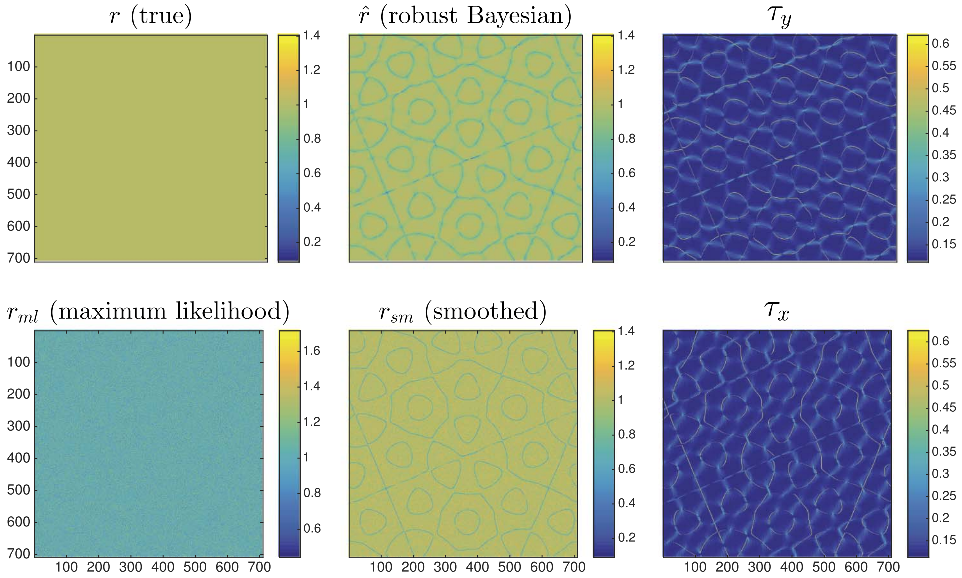


FIG. 6. *True tuning strengths* $\{r_i\}$, *the estimated tuning strengths* $\{r_{i,\mathrm{ml}}, r_{i,\mathrm{sm}}, \hat{r}_i\}$ *and posterior means of local smoothing parameters* $\{\tau_{ij}\}_{i\sim j}$. *Each pixel is a neuron, and its (estimated) tuning strength is given by the length of its (estimated)* $\boldsymbol{\beta}_i$, *for example,* $r_i = \|\boldsymbol{\beta}_i\|_2$, $\hat{r}_i = \|\hat{\boldsymbol{\beta}}_i\|_2$, *etc. The proximity network is a* 710×710 *regular grid with edges between a node and its four (horizontal and vertical) neighbors. The local smoothing parameters defined on edges among vertical and horizontal edges are designated by* $\{\tau_y\}$ *and* $\{\tau_x\}$, *respectively. The* r_{sm} *(smoothed) and* $\hat{r}$ *(robust Bayesian) tuning strength maps underestimate the true value at points where posterior means of local smoothing parameters* $\{\tau_x, \tau_y\}$ *take significant values. These points correspond to sharp breaks in the orientation preference map* θ *(as illustrated in Figure 5), where local averaging of significantly differently oriented tuning functions leads to a downward bias in estimated tuning strengths.*

maps obtained from cats, primates [Kaschube et al. (2010)] and our synthetically generated data all share this important feature.

We simulated the neural responses of each cell to twenty differently oriented grating stimuli by sampling responses according to equation (12) with $\sigma = 0.4$. The orientations ϕ_ℓ (for $\ell = 1, \dots, 20$) were randomly and uniformly sampled from $(-90°, +90°]$. Our main objective is to estimate (from neural responses $\{\boldsymbol{y}_i\}$ and stimuli $\boldsymbol{X}_\emptyset$) the preferred orientations $\{\theta_i\}$ and tuning strengths $\{r_i\}$. Ordinary linear regression yields maximum likelihood estimates

$$
\begin{aligned}
\boldsymbol{\beta}_{i,\mathrm{ml}} &= (\boldsymbol{X}_\emptyset' \boldsymbol{X}_\emptyset)^{-1} X_\emptyset' y_i, \\
\theta_{i,\mathrm{ml}} &= \arctan\left(\frac{\beta_{2,i,\mathrm{ml}}}{\beta_{1,i,\mathrm{ml}}}\right), \qquad (13) \\
r_{i,\mathrm{ml}} &= \|\boldsymbol{\beta}_{i,\mathrm{ml}}\|_2.
\end{aligned}
$$

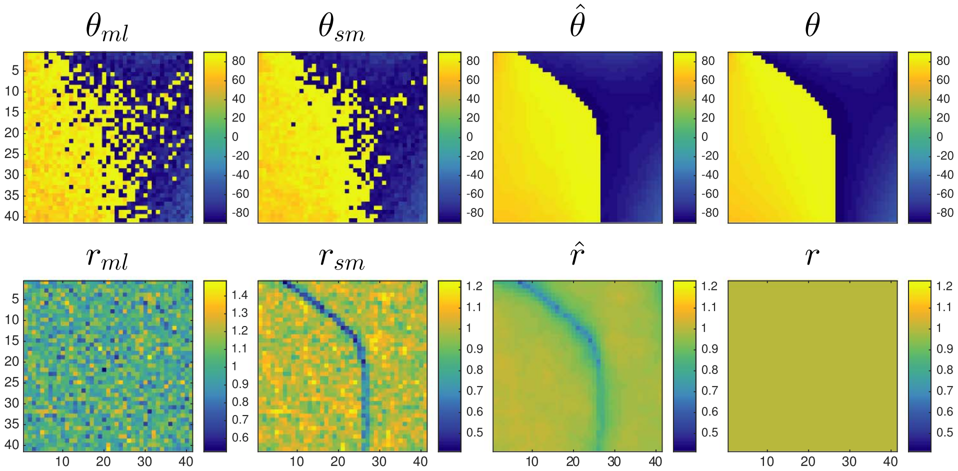


FIG. 7. *A* 40×40 *zoomed-in view of preferred orientations* $\{\theta_i\}$ *and tuning strengths* $\{r_i\}$, *and their estimates.* [*The center of this map is pixel* (241, 60) *in Figure* 5 *and Figure* 6.] *The smoothed* r_{sm} *tuning strength map underestimates the true tuning strength at sharp breaks in the orientation preference map* θ. *This bias is less severe for the Bayesian estimate* $\hat{r}$ *because the robust prior applies less local smoothing at sharp breaks* (*as illustrated in Figure* 6). *Similarly,* $\hat{\theta}$ *provides much more accurate angular estimates than* θ_{sm}.

The maximum likelihood estimates $\theta_{i,\mathrm{ml}}$ and $r_{i,\mathrm{ml}}$ are depicted in Figures 5, 6 and 7. The fine structure around pinwheel centers and the border between clustered preferred orientations is disordered.

We also computed the smoothed estimate $\boldsymbol{\beta}_{\mathrm{sm}}$ based on the following smoothing prior,

$$p(\boldsymbol{\beta}|\gamma) \propto \exp\biggl(-\frac{\gamma}{2}\sum_{i\sim j}\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|_2^2\biggr)$$

$$\propto \exp\biggl(-\frac{\gamma}{2}\boldsymbol{\beta}'\boldsymbol{D}'\boldsymbol{D}\boldsymbol{\beta}\biggr),$$

and the likelihood in (12),

$$p(\boldsymbol{y}|\boldsymbol{\beta}) \propto \exp\biggl(-\frac{1}{2\sigma^2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2\biggr),$$

leading to the posterior expectation of $\boldsymbol{\beta}$:

$$\boldsymbol{\beta}_{\mathrm{sm}}(\gamma) := (\boldsymbol{X}'\boldsymbol{X} + \gamma\boldsymbol{D}'\boldsymbol{D})^{-1}\boldsymbol{X}'\boldsymbol{y}, \tag{14}$$

where $\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{I}_{n\times n} \otimes \boldsymbol{X}'_{\emptyset}\boldsymbol{X}_{\emptyset}$ and $\boldsymbol{X}'\boldsymbol{y} = (\ldots, \boldsymbol{X}'_{\emptyset}\boldsymbol{y}_i, \ldots)$. The smoothed estimate $\boldsymbol{\beta}_{\mathrm{sm}}$ is based on a Gaussian prior that penalizes large local differences quadratically. [In contrast, the robust prior defined in equation (2) penalizes large differences linearly.] The amount of smoothing is dictated by γ; large values of γ lead to over-smoothing and small values of γ lead to under-smoothing. In this example, the true β is known; therefore, for the sake of finding the best achievable smoothing performance, we selected $\gamma = 2.15$ (using a grid search), which minimizes $\|\boldsymbol{\beta}_{\mathrm{sm}}(\gamma) - \boldsymbol{\beta}\|_2$.

The proximity network that we used in this example was defined using the edges between every node and its four nearest (horizontal and vertical) neighbors. The smoothed estimates $\theta_{i,\mathrm{sm}} := \arctan(\frac{\hat{\beta}_{2,i,\mathrm{sm}}}{\hat{\beta}_{1,i,\mathrm{sm}}})$ and $r_{i,\mathrm{sm}} := \|\boldsymbol{\beta}_{i,\mathrm{sm}}\|_2$ are depicted in Figures 5, 6 and 7. In spite of the observation that θ_{sm} is less noisy than θ_{ml}, there are still areas where the fine structure around pinwheel centers and the border between clustered preferred orientations is disordered.

Figure 6 shows that r_{sm} is typically close to the true value of one, except for in neurons that lie at the border between regions with different orientation preferences. This is due to the fact that at regions that mark the border, tuning functions (and their noisy observations) point at significantly different directions, and therefore local averaging decreases the length of the average value. On the other hand, in smooth regions where vectors are pointing in roughly the same direction, local averaging preserves vector length.

The ability of our method to recover orientation preference maps from noisy recordings is shown in Figures 5, 6 and 7. To use the Bayesian formulation of equation (2), we substituted a fixed $\boldsymbol{X}_{\emptyset}$ for all $\boldsymbol{X}_i$. For λ^2, a Gamma($r = 1, \delta = 1$) was used based on the understanding that a priori $\frac{1}{p}\sum_{i\sim j}\|\beta_i - \beta_j\|_2$ should be $O(1)$. As for σ^2, the improper inverse-Gamma($\kappa' = 0, \epsilon' = 0$), that is, $\pi(\sigma^2) \propto 1/\sigma^2$, was used. $\{\hat{\boldsymbol{\beta}}_i\}$, namely, the posterior expectation of $\{\boldsymbol{\beta}_i\}$, is based on 10,000 samples from our efficient Gibbs sampler (after 500 burn-in iterations). The estimates $\hat{\sigma} = 0.4066 \pm 0.0001$ and $\hat{\lambda} = 11.13 \pm 0.01$ (i.e., the mean ± standard deviation) are based on the 10,000 samples. The following estimates of the preferred orientations and tuning strengths,

$$\hat{\theta}_i := \arctan\left(\frac{\hat{\beta}_{2,i}}{\hat{\beta}_{1,i}}\right),$$

$$\hat{r}_i := \|\hat{\boldsymbol{\beta}}_i\|_2,$$

are depicted in Figures 5 and 6. The posterior mean estimates of τ_x and τ_y (depicted in Figure 6) tend to be larger for neurons on the border of regions with similar preferred orientations $\{\theta_i\}$ (and less so around pinwheel centers), leading to minimal local smoothing for those pixels. Figure 6 shows that the Bayesian estimate $\hat{r}$ (like r_{sm}) underestimates the tuning strength for points that mark the border between different orientation preferences. In comparison to r_{sm}, as illustrated in the zoomed-in maps of Figure 7, this problem is less severe for the Bayesian estimate $\hat{r}$ because of the robust prior that decreases the strength of local averaging by increasing the local smoothing parameters $\{\tau_{ij}\}$ in regions marked with discontinuities.

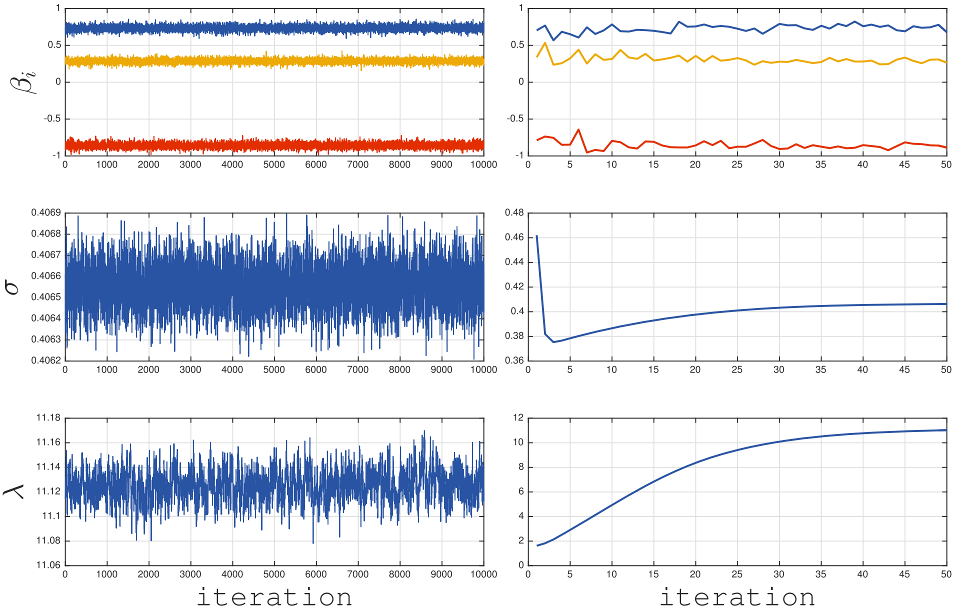


FIG. 8. *The sample path of* 3 *randomly selected pixels* (*top*), σ (*middle*) *and* λ (*bottom*). *The last* 10,000 (*after* 500 *burn-in*) *samples* (*left*) *and the first* 50 *samples* (*right*).

As we can see in Figure 7, the sharp border between similar orientation preferences is not over-smoothed while the noise among nearby neurons with similar orientation preferences is reduced. As a consequence of robustness, information is shared less among cells that lie at the border, but for cells that lie inside regions with smoothly varying preferred orientation, local smoothing is stronger. Moreover, in this example the chain appears to mix well (see Figure 8), and the Gibbs sampler is computationally efficient, requiring just a few seconds on a laptop (per iteration) to sample a surface described by $>10^6$ parameters.

Finally, let us add that it is well known that the semiregular, smoothly varying arrangement (with local discontinuities) of orientation preference maps is not a general feature of cortical architecture [Van Hooser et al. (2005)]. In fact, numerous electrophysiological and imaging studies [Girman, Sauvé and Lund (1999), Metin, Godement and Imbert (1988), Murphy and Berman (1979), Tiao and Blakemore (1976)] have found that orientation selective neurons in the visual cortex of many rodents are randomly arranged. A question that arises is whether the model would over-smooth if the neurons are not arranged smoothly in terms of their maps. In order to answer this question, we generated a randomly arranged orientation preference map, and applied our algorithm to the simulated neural activity in response to the same grating stimuli $\boldsymbol{X}_{\emptyset}$ used above. We also used the same noise variance ($\sigma = 0.4$) and the same priors for λ, σ and $\{\boldsymbol{\beta}\}_{i=1,\dots,n}$. Results are depicted in Figure 9. Since the preferred orientations lack spatial organization, the Bayesian estimate $\hat{\theta}$ of preferred orientations reverts to its respective θ_{ml}.

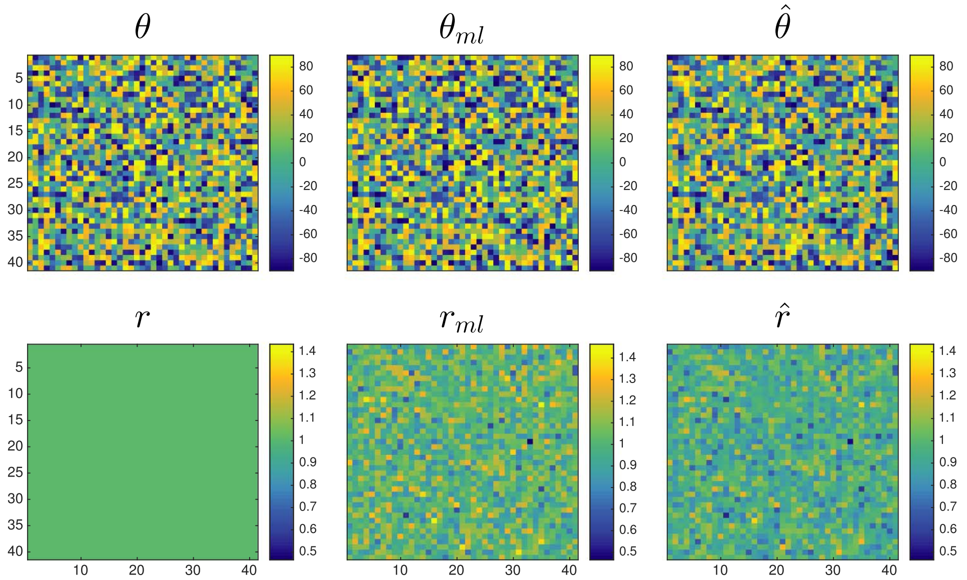


FIG. 9. *A* 40×40 *zoomed-in view of the* 710×710 (*not shown*) *randomly arranged preferred orientations* $\{\theta_i\}$ *and tuning strengths* $\{r_i\}$, *and their estimates. The orientation at each pixel was randomly drawn from a uniform distribution on* $(-90°, +90°]$. *Since the preferred orientations lack spatial organization, the Bayesian estimate* $\hat{\theta}$ *of preferred orientations reverts to its respective* θ_{ml}. *The posterior estimates are based on* 10,000 *consecutive iterations of the Gibbs sampler* (*after* 500 *burn-in iterations*).

4.2. *Real data.*

4.2.1. *Phasic tuning in motor neurons.* We next tested the method's performance on real neural imaging data obtained from an isolated mouse spinal cord preparation [schematized in Figure 10(a)]. In these data, the fluorescent activity sensor GCaMP3 was expressed in motor neurons that innervate leg muscles. After application of a cocktail of rhythmogenic drugs, all motor neurons in the preparation fire in a periodic bursting pattern mimicking that seen during walking [Machado et al. (2015)]. Under these conditions, we acquired sequences of fluorescent images and then applied a model-based constrained deconvolution algorithm to infer the timing of neuronal firing underlying each fluorescent activity time series extracted from the pixels corresponding to individual neurons [Pnevmatikakis et al. (2014a)].

Each mouse leg is controlled by ∼50 different muscles, each of which is innervated by motor neurons that fire in distinct patterns during locomotor behavior [Akay et al. (2014), Krouchev, Kalaska and Drew (2006)]. Furthermore, all motor neurons that share common muscle targets are spatially clustered together into "pools" within the spinal cord [Romanes (1964)]. Therefore, during the locomotor-like network state monitored in these data, different spatially distinct groups of

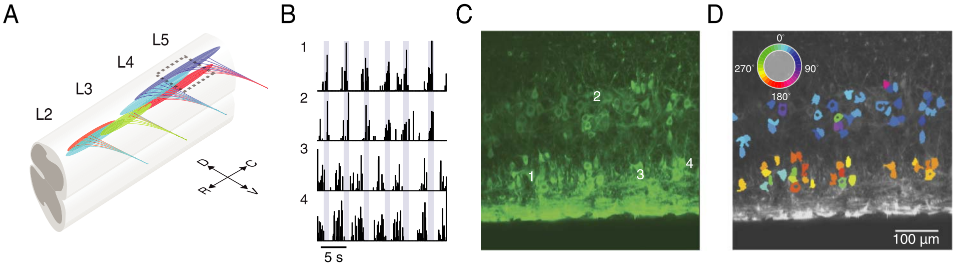
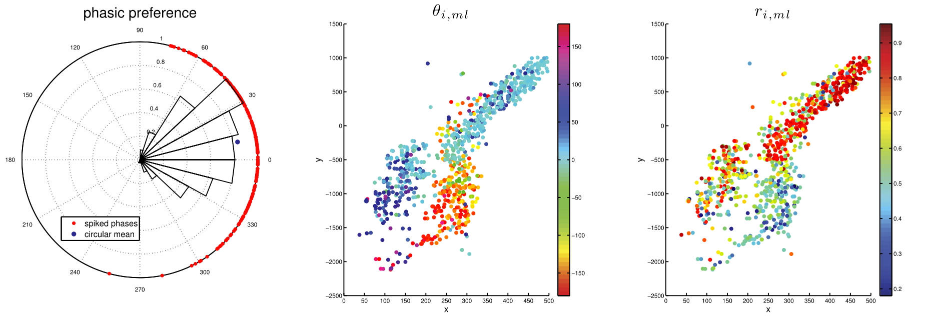


FIG. 10. *Isolated spinal cord imaging preparation.* (a) *Schematic of isolated spinal cord imaging preparation.* (b) *Activity inferred from fluorescence measurements obtained from four motor neurons. Height of black bars indicates intensity of neuronal activity at each time point. Vertical blue bars indicate the onset of each locomotor cycle* (*i.e.*, 0°). (c) *Example fluorescent imaging field with the position of the four neurons shown in* (b) *indicated.* (d) *Each motor neuron shown in* (c) *is represented in color* (*legend in inset*) *corresponding to its estimated tuning value.*

motor neurons are recruited to fire at each moment in time [Figure 10(b)–(d)]. When the activity of each motor neuron is summarized as a single mean phase tuning value [representing the average phase angle of the ∼70 firing events detected per neuron, as seen in Figure 11(a)], a clear spatial map can be derived [Figure 10(d)]. Such maps appear smooth within pools and sharply discontinuous between pools.



(a) Phasic preference of one cell.

(b) Preferred phases of $n = 854$ cells.

(c) Tuning strengths of $n = 854$ cells.

FIG. 11. (a) *Noisy observations of phases at which this cell has fired. The phase of each red dot on the unit circle is a phase at which this cell has fired, and the angular histogram depicts its distribution. The blue dot is the circular mean of all red dots, and its phase and length are the ML estimates of the preferred phase and tuning strength, respectively.* (b, c) *The three-dimensional spatial cell position is projected into the two-dimensional x–y plane. Each dot indicates one cell*; *each cell is color coded with the phase* $\theta_{i,\mathrm{ml}}$ *or tuning strength* $r_{i,\mathrm{ml}}$. *Preferred phases* (*and tuning strengths*) *tend to be similar among nearby cells, but not all nearby cells have similar preferred phases* (*and tuning strengths*).

While phase tuning can be reliably inferred one neuron at a time in these data, fluorescent measurements from each neuron are not always of high quality. As a result, activity events cannot be reliably inferred from all neurons [Machado et al. (2015)]. Additionally, more neurons could have been observed with less data per neuron if phase tuning was estimated more efficiently. Therefore, we applied our robust and scalable Bayesian information sharing algorithm to these data in an attempt to reduce measurement noise and decrease the required data necessary to attain precision tuning map measurements.

In this setting, let us introduce some simplifying notation. We use ℓ_i to denote the total number of spikes that neuron i has fired. As mentioned earlier, $\ell_i \sim 70$ here. Furthermore, we use $\theta_{i,\ell}$ to denote the ℓth phase at which neuron i has fired a spike. Then we convert this phase $\theta_{i,\ell}$ to $\boldsymbol{y}_{i,\ell} := [\cos(\theta_{i,\ell})\,\sin(\theta_{i,\ell})]'$, a point on the unit circle.

We model the neuron's tendency to spike at phases that are concentrated around a certain angle using a two-dimensional vector $\boldsymbol{\beta}_i$. The direction of $\boldsymbol{\beta}_i$ is the preferred phase θ_i, and the length of $\boldsymbol{\beta}_i$ is the tuning strength r_i. If the neuron is highly tuned, that is, there is no variability among phases at which this neuron fires a spike, then $r_i = 1$ and $\boldsymbol{\beta}_i$ lies on the unit circle. On the other hand, if the neuron is weakly tuned, that is, there is large variability among phases at which this neuron fires a spike, then $r_i \sim 0$. We relate observation $\boldsymbol{y}_{i,\ell}$ to the unknown $\boldsymbol{\beta}_i := r_i[\cos\theta_i\,\sin\theta_i]'$ as follows:

$$\boldsymbol{y}_{i,\ell}|\boldsymbol{\beta}_i, \sigma, \nu_i \sim \mathcal{N}(\boldsymbol{\beta}_i, \nu_i^2\sigma^2\boldsymbol{I}) \qquad \text{for } \ell = 1, \dots, \ell_i,$$

where $\boldsymbol{\beta}_i$ is related to θ_i (preferred phase) and r_i (tuning strength) as follows:

$$\theta_i := \arctan\left[\frac{\beta_{2,i}}{\beta_{1,i}}\right],$$

$$r_i := \sqrt{\beta_{2,i}^2 + \beta_{1,i}^2}.$$

There are two points worth mentioning. First, the Gaussian noise model clearly violates the fact that $\{\boldsymbol{y}_{i,\ell}\}$ lie on the unit circle, and should therefore be considered a rather crude approximation. Nevertheless, as demonstrated below, this Gaussian likelihood with our prior in (2) is remarkably effective in estimating the preferred phases $\{\theta_i\}$ with as little as one observed phase per neuron. Second, the vector representation of the ℓ_i spikes that neuron i has fired,

$$\boldsymbol{y}_i = \begin{bmatrix} \boldsymbol{y}_{i,1} \\ \vdots \\ \boldsymbol{y}_{i,\ell_i} \end{bmatrix}_{2\ell_i \times 1},$$

can be related to the unknown $\boldsymbol{\beta}_i$ using the formulation presented in equation (1), where

$$\boldsymbol{X}_i = \begin{bmatrix} \vdots \\ \boldsymbol{I}_{2\times 2} \\ \vdots \end{bmatrix}_{2\ell_i \times 2}.$$

The ML estimate of $\boldsymbol{\beta}_i$, given the Gaussian additive noise model, is the sample mean of the observations $\{\boldsymbol{y}_{i,\ell}\}_{\ell=1,\dots,\ell_i}$, $\boldsymbol{\beta}_{i,\mathrm{ml}} = \frac{1}{\ell_i}\sum_{\ell=1}^{\ell_i} \boldsymbol{y}_{i,\ell}$. The ML estimate of the preferred phase $\theta_{i,\mathrm{ml}} = \arctan[\frac{\beta_{2,i,\mathrm{ml}}}{\beta_{1,i,\mathrm{ml}}}]$ is the circular mean of the observed phases, as depicted in Figure 11(a). The resulting radius $\|\boldsymbol{\beta}_{i,\mathrm{ml}}\|_2$, the ML estimate of r_i, will be 1 if all angles are equal. If the angles are uniformly distributed on the circle, then the resulting radius will be 0, and there is no circular mean. The radius measures the concentration of the angles and can be used to estimate confidence intervals.

In addition to the observed phases, we also have the three-dimensional physical location of all cells. As an illustrative example, the spatial distribution of $\{\theta_{i,\mathrm{ml}}\}$ and $\{r_{i,\mathrm{ml}}\}$ is depicted in Figures 11(b) and (c). The three-dimensional location is projected into the two-dimensional x–y plane. Each dot is a cell, and its color in panel 11(b) and (c) corresponds to $\theta_{i,\mathrm{ml}}$ and $r_{i,\mathrm{ml}}$, respectively. Clearly, nearby cells tend to have similar preferred phases and tuning strengths—but there are many exceptions to this trend. A mixture prior is required to avoid over-smoothing the border between clusters of cells with similar properties while allowing cells within a cluster to share information and reduce noise.

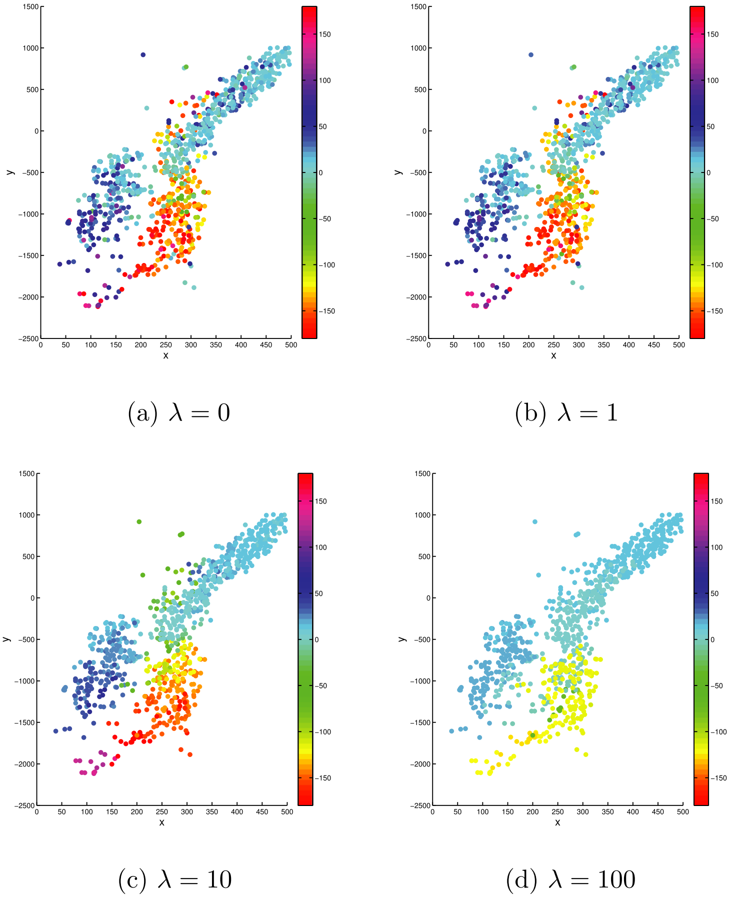
In order to include the physical location of the cells into our Bayesian formulation, we formed a proximity network based on nearest spatially whitened neighbors, as described in Section 2. $\{\hat{\boldsymbol{\beta}}_i\}$, the posterior expectation of $\{\boldsymbol{\beta}_i\}$, is based on 10,000 samples from our efficient Gibbs sampler (after 500 burn-in iterations). For illustration purposes, we experimented with holding the hyperparamter λ fixed in the simulations; the effects of this hyperparameter on the estimates of the preferred phases and tuning strengths,

$$\hat{\theta}_i := \arctan\left(\frac{\hat{\beta}_{2,i}}{\hat{\beta}_{1,i}}\right), \tag{15}$$

$$\hat{r}_i := \|\hat{\boldsymbol{\beta}}_i\|_2, \tag{16}$$

are depicted in Figure 12. It is clear that large λ forces nearby neurons to have more similar preferred phases, whereas for small λ the preferred phases revert to their respective ML estimates.

The ability of our method to recover the preferred phases from as little as one noisy phase $\theta_{i,\ell}$ per neuron is illustrated below. We divide the data into two parts.



(a) $\lambda = 0$

(b) $\lambda = 1$

(c) $\lambda = 10$

(d) $\lambda = 100$

FIG. 12. *Preferred phase estimates for different values of the hyperparameter* λ. *Each dot corresponds to the estimated preferred angle* $\hat{\theta}_i$ *for one cell. For* $\lambda = 0$, *the estimates are equal to the ML estimates. For* $\lambda = 1$, *information sharing is not large enough and estimates are not very different from the ML estimates. For* $\lambda = 10$, *nearby neurons are forced to have similar preferred phases, nonetheless, the sharp border between functionally different clusters of neurons is not over-smoothed. The posterior mean and standard deviation of* λ, *based on* 10,000 *iterations* (*after* 500 *burn-ins*), *are* 5.26 *and* 0.52, *respectively. For* $\lambda = 100$, *smoothing within clusters is stronger and borders are not violated. However, tuning estimates within each cluster suffer from over-smoothing.*

For each cell, there are roughly 70 phases recorded (at which the corresponding neuron fired). For each neuron i, we randomly selected one of the phases $\{\theta_{i,\ell}\}_{\ell=1,\dots,\ell_i}$ for the training set, and let the rest of the phases constitute the testing set:

$$\boldsymbol{y}_{i,\mathrm{train}} := \begin{pmatrix} \cos(\theta_{i,\ell_{\mathrm{train}}}) \\ \sin(\theta_{i,\ell_{\mathrm{train}}}) \end{pmatrix}, \qquad \boldsymbol{y}_{i,\mathrm{test}} := \frac{1}{\ell_i - 1} \sum_{\substack{\ell=1,\dots,\ell_i \\ \ell \neq \ell_{\mathrm{train}}}} \begin{pmatrix} \cos(\theta_{i,\ell}) \\ \sin(\theta_{i,\ell}) \end{pmatrix}.$$

The raw estimates of preferred phases and tuning strengths using training data are computed as follows:

$$\theta_{i,\text{train}} := \arctan\left(\frac{y_{2,i,\text{train}}}{y_{1,i,\text{train}}}\right), \qquad r_{i,\text{train}} := \|\boldsymbol{y}_{i,\text{train}}\|_2;$$

and raw estimates of preferred phases and tuning strengths using testing data are computed likewise. For $\{\boldsymbol{y}_i\}$ in our Gibbs sampler, we use the training data $\{\boldsymbol{y}_{i,\text{train}}\}$. Posterior estimates $\{\hat{\theta}_i, \hat{r}_i, \hat{\nu}_i\}$ for four distinct datasets are depicted and compared against testing data in Figures 13–16. For λ^2 we use a Gamma($r = 1$, $\delta = 1$) prior, and for σ^2 an improper inverse-Gamma($\kappa' = 0, \epsilon' = 0$) prior. Generally speaking, σ and λ are not identifiable. Furthermore, the joint posterior distribution of β and σ is only unimodal given $\{\nu_i^2\}$. We address both challenges by placing a relatively tight prior on $\{\nu_i^2\}$. We use independent inverse-Gamma($\varkappa = 3$, $\varepsilon = 2$) priors for $\{\nu_i^2\}$, making the prior means and variances equal to one. Since the posterior distribution of β and σ is only unimodal given $\{\nu_i\}$, this prior constrains the ν_is such that the posterior distribution stays nearly unimodal. Finally, λ and σ stay nearly identifiable given this tight prior.

The raw training estimates of preferred phases and tuning strengths are very noisy, which is expected given the fact that only one phase per neuron is used. This is an extremely low signal-to-noise limit. In contrast, roughly 70 phases per neuron are used to compute the raw testing estimates. The Bayesian estimates $\{\hat{\theta}_i, \hat{r}_i\}$ are also based on one phase per neuron, but they employ the a priori knowledge that the activity of a neuron carries information about its nearby neurons. As mentioned earlier, this is done by incorporating the proximity network into the Bayesian formulation.

Moreover, as illustrated in the middle panels of Figures 13–16, the Bayesian estimates respect the border of clustered cells with similar phasic preferences and tuning strengths. Information is not invariably shared among nearby cells; instead, it is based on how locally similar the samples of $\{\boldsymbol{\beta}_i\}$ are. If the estimated typical noise is much less than the local difference, then, intuitively speaking, local smoothing should be avoided because the difference seems statistically significant.

In contrast, the raw test estimates $\{\theta_{i,\text{test}}, r_{i,\text{test}}\}$ are computed in isolation (one neuron at a time), but use roughly 70 phases per neuron (high signal-to-noise). The Bayesian estimates are less noisy in comparison to the raw training estimates (low signal-to-noise) and qualitatively resemble the raw test estimates (high signal-to-noise). Unlike the previous synthetic data example, here the true parameters are unknown. In order to quantify the noise reduction, we treat the high signal-to-noise test estimates as the unknown true parameters, and compare them against the Bayesian estimates. Recall that the Bayesian estimates are based on the low signal-to-noise raw train data. We quantify the noise reduction by comparing the testing error $\frac{1}{n}\sum|\hat{\theta}_i - \theta_{i,\text{test}}|$ with the raw error $= \frac{1}{n}\sum|\theta_{i,\text{train}} - \theta_{i,\text{test}}|$. The test error is 10°–16° less than the raw error; for more details see the the captions of

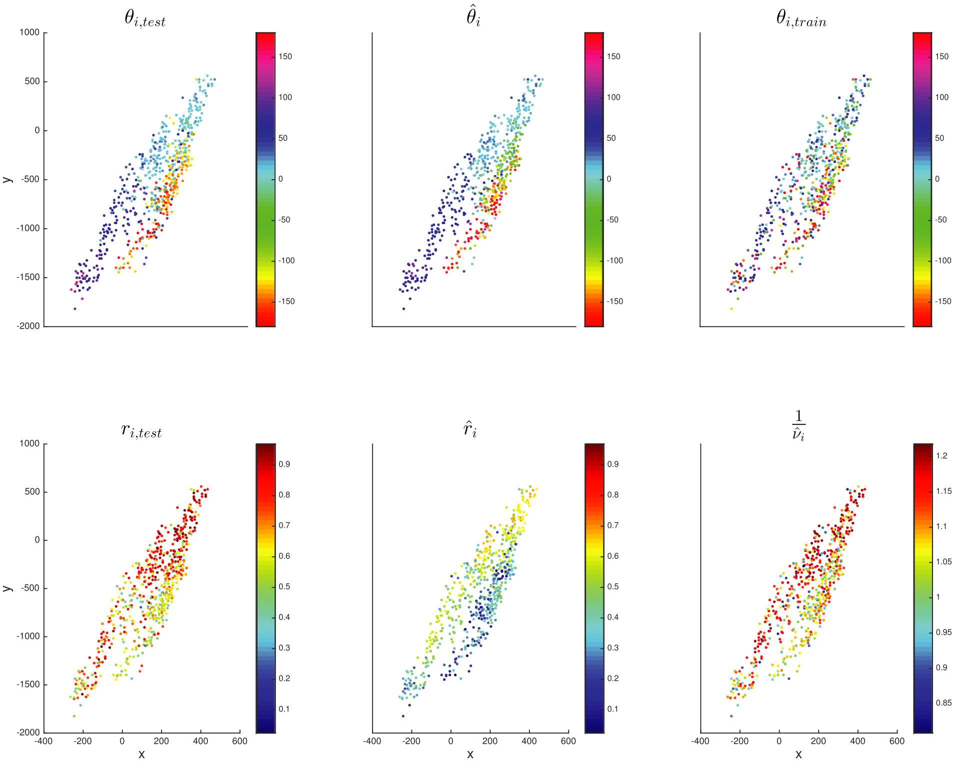


FIG. 13. *Dataset* 1 *with* $n = 584$. *The posterior estimates* $\hat{\sigma} = 0.62 \pm 0.02$ *and* $\hat{\lambda} = 6.43 \pm 0.38$ (*i.e., the mean* ± *standard deviation*) *are based on* 10,000 *samples* (*after* 500 *burn-ins*). *The test set is made of* 59 *observed phases per neuron. The test error is* $\frac{1}{n}\sum|\hat{\theta}_i - \theta_{i,\text{test}}| = 27.6°$, *and the raw error is* $\frac{1}{n}\sum|\hat{\theta}_{i,\text{train}} - \theta_{i,\text{test}}| = 36.9°$.

Figures 13–16. Last, the boxplots in Figure 17 summarize and quantify the noise reduction due to our robust Bayesian information sharing approach. In each case, the new Bayesian approach provides significant improvements on the estimation accuracy.

5. Concluding remarks. We developed a robust and scalable Bayesian smoothing approach for inferring tuning functions from large-scale high-resolution spatial neural activity, and illustrated its application in a variety of neural coding settings. A large body of work has addressed the problem of estimating a smooth spatial process from noisy observations [Besag (1974), Besag and Kooperberg (1995), Rasmussen and Williams (2006), Rue and Held (2005), Wahba (1990)]. These ideas have found many of their applications in problems involving tuning function estimation [Cunningham et al. (2008, 2009), Czanner et al. (2008), Gao et al. (2002), Macke et al. (2010, 2011), Paninski (2010), Paninski et al. (2010), Pnevmatikakis et al. (2014b), Rahnama Rad and Paninski (2010)]. There has also

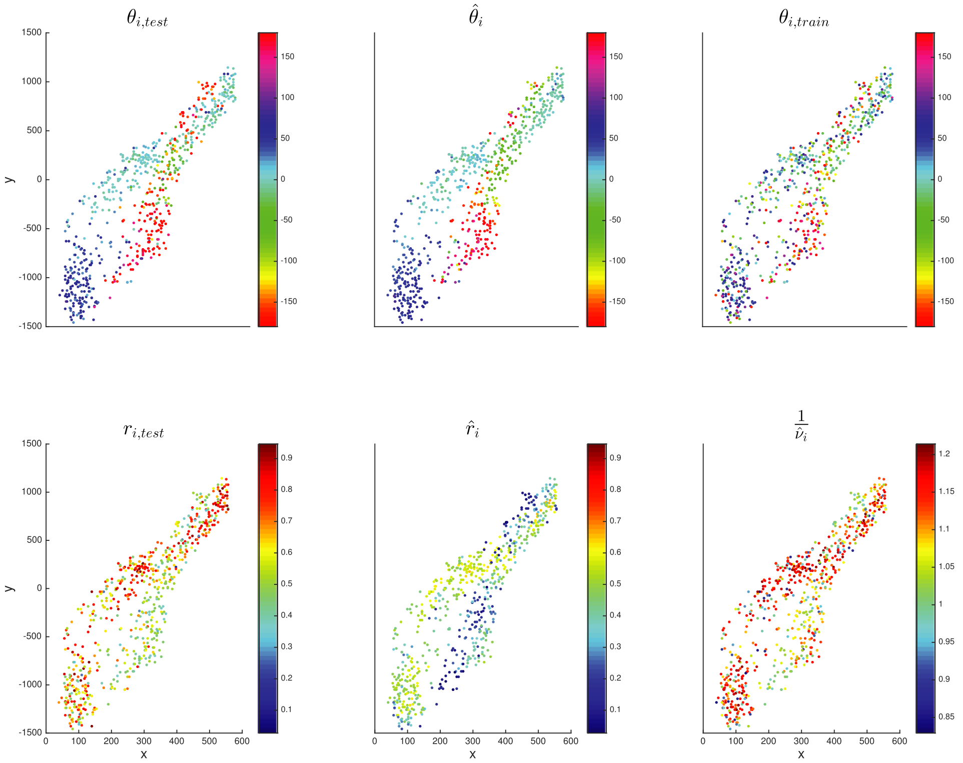


FIG. 14. *Dataset* 16 *with* $n = 676$. *The posterior estimates* $\hat{\sigma} = 0.62 \pm 0.01$ *and* $\hat{\lambda} = 7.04 \pm 0.38$ (*i.e.*, *the mean* ± *standard deviation*) *are based on* 10,000 *samples* (*after* 500 *burn-ins*). *The test set is made of* 60 *phases per neuron. The test error is* $\frac{1}{n}\sum|\hat{\theta}_i - \theta_{i,\text{test}}| = 28.4°$, *and the raw error is* $\frac{1}{n}\sum|\hat{\theta}_{i,\text{train}} - \theta_{i,\text{test}}| = 41.5°$.

been some work on parametric Bayesian tuning function estimation [see Cronin et al. (2010) and references therein]. The main challenge in the present work was the large scale (due to the high spatial resolution) of the data and the functional discontinuities present in neuronal tuning maps [e.g., Machado et al. (2015), Ohki et al. (2005), Shmuel and Grinvald (1996)].

In order to address these challenges, we proposed a robust prior as part of a computationally efficient block Gibbs sampler that employs fast Gaussian sampling techniques [Hoffman (2009), Hoffman and Ribak (1991), Papandreou and Yuille (2010)] and the Bayesian formulation of the Lasso problem [Casella et al. (2010), Park and Casella (2008)]. This work focused especially on the conceptual simplicity and computational efficiency of the block Gibbs sampler: we emphasized the robustness properties of the Bayesian Lasso, the unimodality of the posterior and the use of efficient linear algebra methods for sampling, which avoid the Cholesky decomposition or other expensive matrix decompositions. Using in vitro recordings from the spinal cord, we illustrated that this approach can effectively

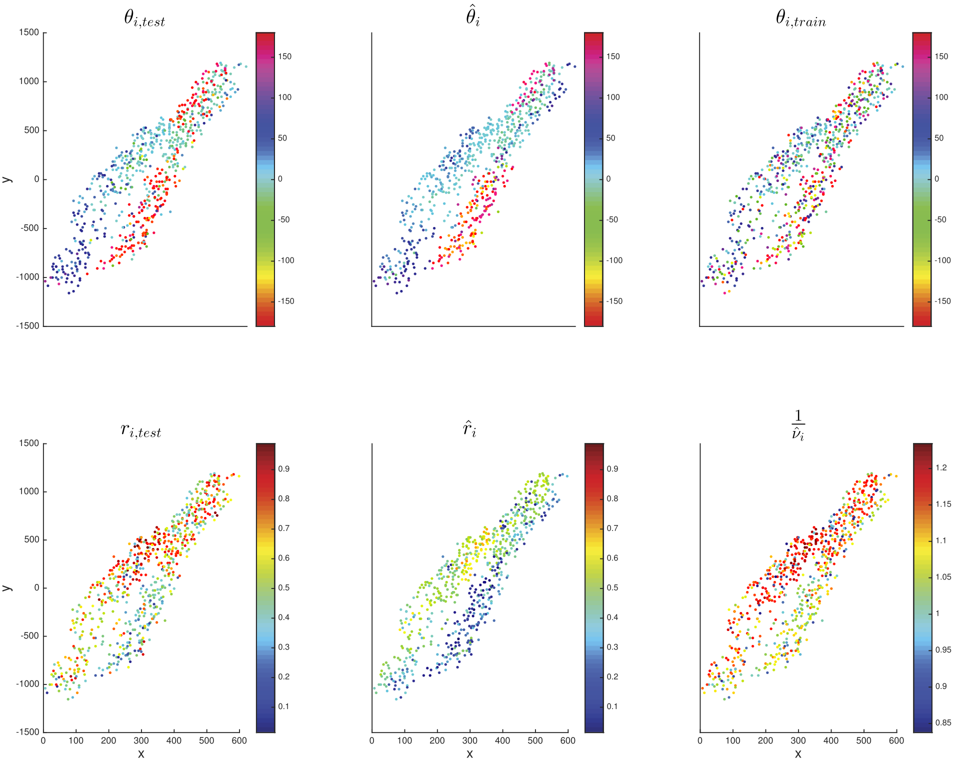


FIG. 15. *Dataset* 23 *with* $n = 695$. *The posterior estimates* $\hat{\sigma} = 0.69 \pm 0.02$ *and* $\hat{\lambda} = 7.39 \pm 0.39$ (*i.e., the mean* ± *standard deviation*) *are based on* 10,000 *samples* (*after* 500 *burn-ins*). *The test set is made of* 73 *phases per neuron. The test error is* $\frac{1}{n}\sum|\hat{\theta}_i - \theta_{i,\mathrm{test}}| = 39.45^\circ$, *and the raw error is* $\frac{1}{n}\sum|\hat{\theta}_{i,\mathrm{train}} - \theta_{i,\mathrm{test}}| = 51.55^\circ$.

infer tuning functions from noisy observations, given a negligible portion of the data and reasonable computational time.

It is worth mentioning that in another line of work, smoothness inducing priors were used to fit spatio-temporal models to fMRI data [Groves, Chappell and Woolrich (2009), Harrison and Green (2010), Penny, Trujillo-Barreto and Friston (2005), Quiros, Diez and Gamerman (2010), Woolrich (2012)]. Although these priors handle spatial correlation in the data, they do not always successfully account for spatial discontinuities and the large scale of the data. Woolrich et al. (2004) used automatic relevance determination (ARD) [MacKay (1995)] to allow for spatially nonstationary noise where the level of smoothness at each voxel was estimated from the data. It is known [Wipf and Nagarajan (2008)] that ARD can converge slowly to suboptimal local minima. On the other hand, wavelet bases with a sparse prior, defined by a mixture of two Gaussian components, allowed Guillaume and Penny (2007) to present a statistical framework for modeling transient, nonstationary or spatial varying phenomenon. They used variational Bayes

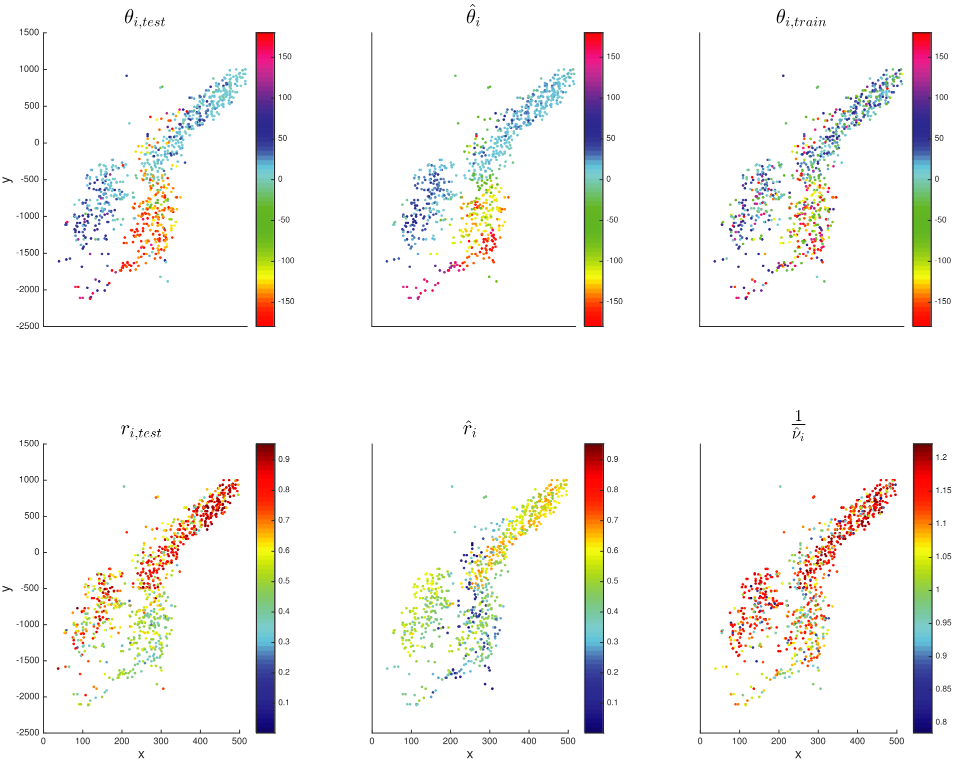


FIG. 16. *Dataset* 26 *with* $n = 854$. *The posterior estimates* $\hat{\sigma} = 0.62 \pm 0.01$ *and* $\hat{\lambda} = 7.52 \pm 0.40$ (*i.e.*, *the mean* ± *standard deviation*) *are based on* 10,000 *samples* (*after* 500 *burn-ins*). *The test set is made of* 82 *phases per neuron. The test error is* $\frac{1}{n}\sum|\hat{\theta}_i - \theta_{i,\text{test}}| = 27.1°$, *and the raw error is* $\frac{1}{n}\sum|\hat{\theta}_{i,\text{train}} - \theta_{i,\text{test}}| = 42.9°$.

approximations together with fast orthogonal wavelet transforms to efficiently compute the posterior distributions. As mentioned in their paper, a main drawback is that wavelet denoising with an orthogonal transform exhibits Gibbs phenomena around discontinuities, leading to inefficient modeling of singularities, such as edges. In Grosenick et al. (2013), Harrison et al. (2015), Slawski (2012), Slawski, Zu Castell and Tutz (2010), van Gerven et al. (2010), smoothness (and matrix factorization) approaches were combined with various global sparsity-inducing priors (or regularizers) to smooth (or factorize) the spatio-temporal activity of voxels that present significant effects, and to shrink to zero voxels with insignificant effects. In Harrison et al. (2007), nonstationary Gaussian processes were used as adaptive filters with the computational disadvantage of inverting large covariance matrices. Finally, in a recent work, Siden et al. (2016) design an efficient Monte Carlo sampler to perform spatial whole-brain Bayesian smoothing. Costly Cholesky decompositions are avoided by efficiently employing the sparsity of precision matrices and preconditioned conjugate gradient methods. The prior in Siden et al. (2016)

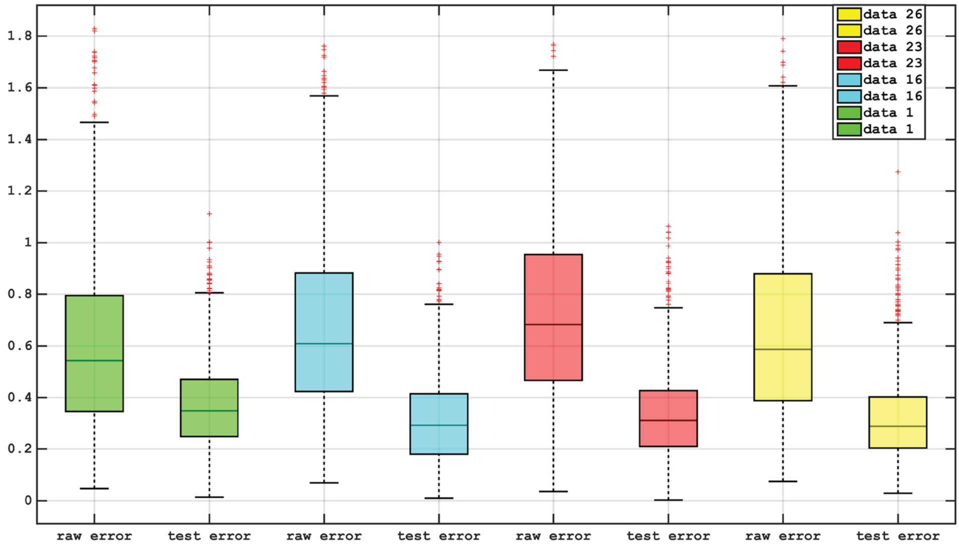


FIG. 17. *Tukey boxplots comparing the raw error and test error for the four datasets illustrated in Figures 13–16. The raw and test errors (for cell i) are defined as $\|\mathbf{y}_{i,\text{train}} - \mathbf{y}_{i,\text{test}}\|_2$ and $\|\hat{\boldsymbol{\beta}}_i - \mathbf{y}_{i,\text{test}}\|_2$, respectively.*

assigns a spatially homogenous level of smoothness which performs less favorably in situations involving outliers and sharp breaks in the functional map.

There is also a vast literature addressing the recovery of images from noisy observations [see Buades, Coll and Morel (2005), Motwani et al. (2004) and references therein]. Most of these techniques use some sort of regularizer or prior to successfully retain image discontinuities and remove noise.

Early examples include the auxiliary line process-based quadratic penalty in Geman and Geman (1984) and the $\sum_i \frac{1}{1+|\nabla_i \boldsymbol{\beta}|}$ log prior in Geman and Reynolds (1992), where the gradient at i is denoted by ∇_i. The line process indicates sharp edges and suspends or activates the smoothness penalty associated with each edge. The log prior $\sum_i \frac{1}{1+|\nabla_i \boldsymbol{\beta}|}$ encourages the recovery of discontinuities while rendering auxiliary variables of the line process as unnecessary. These log priors are nonconcave. These nonconcave maximum a posteriori optimization problems are generally impractical to maximize. Different techniques were designed based on simulated annealing [Geman and Geman (1984), Geman and Reynolds (1992), Geman and Yang (1995)], coarse-to-fine optimization [Bouman and Liu (1988)] and alternate maximization between image and auxiliary contour variables [Charbonnier et al. (1997)] to compute (nearly) global optimums at the expense of a prohibitively large amount of computation. Moreover, it is known that a small perturbation in the data leads to abrupt changes in the denoised image [Bouman and Sauer (1993)]. This is due to the nonconcavity of the problem.

Image denoising methods based on concave log priors (or regularizers) that enjoy edge preserving properties were designed in Bouman and Sauer (1993), Green (1990), Rudin, Osher and Fatemi (1992), Stevenson and Delp (1990a), Stevenson and Delp (1990b). These log priors typically take the form of $-\sum_i \phi(\nabla_i \boldsymbol{\beta})$ for some concave function $\phi(\cdot)$. Examples of $\phi(x)$ include the Huber function [Huber (1964)] in Stevenson and Delp (1990a, 1990b), $\log\cosh(\frac{x}{T})$ in Green (1990), $|x|^p$ where $1 \leq p < 2$ [Bouman and Sauer (1993)] and $|x|$ in the TV penalty [Besag (1993), Rudin, Osher and Fatemi (1992)]. Various methods have been proposed for computing optimal or nearly optimal solutions to these image recovery problem, for example, Afonso, Bioucas-Dias and Figueiredo (2010), Barbero and Sra (2011), Bouman and Sauer (1993), Chambolle (2004), Defrise, Vanhove and Liu (2011), Green (1990), Oliveira, Bioucas-Dias and Figueiredo (2009), Rudin, Osher and Fatemi (1992), Stevenson and Delp (1990a, 1990b), Vogel and Oman (1996, 1998), Wang et al. (2009).

In addition to these approaches, another significant contribution has been to consider wavelet, ridgelet and curvelet-based priors/regularizers, for example, Candes (1999a, 1999b), Donoho and Johnstone (1994), Portilla et al. (2003), Starck, Candes and Donoho (2002), which present noticeable improvements in image reconstruction problems. More recently, the nonlocal means method [Buades, Coll and Morel (2005), Dabov et al. (2007), Lebrun, Buades and Morel (2013)] presents a further improvement. However, most of these methods' favorable performance relies heavily on parameters which have been fine tuned for specifically additive noisy observations of two-dimensional arrays of pixels of real world images. In other words, they are specifically tailored for images. It is not clear if and how these approaches can be modified to retain their efficiency while being applied to a broader class of spacial observations lying on generic graphs. Moreover, the denoised images rarely come equipped with confidence intervals. But our sampling-based approach allows for proper quantification of uncertainty, which could in turn be used to guide online experimental design; for example, in the spinal cord example analyzed here, we could choose to record more data from neurons with the largest posterior uncertainty about their tuning functions.

In principle, many of abovementioned approaches can be formulated as Bayesian, with the aid of the Metropolis–Hastings (MH) algorithm, to compute posterior means and standard deviations [Lassas and Siltanen (2004), Louchet and Moisan (2013)]. However, generic MH approaches can lead to unnecessary high computational cost. For example, in Lassas and Siltanen (2004) a TV prior and Gaussian noise model was used to denoise a one-dimensional pulse; it was reported that the chain resulting from the MH algorithm suffers from very slow convergence. One contribution of the present paper is to show that by using a hierarchical representation of our prior in equation (5) costly MH iterations can be avoided in all steps of our block Gibbs sampler. Additionally, we show how our model can take into account nonuniform noise variance (quite common in neuroscience applications) without increasing the computational complexity. Finally, we emphasize the importance of conditioning $\boldsymbol{\beta}$ on σ in equation (2), which has been neglected in previous Bayesian formulations of the TV prior [Lassas and Siltanen (2004), Louchet and Moisan (2013)]. This is important because it guarantees a unimodal posterior of $\boldsymbol{\beta}$ and σ given $\{\nu_i\}_{i=1,\dots,n}$ and λ.

We should also note that a number of fully Bayesian methods have been developed that present adaptive smoothing approaches for modeling nonstationary spatial data. These methods are predicated on the idea that, to enhance spatial adaptivity, local smoothness parameters should a priori be viewed as a sample from a common ensemble. Conditional on these local smoothing parameters, the prior is a Gaussian Markov random field (GMRF) with a rank deficient precision matrix [Fahrmeir, Kneib and Lang (2004), Lang and Brezger (2004), Lang, Fronk and Fahrmeir (2002), Rue and Held (2005), Yue, Loh and Lindquist (2010), Yue and Speckman (2010), Yue, Speckman and Sun (2012)]. The hyperprior for the local smoothing parameters can be specified in two ways. The simpler formulation assumes the local smoothing parameters to be independent [Brezger, Fahrmeir and Hennerfeind (2007), Fahrmeir, Kneib and Lang (2004), Lang and Brezger (2004), Lang, Fronk and Fahrmeir (2002)]. For example, Lang, Fronk and Fahrmeir (2002) presented a nonparametric prior for fitting unsmooth and highly oscillating functions based on a hierarchical extension of state-space models where the noise variance of the unobserved states is locally adaptive. The main computational burden lies on the Cholesky decomposition [Brezger, Fahrmeir and Hennerfeind (2007)] or other expensive matrix decompositions of the precision matrix. In a more complex formulation, the log-smoothing parameters follow another GMRF on the graph defined by edges $i \sim j$ [Yue, Loh and Lindquist (2010), Yue and Speckman (2010), Yue, Speckman and Sun (2012)]. In both formulations, local smoothing parameters are conditionally dependent, rendering Metropolis-within-Gibbs sampling necessary. These methods often provide superior estimation accuracy for functions with high spatial variability on regular one-dimensional and two-dimensional lattices, but at a prohibitively higher computational cost which makes them less attractive for the high-dimensional datasets considered in this paper. One interesting direction for future work would be to combine the favorable properties of these approaches with those enjoyed by our scalable and robust Bayesian method.

Finally, important directions for future work involve extensions that allow the treatment of point processes, or other non-Gaussian data, and correlated neural activities. Since our prior can be formulated in a hierarchical manner, when dealing with non-Gaussian likelihoods, it is only step 2 of our Gibbs sampler that needs modification. In step 2, all MCMC algorithms suited for Gaussian priors and non-Gaussian likelihoods can be integrated into our efficient Gibbs sampler. For example, the elliptical slice sampler [Murray, Adams and MacKay (2010)] or Hamiltonian Monte Carlo methods [Ahmadian, Pillow and Paninski (2011), Duane et al. (1987), Girolami, Calderhead and Chin (2011), Robert and Casella (2004), Roberts and Stramer (2002)] are well suited for sampling from posteriors arising from a Gaussian prior and likelihoods from the exponential family. With regard to correlated neural activities, it would be interesting to see how tools developed in Buesing, Macke and Sahani (2012), Vidne et al. (2012) can be incorporated into our Gibbs sampler to make inference about models which can account for correlated observations.

APPENDIX: UNIMODALITY OF THE POSTERIOR

Here we demonstrate that the joint posterior of $\boldsymbol{\beta}$ and σ^2 given $\{\nu_i\}_{i=1,\dots,n}$ and λ is unimodal under the prior in equation (2) and equation (3). Note that our discussion here is very similar to that of Park and Casella (2008). The joint prior is

$$p(\boldsymbol{\beta}, \sigma^2|\lambda) = \frac{\epsilon^\kappa}{\Gamma(\kappa)}(\sigma^2)^{-\kappa-1}e^{-\epsilon/\sigma^2}\prod_{i\sim j}\left(\frac{\lambda}{2\sigma}\right)^m \exp\left(-\frac{\lambda}{\sigma}\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|_2\right).$$

The log posterior is

$$\begin{aligned}&-\left(\kappa + 1 + \frac{nd + pm}{2}\right)\log\sigma^2 - \frac{\epsilon}{\sigma^2} - \frac{\lambda}{\sqrt{\sigma^2}}\sum_{i\sim j}\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|_2\\ &\quad - \frac{1}{2\sigma^2}\sum_{i=1}^n \frac{\|\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}_i\|_2^2}{\nu_i^2},\end{aligned} \tag{17}$$

ignoring all the terms independent of $\boldsymbol{\beta}$ and σ^2. The mapping (and its inverse)

$$\boldsymbol{\phi}_i \leftrightarrow \frac{\boldsymbol{\beta}_i}{\sqrt{\sigma^2}}, \qquad \rho \leftrightarrow \frac{1}{\sqrt{\sigma^2}} \tag{18}$$

is continuous. Therefore, unimodality in the mapped coordinates is equivalent to unimodality in the original coordinates. The log posterior (17) in the new coordinates is

$$\begin{aligned}&(2\kappa + 2 + nd + pm)\log\rho - \epsilon\rho^2 - \lambda\sum_{i\sim j}\|\boldsymbol{\phi}_i - \boldsymbol{\phi}_j\|_2\\ &\quad - \frac{1}{2}\sum_{i=1}^n\|\rho\underline{\boldsymbol{y}}_i - \underline{\boldsymbol{X}}_i\boldsymbol{\phi}_i\|_2^2,\end{aligned} \tag{19}$$

where we have earlier defined in equation (7)

$$\underline{\boldsymbol{y}}_i := \frac{\boldsymbol{y}_i}{\nu_i}, \qquad \underline{\boldsymbol{X}}_i := \frac{\boldsymbol{X}_i}{\nu_i}.$$

The log posterior in equation (19) is clearly concave in $(\boldsymbol{\phi}_1, \dots, \boldsymbol{\phi}_n, \rho)$, and hence the posterior is unimodal.

Acknowledgments. We are grateful to A. Brezger, B. Babadi, J. Jahani, A. Maleki, K. Miller, C. Smith, A. Pakman and R. Yue for fruitful conversations, and to M. Schnabel for sharing the code that produced the synthetic orientation maps in Figure 5. We are also grateful to the laboratory of Thomas M. Jessell in the Department of Biochemistry and Molecular Biophysics at Columbia University for providing the spinal cord data presented in Section 4.2.1. We also would like to thank the referees for carefully reading the original manuscript and raising important issues which improved the paper significantly.

REFERENCES

AFONSO, M. V., BIOUCAS-DIAS, J. M. and FIGUEIREDO, M. A. T. (2010). Fast image recovery using variable splitting and constrained optimization. *IEEE Trans. Image Process.* **19** 2345–2356. MR2798930

AHMADIAN, Y., PILLOW, J. W. and PANINSKI, L. (2011). Efficient Markov chain Monte Carlo methods for decoding neural spike trains. *Neural Comput.* **23** 46–96. MR2768274

AHRENS, M., ORGER, M., ROBSON, D., LI, J. and KELLER, P. (2013). Whole-brain functional imaging at cellular resolution using light-sheet microscopy. *Nat. Methods* **10** 413–420.

AKAY, T., TOURTELLOTTE, W. G., ARBER, S. and JESSELL, T. M. (2014). Degradation of mouse locomotor pattern in the absence of proprioceptive sensory feedback. *Proc. Natl. Acad. Sci. USA* **111** 16877–16882.

ANDREWS, D. F. and MALLOWS, C. L. (1974). Scale mixtures of normal distributions. *J. Roy. Statist. Soc. Ser. B* **36** 99–102. MR0359122

BARBERO, A. and SRA, S. (2011). Fast Newton-type methods for total variation regularization. In *Proceedings of the* 28*th International Conference on Machine Learning* (*ICML*-11) 313–320.

BARDSLEY, J. M., SOLONEN, A., HAARIO, H. and LAINE, M. (2014). Randomize-then-optimize: A method for sampling from posterior distributions in nonlinear inverse problems. *SIAM J. Sci. Comput.* **36** A1895–A1910. MR3248038

BESAG, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *J. Roy. Statist. Soc. Ser. B* **36** 192–236. MR0373208

BESAG, J. (1993). Towards Bayesian image analysis. *J. Appl. Stat.* **20** 107–119.

BESAG, J. and KOOPERBERG, C. (1995). On conditional and intrinsic autoregressions. *Biometrika* **82** 733–746. MR1380811

BOUCHARD, K. E., MESGARANI, N., JOHNSON, K. and CHANG, E. F. (2013). Functional organization of human sensorimotor cortex for speech articulation. *Nature* **495** 327–332.

BOUMAN, I. and LIU, B. (1988). A multiple resolution approach to regularization. In *Proceedings SPIE* 1001, *Visual Communications and Image Processing* '88 512–520.

BOUMAN, C. and SAUER, K. (1993). A generalized Gaussian image model for edge-preserving MAP estimation. *IEEE Trans. Image Process.* **2** 296–310.

BOYD, S., PARIKH, N., CHU, E., PELEATO, B. and ECKSTEIN, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Faund. Trends Mach. Learn.* **3** 1–122.

BRANDT, A. (1977). Multigrid Monte Carlo method. Conceptual foundations. *Math. Comp.* **31** 333–390.

BREZGER, A., FAHRMEIR, L. and HENNERFEIND, A. (2007). Adaptive Gaussian Markov random fields with applications in human brain mapping. *J. Roy. Statist. Soc. Ser. C* **56** 327–345. MR2370993

BUADES, A., COLL, B. and MOREL, J. M. (2005). A review of image denoising algorithms, with a new one. *Multiscale Model. Simul.* **4** 490–530.

BUESING, L., MACKE, J. H. and SAHANI, M. (2012). Learning stable, regularised latent models of neural population dynamics. *Network* **23** 24–47.

CANDES, E. J. (1999a). Harmonic analysis of neural networks. *Appl. Comput. Harmon. Anal.* **6** 197–218.

CANDES, E. J. (1999b). Curvelets—a surprisingly effective nonadaptive representation for objects with edges. In *Curve and Surface Fitting*: *Saint-Malo* (A. Cohen, C. Rabut and L. L. Schumaker, eds.) Vanderbilt Univ. Press, Nashville, TN.

CASELLA, G. (2001). Empirical Bayes Gibbs sampling. *Biostat.* **2** 485–500.

CASELLA, G., GHOSH, M., GILL, J. and KYUNG, M. (2010). Penalized regression, standard errors, and Bayesian lassos. *Bayesian Anal.* **5** 369–411.

CHAMBOLLE, A. (2004). An algorithm for total variation minimization and applications. *J. Math. Imaging Vision* **20** 89–97. MR2049783

CHARBONNIER, P., BLANC-FERAUD, L., AUBERT, G. and BARLAUD, M. (1997). Deterministic edge-preserving regularization in computed imaging. *IEEE Trans. Image Process.* **6** 298–311.

CRONIN, B., STEVENSON, I. H., SUR, M. and KORDING, P. (2010). Hierarchical Bayesian modeling and Markov Chain Monte Carlo sampling for tuning-curve analysis. *J. Neurophysiol.* **103** 591–602.

CUNNINGHAM, J., YU, B., SHENOY, K. and SAHANI, M. (2008). Inferring neural firing rates from spike trains using Gaussian processes. In *Advances in Neural Information Processing Systems* 20 (J. C. Platt, D. Koller, Y. Singer and S. T. Roweis, eds.). Curran Associates, Red Hook, NY.

CUNNINGHAM, J. P., GILJA, V., RYU, S. and SHENOY, K. (2009). Methods for estimating neural firing rates, and their application to brain-machine interface. *Neural Netw.* **22** 1235–1246.

CZANNER, G., EDEN, U., WIRTH, S., YANIKE, M., SUZUKI, W. and BROWN, E. (2008). Analysis of between-trial and within-trial neural spiking dynamics. *J. Neurophysiol.* **99** 2672–2693.

DABOV, K., FOI, A., KATKOVNIK, V. and EGIAZARIAN, K. (2007). Image denoising by sparse 3-D transform-domain collaborative filtering. *IEEE Trans. Image Process.* **16** 2080–2095. MR2460626

DAVIS, T. (2006). *Direct Methods for Sparse Linear Systems*. SIAM, Philadelphia, PA.

DAYAN, P. and ABBOTT, L. F. (2001). *Theoretical Neuroscience*. *Computational Neuroscience*. MIT Press, Cambridge, MA. MR1985615

DEFRISE, M., VANHOVE, C. and LIU, X. (2011). An algorithm for total variation regularization in high-dimensional linear problems. *Inverse Probl.* **27** 065002.

DOI, E., GAUTHIER, J. L., FIELD, G. D., SHLENS, J., SHER, A., GRESCHNER, M., MACHADO, T. A., JEPSON, L. H., MATHIESON, K., GUNNING, D. E., LITKE, A. M., PANINSKI, L., CHICHILNISKY, E. J. and SIMONCELLI, E. P. (2012). Efficient coding of spatial information in the primate retina. *J. Neurosci.* **32** 16256–16264.

DONOHO, D. and JOHNSTONE, I. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81** 425–455.

DUANE, S., KENNEDY, A. D., PENDELTON, B. J. and ROWETH, D. (1987). Hybrid Monte Carlo. *Phys. Lett. B* **55**.

ELTOFT, T., KIM, T. and LEE, T. (2006). On the multivariate Laplace distribution. *IEEE Signal Process. Lett.* **13** 300–303.

FAHRMEIR, L., KNEIB, T. and LANG, S. (2004). Penalized structured additive regression for space-time data: A Bayesian perspective. *Statist. Sinica* **14** 731–761. MR2087971

FEINBERG, E. H. and MEISTER, M. (2014). Orientation columns in the mouse superior colliculus. *Nature* **519** 229–232.

GAO, Y., BLACK, M., BIENENSTOCK, E., SHOHAM, S. and DONOGHUE, J. (2002). Probabilistic inference of arm motion from neural activity in motor cortex. In *Advances in Neural Information Processing Systems* 14 (Z. G. Thomas, G. Dietterich and S. Becker, eds.) 213–220.

GELMAN, A., CARLIN, J., STERN, H. and RUBIN, D. (2003). *Bayesian Data Analysis*. CRC Press, Boca Raton, FL.

GEMAN, S. and GEMAN, D. (1984). Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.* **6** 721–741.

GEMAN, D. and REYNOLDS, G. (1992). Constrained restoration and the recovery of discontinuities. *IEEE Trans. Pattern Anal. Mach. Intell.* **14** 367–383.

GEMAN, D. and YANG, C. (1995). Nonlinear image recovery with half-quadratic regularization. *IEEE Trans. Image Process.* **4** 932–946.

GEORGOPOULOS, A., KETTNER, R. and SCHWARTZ, A. (1986). Neuronal population coding of movement direction. *Science* **233** 1416–1419.

GILAVERT, C., MOUSSAOUI, S. and IDIER, J. (2015). Efficient Gaussian sampling for solving large-scale inverse problems using MCMC. *IEEE Trans. Signal Process.* **63** 70–80. MR3286325

GIRMAN, S. V., SAUVÉ, Y. and LUND, R. D. (1999). Receptive field properties of single neurons in rat primary visual cortex. *J. Neurophysiol.* **82** 301–311.

GIROLAMI, M., CALDERHEAD, B. and CHIN, S. (2011). Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *J. Roy. Statist. Soc. Ser. B* **73** 123–214. MR2814492

GOODMAN, J. and SOKAL, A. D. (1989). Multigrid Monte Carlo method. Conceptual foundations. *Phys. Rev. D* **40** 2035–2071.

GREEN, P. J. (1990). Bayesian reconstructions from emission tomography data using a modified EM algorithm. *IEEE Trans. Med. Imag.* **9** 84–94.

GROSENICK, L., KLINGENBERG, B., KATOVICH, K., KNUTSON, B. and TAYLOR, J. E. (2013). Interpretable whole-brain prediction analysis with GraphNet. *NeuroImage* **73** 304–321.

GROVES, A. R., CHAPPELL, M. A. and WOOLRICH, M. W. (2009). Combined spatial and non-spatial prior for inference on MRI time-scales. *NeuroImage* **45** 795–809.

GUILLAUME, F. and PENNY, W. D. (2007). Bayesian fMRI data analysis with sparse spatial basis function priors. *NeuroImage* **34** 1108–1125.

HAFTING, T., FYHN, M., MOLDEN, S., MOSER, M. B. and MOSER, E. I. (2005). Microstructure of a spatial map in the enthorhinal cortex. *Nature* **436** 801–806.

HALLAC, D., LESKOVEC, J. and BOYD, S. (2015). Network lasso: Clustering and optimization in large graphs. In *SIGKDD* 387–396.

HAMEL, E. J. O., GREWE, B. F., PARKER, J. G. and SCHNITZER, M. J. (2015). Cellular level brain imaging in behaving mammals: An engineering approach. *Neuron* **86** 140–159. DOI:10.1016/j.neuron.2015.03.055.

HARRISON, L. M. and GREEN, G. G. R. (2010). A Bayesian spatiotemporal model for very large data sets. *NeuroImage* **50** 1126–1141.

HARRISON, L. M., PENNY, W., ASHBURNER, J., TRUJILLO-BARRETO, N. and FRISTON, K. J. (2007). Diffusion-based spatial priors for imaging. *NeuroImage* **38** 677–695.

HARRISON, S. J., WOOLRICH, M. W., ROBINSON, E. C., GLASSER, M. F., BECKMAN, C. F., JENKINSON, M. and SMITH, S. M. (2015). Large-scale probabilistic functional modes from resting state fMRI. *NeuroImage* **109** 217–231.

HENRY, G. H., DREHER, B. and BISHOP, P. O. (1974). Orientation specificity of cells in cat striate cortex. *J. Neurophysiol.* **37** 1394–1409.

HOFFMAN, Y. (2009). Gaussian fields and constrained simulations of the large-scale structure. In *Data Analysis in Cosmology* 565–583. Springer, Berlin.

HOFFMAN, Y. and RIBAK, E. (1991). Constrained realization of Gaussian fields—a simple algorithm. *Astrophys. J.* **380** L5–L8.

HUBEL, D. H. and WIESEL, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol.* **160** 106–154.

HUBEL, D. H. and WIESEL, T. N. (1968). Receptive fields and functional architecture of the monkey striate cortex. *J. Neurophysiol.* **195** 215–243.

HUBER, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Stat.* **35** 73–101. MR0161415

ISSA, J. B., HAEFFELE, B. D., AGARWAL, A., BERGLES, D. E., YOUNG, E. D. and YUE, D. T. (2014). Multiscale optical Ca 2+ imaging of tonal organization in mouse auditory cortex. *Neuron* **83** 944–959.

KASCHUBE, M., SCHNABEL, M., LÖWEL, S., COPPOLA, D. M., WHITE, L. E. and WOLF, F. (2010). Universality in the evolution of orientation columns in the visual cortex. *Science* **330** 1113–1116.

KEIL, W., KASCHUBE, M., SCHNABEL, M., KISVARDAY, Z., LOWEL, S., COPPOLA, D. M., WHITE, L. E. and WOLF, F. (2012). Response to comment on "Universality in the evolution of orientation columns in the visual cortex." *Science* **336**.

KIM, S. J., KOH, S., BOYD, S. and GORINEVSKY, D. (2009). ℓ_1 trend filtering. *SIAM Rev.* **51** 339–360.

KROUCHEV, N., KALASKA, J. F. and DREW, T. (2006). Sequential activation of muscle synergies during locomotion in the intact cat as revealed by cluster analysis and direct decomposition. *J. Neurophysiol.* **96** 1991–2010.

LANG, S. and BREZGER, A. (2004). Bayesian P-splines. *J. Comput. Graph. Statist.* **13** 183–212. MR2044877

LANG, S., FRONK, E.-M. and FAHRMEIR, L. (2002). Function estimation with locally adaptive dynamic models. *Comput. Statist.* **17** 479–499. MR1952693

LASSAS, M. and SILTANEN, S. (2004). Can one use total variation prior for edge-preserving Bayesian inversion? *Inverse Probl.* **20** 1537.

LEBRUN, M., BUADES, A. and MOREL, J. M. (2013). A nonlocal Bayesian image denoising algorithm. *SIAM J. Imaging Sci.* **3** 1665–1688.

LEYTON, A. S. and SHERRINGTON, C. S. (1917). Observations on the excitable cortex of the chimpanzee, orangutan, and gorilla. *Q.j. Exp. Physiol.* **11** 135–222.

LOUCHET, C. and MOISAN, L. (2013). Posterior expectation of the total variation model: Properties and experiments. *SIAM J. Imaging Sci.* **6** 2640–2684. MR3143828

MACHADO, T. A., PNEVMATIKAKIS, E., PANINSKI, L., JESSELL, T. and MIRI, A. (2015). Primacy of flexor locomotor pattern revealed by ancestral reversion of motor neuron identity. *Cell* **162** 338–350.

MACKAY, D. (1995). Probable networks and plausible predictions—a review of practical Bayesian methods for supervised neural networks. *Network* **6** 469–505.

MACKE, J. H., GERWINN, S., WHITE, L. E., KASCHUBE, M. and BETHGE, M. (2010). Bayesian estimation of orientation preference maps. *Adv. Neural Inf. Process. Syst.* **22** 1195–1203.

MACKE, J. H., GERWINN, S., WHITE, L. E., KASCHUBE, M. and BETHGE, M. (2011). Gaussian process methods for estimating cortical maps. *NeuroImage* **56** 570–581.

METIN, C., GODEMENT, P. and IMBERT, M. (1988). The primary visual cortex in the mouses: Receptive field properties and functional organization. *Exp. Brain Res.* **69**.

MOTWANI, M. C., GADIYA, M. C., MOTWANI, R. C. and HARRIS, F. C. (2004). Survey of image denoising techniques. In *Global Signal Processing Expo*, Santa Clara, CA.

MURPHY, E. H. and BERMAN, N. (1979). The rabit and the cat: A comparison of some features of response properties of single cells in the primary visual cortex. *J. Comp. Neurol.* **188**.

MURRAY, I., ADAMS, R. P. and MACKAY, D. (2010). Elliptical slice sampling. In *The Proceedings of the* 13*th International Conference on Artificial Intelligence and Statistics* (*AISTATS*) **9** 541–548.

OHKI, K., CHUNG, S., CH'NG, Y., KARA, P. and REID, C. (2005). Functional imaging with cellular resolution reveals precise micro-architecture in visual cortex. *Nature* **433** 597–603.

OHKI, K., CHUNG, S., KARA, P., HUBENER, M., BONHOEFFER, T. and REID, R. C. (2006). Highly ordered arrangement of single neurons in orientation pinwheels. *Nature* **442** 925–928. DOI:10.1038/nature05019.

OLIVEIRA, J., BIOUCAS-DIAS, J. M. and FIGUEIREDO, M. (2009). Adaptive total variation image deblurring: A majorization-minimization approach. *Signal Process.* **89** 1683–1693.

PANINSKI, L. (2010). Fast Kalman filtering on quasilinear dendritic trees. *J. Comput. Neurosci.* **28** 211–228. MR2609429

PANINSKI, L., AHMADIAN, Y., FERREIRA, D. G., KOYAMA, S., RAHNAMA RAD, K., VIDNE, M., VOGELSTEIN, J. and WU, W. (2010). A new look at state-space models for neural data. *J. Comput. Neurosci.* **29** 107–126. MR2721336

PAPANDREOU, G. and YUILLE, A. (2010). Gaussian sampling by local perturbations. In *Advances in Neural Information Processing Systems* 23 (J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel and A. Culotta, eds.) 1858–1866. Curran Associates, Red Hook, NY.

PARK, T. and CASELLA, G. (2008). The Bayesian lasso. *J. Amer. Statist. Assoc.* **103** 681–686. MR2524001

PENFIELD, W. and RASMUSSEN, T. (1950). The cerebral cortex of man; a clinical study of localization of function. *J. Amer. Med. Assoc.* **144**.

PENNY, W. D., TRUJILLO-BARRETO, N. and FRISTON, K. J. (2005). Bayesian fMRI time series analysis with spatial priors. *NeuroImage* **24** 350–362.

PNEVMATIKAKIS, E. A., GAO, Y., SOUDRY, D., PFAU, D., LACEFIELD, C., POSKANZER, K., BRUNO, R., YUSTE, R. and PANINSKI, L. (2014a). A structured matrix factorization framework for large scale calcium imaging data analysis. Preprint. Available at arXiv:1409.2903.

PNEVMATIKAKIS, E. A., RAHNAMA RAD, K., HUGGINS, J. and PANINSKI, L. (2014b). Fast Kalman filtering and forward-backward smoothing via low-rank perturbative approach. *J. Comput. Graph. Statist.* **23** 316–339. MR3215813

PORTILLA, J., STRELA, V., WAINWRIGHT, M. J. and SIMONCELLI, E. P. (2003). Image denoising using scale mixtures of Gaussians in the wavelet domain. *IEEE Trans. Image Process.* **12** 1338–1351. MR2026777

PORTUGUES, R., FEIERSTEIN, C. E., ENGERT, F. and ORGER, M. B. (2014). Whole-brain activity maps reveal stereotyped, distributed networks for visuomotor behavior. *Neuron* **81** 1328–1343. DOI:10.1016/j.neuron.2014.01.019.

PRESS, W., TEUKOLSKY, S., VETTERLING, W. and FLANNERY, B. (1992). *Numerical Recipes in C*. Cambridge Univ. Press, Cambridge.

PREVEDEL, R., YOON, Y., HOFFMANN, M., PAK, N., WETZSTEIN, G., KATO, S., SCHRÖDEL, T., RASKAR, R., ZIMMER, M., BOYDEN, E. et al. (2014). Simultaneous whole-animal 3D imaging of neuronal activity using light-field microscopy. *Nat. Methods* **11** 727–730.

QUIROS, A., DIEZ, R. M. and GAMERMAN, D. (2010). Bayesian spatiotemporal model of fMRI data. *NeuroImage* **49** 442–456.

RAHNAMA RAD, K. and PANINSKI, L. (2010). Efficient estimation of two-dimensional firing rate surfaces via Gaussian process methods. *Network* **21** 142–168.

RASMUSSEN, C. and WILLIAMS, C. (2006). *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA.

REICHL, L., LÖWEL, S. and WOLF, F. (2009). Pinwheel stabilization by ocular dominance segregation. *Phys. Rev. Lett.* **102** 208101.

RIEKE, F., WARLAND, D., DE RUYTER VAN STEVENINCK, R. and BIALEK, W. (1997). *Spikes: Exploring the Neural Code*. MIT Press, Cambridge, MA.

ROBERT, C. P. and CASELLA, G. (2004). *Monte Carlo Statistical Methods*, 2nd ed. Springer, New York. MR2080278

ROBERTS, G. O. and STRAMER, O. (2002). Langevin diffusions and Metropolis-Hastings algorithms. *Methodol. Comput. Appl. Probab.* **4** 337–357. MR2002247

ROMANES, G. J. (1964). The motor pools of the spinal cord. *Prog. Brain Res.* **11** 93–119.

RUDIN, L. I., OSHER, S. and FATEMI, E. (1992). Nonlinear total variation based noise removal algorithms. *Phys. D* **60** 259–268. MR3363401

RUE, H. and HELD, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*. Taylor & Francis, London.

SCHNABEL, M., KASCHUBE, M., LOWEL, S. and WOLF, F. (2007). Random waves in the brain: Symmetries and defect generation in the visual cortex. *Eur. Phys. J. Spec. Top.* **145** 137–157.

SCOTT, S. H. (2000). Population vectors and motor cortex: Neural coding or epiphenomenon? *Nat. Neurosci.* **3** 307–308.

SHMUEL, A. and GRINVALD, A. (1996). Functional organization for direction of motion and its relationship to orientation maps in cat area 18. *J. Neurosci.* **16** 6945–6964.

SIDEN, P., EKLUND, A., BOLIN, D. and VILLANI, M. (2016). Fast Bayesian whole-brain fMRI analysis with spatial 3D priors. Preprint. Available at arXiv:1606.00980v1 [stat.CO].

SLAWSKI, M. (2012). The structured elastic net for quantile regression and support vector classification. *Stat. Comput.* **22** 153–168. MR2865062

SLAWSKI, M., ZU CASTELL, W. and TUTZ, G. (2010). Feature selection guided by structural information. *Ann. Appl. Stat.* **4** 1055–1080. MR2758433

STARCK, J., CANDES, E. and DONOHO, D. (2002). The curvevlet transform for image denoising. *IEEE Trans. Image Process.* **11** 670–684.

STEVENSON, R. and DELP, E. (1990a). Fitting curves with discontinuities. In *Proceedings of International Workshop on Robust Comput. Vision*, 127–136.

STEVENSON, R. and DELP, E. (1990b). Surface reconstruction with discontinuities. *Proc. SPIE Int. Soc. Opt. Eng.* **1610** 46–57.

SWINDALE, N. V. (1998). Orientation tuning curves: Empirical description and estimation of parameters. *Biol. Cybernet.* **78** 45–56.

SWINDALE, N. V. (2008). Visual map. *Scholarpedia* **3** 4607.

TIAO, Y. C. and BLAKEMORE, C. (1976). Functional organization in the visual cortex of the golden hamster. *J. Comp. Neurol.* **168** 459–481.

TIBSHIRANI, R. J. and TAYLOR, J. (2011). The solution path of the generalized lasso. *Ann. Statist.* **39** 1335–1371. MR2850205

VAN GERVEN, M. A. J., CSEKE, B., DE LANGE, F. P. and HESKES, T. (2010). Efficient Bayesian multivariate fMRI analysis using a sparsifying spatio-temporal prior. *NeuroImage* **50** 150–161.

VAN HOOSER, S. D., HEIMEL, J. A. F., CHUNG, S., NELSON, S. B. and TOTH, L. J. (2005). Orientation selectivity without orientation maps in visual cortex of a highly visual mammal. *J. Neurosci.* **25** 19–28.

VIDNE, M., AHMADIAN, Y., SHLENS, J., PILLOW, J. W., KULKARNI, J., LITKE, A. M., CHICHILNISKY, E. J., SIMONCELLI, E. and PANINSKI, L. (2012). Modeling the impact of common noise inputs on the network activity of retinal ganglion cells. *J. Comput. Neurosci.* **33** 97–121. MR2956393

VOGEL, C. R. and OMAN, M. E. (1996). Iterative methods for total variation denoising. *SIAM J. Sci. Comput.* **17** 227–238. MR1375276

VOGEL, C. R. and OMAN, M. E. (1998). Fast, robust total variation-based reconstruction of noisy, blurred images. *IEEE Trans. Image Process.* **7** 813–824. MR1667392

WAHBA, G. (1990). *Spline Models for Observational Data*. SIAM, Philadelphia, PA.

WANDELL, B. (1995). *Foundations of Vision*. Sinauer, Boston, MA.

WANG, Y., YANG, J., YIN, W. and ZHANG, Y. (2009). A new alternative minimization algorithm for total variation image reconstruction. *SIAM J. Imaging Sci.* **1** 248–272.

WANG, Y.-X., SHARPNACK, J., SMOLA, A. J. and TIBSHIRANI, R. J. (2016). Trend filtering on graphs. *J. Mach. Learn. Res.* **17** 1–41. MR3543511

WEST, M. (1987). On scale mixtures of normal distributions. *Biometrika* **74** 646–648. MR0909372

WILSON, S. and MOORE, C. (2015). S1 somatotopic maps. *Scholarpedia* **10** 8574.

WIPF, D. and NAGARAJAN, S. (2008). A new view of automatic relevance determination. In *Advances in Neural Information Processing Systems* 1625–1632. Curran Associates, Red Hook.

WOOLRICH, M. W. (2012). Bayesian inference in fMRI. *NeuroImage* **62** 801–810.

WOOLRICH, M. W., JENKINSON, M., BRADY, J. M. and SMITH, S. M. (2004). Fully Bayesian spatio-temporal modeling of fMRI data. *IEEE Trans. Med. Imag.* **23** 213–231.

YUAN, M. and LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 49–67. MR2212574

YUE, Y., LOH, J. M. and LINDQUIST, M. A. (2010). Adaptive spatial smoothing of fMRI images. *Stat. Interface* **3** 3–13. MR2609707

YUE, Y. and SPECKMAN, P. L. (2010). Nonstationary spatial Gaussian Markov random fields. *J. Comput. Graph. Statist.* **19** 96–116. MR2654402

YUE, Y. R., SPECKMAN, P. L. and SUN, D. (2012). Priors for Bayesian adaptive spline smoothing. *Ann. Inst. Statist. Math.* **64** 577–613. MR2880870

K. RAHNAMA RAD
DEPARTMENT OF INFORMATION SYSTEMS
AND STATISTICS BARUCH COLLEGE
CITY UNIVERSITY OF NEW YORK
NEW YORK, NEW YORK 10010
USA
E-MAIL: kamiar.rahnamarad@baruch.cuny.edu

T. A. MACHADO
COGNESCENT CORPORATION
1140 BROADWAY, SUITE 307
NEW YORK, NEW YORK 10001
USA
E-MAIL: tim.machado@gmail.com

L. PANINSKI
DEPARTMENT OF STATISTICS
AND CENTER FOR THEORETICAL NEUROSCIENCE
COLUMBIA UNIVERSITY
NEW YORK, NEW YORK 10027
USA
E-MAIL: liam@stat.columbia.edu