Stat

The ISI's Journal for the Rapid
Dissemination of Statistics Research

(wileyonlinelibrary.com) DOI: 10.1002/sta4.137

A Bayesian supervised dual-dimensionality reduction model for simultaneous decoding of LFP and spike train signals

Andrew Holbrook[a*], Alexander Vandenberg-Rodes[a], Norbert Fortin[b] and Babak Shahbaba[a]

Received 20 November 2016; Accepted 4 January 2017

Neuroscientists are increasingly collecting multimodal data during experiments and observational studies. Different data modalities—such as electroencephalogram, functional magnetic resonance imaging, local field potential (LFP) and spike trains—offer different views of the complex systems contributing to neural phenomena. Here, we focus on joint modelling of LFP and spike train data and present a novel Bayesian method for neural decoding to infer behavioural and experimental conditions. This model performs supervised dual-dimensionality reduction: it learns low-dimensional representations of two different sources of information that not only explain variation in the input data itself but also predict extraneuronal outcomes. Despite being one probabilistic unit, the model consists of multiple modules: exponential principal components analysis (PCA) and wavelet PCA are used for dimensionality reduction in the spike train and LFP modules, respectively; these modules simultaneously interface with a Bayesian binary regression module. We demonstrate how this model may be used for prediction, parametric inference and identification of influential predictors. In prediction, the hierarchical model outperforms other models trained on LFP alone, spike train alone and combined LFP and spike train data. We compare two methods for modelling the loading matrix and find them to perform similarly. Finally, model parameters and their posterior distributions yield scientific insights. Copyright © 2017 John Wiley & Sons, Ltd.

Keywords: Bayesian methods; discrimination; high-dimensional data; statistical learning

1 Introduction

Both in behavioural research and in clinical medicine, the relationship between neuronal signals and extraneuronal phenomena is often of interest. This is true whether a scientist wishes to test a hypothesis on memory function or whether a physician wants to detect a vulnerability towards stroke. Neural decoding is the science of deducing extraneuronal phenomena from the activity of neurons or groups of neurons (Brown et al., 1998). This activity can be registered in a number of modalities, including functional magnetic resonance imaging, electroencephalogram, local field potential (LFP) and spike trains. In general, neural decoding becomes more difficult as the number of neurons considered increases and as more distinct data modalities are integrated. On the other hand, the benefit of increasing neuron count—and from incorporating different signal types—is large. *A priori,* large groups of neurons have more

[a]Department of Statistics, University of California, Irvine, CA 92697, USA
[b]Department of Neurobiology and Behavior, University of California, Irvine, CA 92697, USA
*Email: aholbroo@uci.edu

capacity to encode complex behaviour than does a single neuron. Because signals captured by different data modalities are obtained by measuring different neurobiological quantities, incorporating multiple data modalities into a single statistical analysis offers an increased perspective that may help detect latent patterns otherwise missed.

Our contribution is a Bayesian hierarchical model that performs supervised dual-dimensionality reduction (sDDR) of both LFP and spike count data. Given a collection of binary outcomes y_i that may be plausibly linked to or reflected in neuronal activity, we assume that the probability $p(y_i = 1)$ is functionally linked to low-dimensional representations of both kinds of neuronal data:

$$g\{p(y_i = 1)\} = \beta + \beta_S^T z_i^S + \beta_L^T z_i^L . \tag{1}$$

Here, z_i^S and z_i^L are low-dimensional representations of the spike counts and LFP data, respectively. In the following, we explain the precise nature of these low-dimensional quantities and specify exactly how they may be obtained. Importantly, the model components are *not* found sequentially or by a multistage approach: the low-dimensional signals, z_i^S and z_i^L, are learned simultaneously with the coefficients of the predictive model (1). The model is thus able to relate low-dimensional representations of these input data to extraneuronal outcomes and to measure uncertainty while doing so. We demonstrate how this model may be used for neural decoding and the identification of influential predictors.

Our Bayesian hierarchical model captures variability in the input data in a way that is predictive of experimental conditions. We represent both the spike train and LFP data as being generated from exponential family distributions with low-dimensional covariance structure. This is carried out using probabilistic extensions of principal components analysis (PCA) (Pearson, 1901; Jolliffe, 2002), a non-probabilistic linear dimensionality reduction technique in which the eigenvalue decomposition of the empirical covariance matrix is considered. Generative versions include probabilistic PCA (PPCA) and factor analysis (Tipping & Bishop, 1999; Johnson & Wichern, 1992). Both of these methods model high-dimensional data as generated by a multivariate Gaussian distribution with covariance the sum of a low-rank matrix and a diagonal matrix (restricted to a multiple of the identity under PPCA). Exponential family PCA (ePCA) (Collins et al., 2001) is a generalized linear model extension of PCA, and we use this to model the non-negative integer constrained vectors of spike counts. A probabilistic wavelet PCA (wPCA) is used to model the LFP signals. See Bakshi (1998) and Feng et al. (2000) for non-probabilistic implementations.

Besides performing dimensionality reduction, the sDDR model also discriminates between extraneuronal conditions. It uses a variation of supervised PCA (Yu et al., 2006), which is an alternative to principal components regression and partial least squares (PLS) (Wold, 2006; Gustafsson, 2001). Applying logistic regression to a PCA derived, low-dimensional representation of data x for classifying outcome variables y is known as principal components (logistic) regression (PCR). PCR is rarely a recommended method, for the simple reason that the directions (eigenvectors) that best explain variability in x have little reason *a priori* to explain the variation in the outcomes y (Jolliffe, 1982). PLS has both probabilistic and non-probabilistic variations and also has been extended to PLS-discriminant analysis (PLS-DA), a deterministic method that handles discrete outcomes y (Barker & Rayens, 2003). PLS-DA favours the directions that maximize discrimination for y, but without taking into account the within-class covariances. In their probabilistic versions, PLS is easily extended to handle discrete data using the exponential family, and supervised PCA can be viewed as a special case of PLS (Murphy, 2012) (Section 3.3).

The sDDR model is applied to data from an experiment testing the capacity for non-spatial sequential memory in rats. During this experiment, the rats are presented with "correct" and "incorrect" sequences of odours and they must classify them accordingly. Meanwhile, spike trains are recorded from over 50 neurons (Figure 1), and LFP signals are recorded from 12 channels (Figure 2). Because of the LFP signals' non-stationary functional nature, the model performs wPCA on the LFP data. Because the spike counts are vectors of non-negative integers, we use ePCA with the log-link function on the spike train data. Latent variables from these two generalized PCA models feed into a logistic

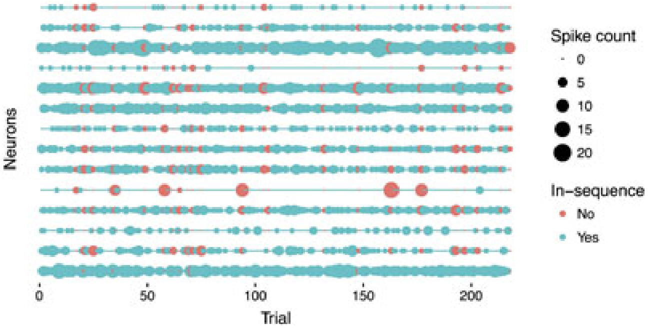

Figure 1. Trial-wise spike counts for 14 of the 52 neurons recorded over the course of the session. Counts range from 0 to 24. Individual neurons vary greatly in spike profile. In-sequence trials are colored blue, out-of-sequence trials are colored orange.

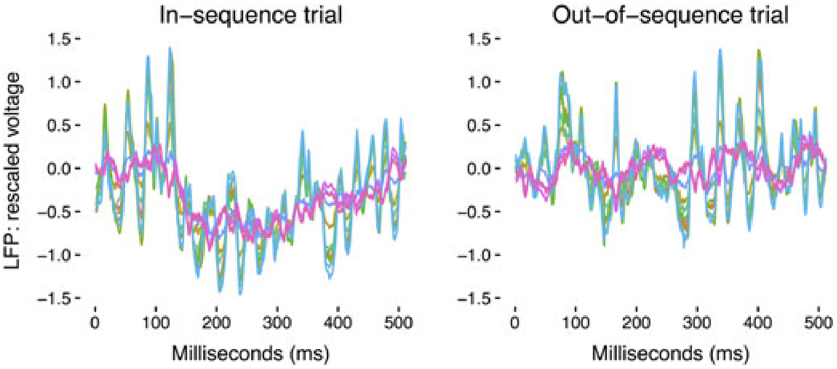

Figure 2. Two examples of LFP time series, each color pertains to one of twelve channels. The left and right plots come from trials 16 and 17, respectively. In trial 16, the correct odor was presented, and in trial 17 an incorrect odor was presented. In both plots, the time series seem to belong to two separate groups by channel: one of lower amplitude and one of higher amplitude. Both amplitude groups show interesting low frequency patterns. Indeed, the in-sequence signals are far from stationary. It is for these reasons a wavelet transform is used: it accounts for both changes in scale and shifts in time.

regression with sequential status (correct/incorrect) response. The model renders accurate predictions and interpretable parameters. If regression coefficients are larger for latent variables associated with one of the data modalities than they are for the other, there is evidence that that modality holds more information regarding the outcome (sequence status). We show that a certain functional on a few model parameters renders a useful metric for how predictive individual neurons and wavelet coefficients are of sequence status. This leads to easy variable selection and identification of influential predictors.

2 Scientific background and experimental setup

The hippocampus is central to the memory for sequences of events, but the basic neuronal mechanisms underlying sequential memory capacity are not well understood. There is significant evidence for the ability of hippocampal neurons to encode sequences of locations (Skaggs & McNaughton, 1996; Mehta et al., 2000; Dragoi & Buzsáki, 2006; Foster & Wilson, 2006; Gupta et al., 2012). Direct evidence, however, for a neural role in the memory of sequential relationships among non-spatial events remains absent. To address this issue, Allen et al. (2016) recorded neural activity in the CA1 region of the hippocampus as rats performed a sequential memory task. The task involved the presentation of repeated sequences of odours at a single port and tested the rats' ability to identify each odour as

(wileyonlinelibrary.com) DOI: 10.1002/sta4.137

"in sequence" or "out of sequence." LFP signals and multidimensional spike trains were recorded in the hippocampi of six rats, who had previously been trained on a particular "correct" sequence (A, B, C, D and E) of odours. Each trial involved the rat smelling one of the five odours through a port. The rat signalled whether the odour was in sequence or out of sequence by choosing to withdraw its nose from the port either after or before one second, respectively.

Roughly 88% of the trials were in sequence. This paper only considers data from a single session featuring rat "Super Chris." The session consisted of 218 trials lasting anywhere from 0.48 to 1.74 seconds each. For each trial, the data feature spike counts from 52 neurons, LFP signals from 12 channels and a binary indicator for whether the odour presented was in sequence (1) or out of sequence (0). In order to minimize differences in motor neuron activity across trials, the spike counts for each trial are the total number of spikes in the 0.5-second interval immediately preceding port withdrawal. We are interested in a supervised learning problem: *can we decode the rat's response from the neuronal data alone? If so, is there evidence for one data modality having more predictive capacity than the other? Are the two data modalities complementary with respect to outcome?* We assert that our Bayesian sDDR model can help answer these questions.

3 Bayesian linear dimensionality reduction and extensions

The most prevalent linear dimensionality reduction methods fall into the factor analysis framework. Such models specify the N observed continuous data points $x_1, \dots, x_N \in \mathbb{R}^d$ as

$$x_i = Fz_i + \mu + \epsilon_i, \tag{2}$$

where $z_i \in \mathbb{R}^k$ ($k < d$) are the latent factors, F is the $d \times k$ factor loading matrix and ϵ_i are i.i.d. $\mathrm{N}_d(0, \Psi)$, with Ψ a diagonal covariance matrix. Typically, the parameters F, μ and Ψ are optimized over, by using either expectation–maximization or closed-form expressions available when Ψ is restricted to be a multiple of the identity (Tipping & Bishop, 1999). If we place $N(0, I_k)$ priors on z_i, this latent factor is easily integrated out, leaving the sampling distribution written as

$$x_i \sim N(\mu, FF^T + \Psi). \tag{3}$$

This formulation assumes that the data lie close to an affine subspace spanned by the column vectors of F. There is, however, a continuum of subspaces that approximately spans the data. Picking a single subspace can dramatically understate variation in the data and lead to overfitting. One approach is to instead use the Bayesian framework to obtain a *posterior* distribution over model parameters F, μ and Ψ given x

$$\pi(F, \mu, \Psi|x) = \frac{f(x|F, \mu, \Psi)\, \pi(F, \mu, \Psi)}{\int f(x|F, \mu, \Psi)\, \pi(F, \mu, \Psi)\, dF\, d\mu\, d\Psi}, \tag{4}$$

where $f(x|F, \mu, \Psi)$ is the Gaussian density function of x given model parameters and $\pi(F, \mu, \Psi)$ is the prior distribution. Obtaining the posterior distribution means integrating over high-dimensional, high-density regions inhabited by F and helps avoid overfitting in a natural way. But even with much simpler models, the integral over model parameters is almost always intractable. The common solution is to sample from the posterior distribution using Markov chain Monte Carlo (MCMC) methods. Unfortunately, using MCMC for large matrices such as F is not straightforward, and the way F is modelled will delineate which sampling tools are available. Two different models for the loading matrix and their respective MCMC methods are discussed in Section 3.4.

3.1 Exponential family principal components analysis

A significant limitation of standard PCA appears when one tries to apply it to binary, count or categorical data. Taking a cue from generalized linear models, Collins et al. (2001) model each data point $x_{j,i}$ as being generated by an exponential family distribution:

$$f(x|\theta) = h(x)\exp\left(x\,\theta - b(\theta)\right). \tag{5}$$

Here, the natural parameter θ is related to the mean $\xi = \mathrm{E}(x\,|\,\theta)$ through the canonical link function: $\theta = g(\xi)$, where $g^{-1}(\theta) = b'(\theta)$. Dimension reduction is then applied to the natural parameter θ, which for many distributions of interest (e.g. Bernoulli or Poisson) can take any value on the real line, unlike mean ξ.

Because one kind of data we deal with is from spike trains, we focus on non-negative, integer-valued output and the log link. With $x_i \in (\{0\} \cup \mathbb{Z}^+)^d$, we have the canonical log-link function $g(\xi) = \log \xi$ and

$$x_{j,i} \sim \text{Poisson}(\xi_{j,i}), \quad \text{with} \tag{6}$$

$$\xi_{j,i} = g^{-1}\left(\sum_{\ell=1}^{k} F_{j,\ell} z_{i,\ell} + \mu_j\right), \tag{7}$$

where the $x_{j,i}$ are conditionally independent given the parameters (F, z, μ). In shorthand, this scheme may be written

$$x_i \sim \text{Pois}_{\otimes}\left(\exp\{U\Lambda\, z_i + \mu\}\right). \tag{8}$$

Here, $\otimes$ indicates the element-wise factorization of the multivariate conditional density function into univariate Poisson distributions. Parameters z and μ play the same role they do in (2) and can be given the same priors. Priors and constraints for F are discussed in Section 3.4. Note that the form of (4) in no way restricts f to being a Gaussian density: MCMC is performed in ePCA in the exact same way it is in run-of-the-mill Bayesian PCA.

3.2 Wavelet transform and wavelet principal components analysis

A wavelet (little wave) is a function approximately localized in both time and frequency. Thus, expressing a signal in terms of wavelets, instead of sinusoids as with the Fourier transform, can allow for more compact representations of non-stationary behaviour. A wavelet *basis*, generated by mother and father wavelets ψ and ϕ, has a particular dyadic structure describing the *multi-resolution analysis* (Daubechies, 1992). Given a signal $f(\cdot)$ on the unit interval, the wavelet transform returns the coefficients in the expansion

$$f(t) = s_0\phi(t) + \sum_{j=0}^{\infty}\sum_{k=0}^{2^j-1} d_{j,k}\psi_{j,k}(t), \tag{9}$$

where s_0 is the coarsest *smoothing* coefficient, $d_{j,k}$ is the *detail* coefficient at scale j and location k and $\psi_{j,k}(t) = 2^{j/2}\psi(2^j t - k)$ is the rescaled and shifted mother wavelet. Given 2^J equally spaced observations of f, the outer sum in (9) is truncated at scale J. In the classification setting, the resulting collection of 2^J wavelet coefficients can be treated as features, although dimension reduction is still necessary. In the Bayesian context, Vannucci et al. (2005) train a probit model that includes a spike and slab prior for variable selection on the wavelet coefficient, while Wang et al. (2007) instead use shrinkage priors. However, Crouse et al. (1998) noted significant correlation between neighbouring wavelet detail coefficients $d_{j,k}$ in real-world signals and proposed a hidden Markov tree model for the dependencies both across locations k and scales j. Classification is then based on the learned hidden Markov tree parameters. For dimension reduction, one can apply PCA to a *modified* set of wavelet features (Gupta & Jacobson, 2006). Because the wavelet transform is an orthogonal transformation, applying PCA to the raw wavelet coefficients would have the same effect as applying PCA to the data itself. In our setting, we observe 512 samples at 1000 Hz and apply the wavelet transform using the Daubechies extremal phase wavelet with 10 vanishing moments. As only the low frequencies are thought to be informative, up to about 50 Hz, we throw away the three finest scales leaving us with 64 coefficients.

(wileyonlinelibrary.com) DOI: 10.1002/sta4.137

3.3 Supervised exponential principal components analysis

Suppose now that *paired* data (x, y) are collected. Often the goal is to fit a joint model for the data, such that future x data can be used to predict future y—the supervised learning problem. Two classical linear dimensionality reduction methods for paired data are linear discriminant analysis (Fisher, 1936) and canonical correlation analysis (CCA) (Hotelling, 1936): the former method is applied when y is a discrete label, the latter when y takes on real multivariate values. Linear discriminant analysis is a directed method: a function on x is used to predict label y. CCA is undirected: x and y are modelled as exchangeable (ignoring differences in dimensionality). Their variability is explained by one latent variable each, one shared latent variable and additive noise (Murphy, 2012). PLS (Wold, 2006; Gustafsson, 2001) is halfway between CCA and supervised PCA, in that it is directed, but not fully. Ignoring additive noise, the PLS scheme features y fully explained by a latent variable that also underlies x. In this sense, it is directed. On the other hand, variability in x is explained by an additional latent variable that does not interface with y. In this way, it is undirected.

Supervised PCA (Bernardo et al., 2003; Yu et al., 2006), on the other hand, is fully directed: univariate y is regressed over the latents underlying x. Expanding on the notation of (2), we write

$$\begin{aligned} x_i &= F z_i + \mu + \epsilon_i^x , \\ y_i &= \beta z_i + \beta_0 + \epsilon_i^y . \end{aligned} \tag{10}$$

Because we are interested in predicting sequence status (in sequence/out of sequence), and because the spike counts are non-negative integers, we need a generalized linear model extension of supervised PCA. Consider the following supervised ePCA model:

$$x_i \sim p(x|\,\theta_j), \text{ an exponential family vector} \tag{11}$$

$$\theta_i = g_x^{-1}(Fz_i + \mu) \tag{12}$$

$$y_i \sim p(y|\,\eta_i), \text{ an exponential family random variable} \tag{13}$$

$$\eta_i = g_y^{-1}(\beta^T z_i + \beta_0). \tag{14}$$

We keep the prior specification for F, Z and μ in (12) as before—only the β coefficients are new. Later, this model will be applied in a number of variations: to count data coming from neural spike trains, where we specify x_j as Poisson with canonical link function $g_x(\xi) = \log \xi$; to wavelet coefficient data where x_j is normally distributed and given the identity link; and to both modalities combined by creating a hierarchical extension (Section 4). In all three cases, the binary behavioural response, y, is modelled using the logit link function $g_y(p) = \log \frac{p}{1-p}$. An example of this type of model is supervised logistic PCA, which was applied to genomic data by Yu et al. (2006). The parameters there were learned by maximum likelihood estimation.

3.4 Modelling the loading matrix

Historically, the loading matrix has been modelled in a number of different ways. Most of these models involve some sort of constraint on the loading matrix or a factorization of the matrix into two constrained matrices: $F = U\Lambda$. Here, U is a $d \times k$ matrix, the columns of which represent the low-dimensional directions of variation in x. Λ is a positive-diagonal $k \times k$ matrix parameterizing the scales of variation in those principal directions. For example, the maximum likelihood loading matrix from PPCA factorizes into U—the first k columns of the orthogonal matrix obtained by symmetrizing the empirical covariance matrix—times a scale matrix (Tipping & Bishop, 1999). In Collins et al. (2001), $\Lambda = I_k$ and mean μ are set to zero, but U is an unconstrained $d \times k$ matrix with all parameters learned via (penalized) maximum likelihood. Two competing methods for Bayesian PCA model the loading matrix differently: the first gives the columns of the loading matrix F independent spherical Gaussian priors (Bishop, 1999). Their treatment corresponds to making the elements of U i.i.d. standard normal and giving flat priors to the positively constrained elements of Λ. The second method models U as being an element of the Stiefel manifold (i.e. an orthonormal matrix) and gives it uniform prior with respect to the Haar measure (Holbrook et al., 2016; Hoff, 2007; Chan et al.). The decision to model U as a matrix with standard normal entries, or as an orthonormal matrix, does not change the general form of any of the models presented here. That said, how this decision affects model fitness is of particular interest. This question is addressed in Section 5.2.

3.4.1 The spherical Gaussian loading matrix.

Modelling the loading matrix as having spherical Gaussian columns was proposed in the original Bayesian PCA paper, Bishop (1999). If $F_1, \dots, F_k$ are the columns of F in equation (2), this prior may be written as

$$F_i \overset{\text{ind}}{\sim} N_k(0, \lambda_i^2 I), \quad i = 1, \dots, k. \tag{15}$$

If one specifies $k \times k$ matrix Λ to be diagonal with elements λ_j, $j = 1, \dots, k$, the model may also be written as

$$x_i = U\Lambda z_i + \mu + \epsilon_i\,, \quad U_{ij} \overset{\text{i.i.d.}}{\sim} N(0,1), \tag{16}$$

where all elements of U have standard normal distributions. One may give λ_i^2 a diffuse, positively constrained prior. (The author opts for flat priors in Bishop (1999).) Because (16) is unchanged when applying a permutation to the columns of U and the entries of Λz_i, we further specify $\lambda_1 < \cdots < \lambda_k$, to make the loading matrix $U\Lambda$ identifiable. Hybrid Monte Carlo (HMC) (Neal, 2011) is an effective computational inference method for this model.

3.4.2 The orthonormal loading matrix.

Under the assumption of dimension reduction ($d > k$), the singular value decomposition (SVD) $F = U\Lambda V^T$ can be modified so that U and V are respectively $d \times k$ and $k \times k$ matrices with orthonormal columns, while Λ is a $k \times k$ diagonal matrix with non-negative entries (the singular values) in decreasing order. Assuming the singular values are all distinct, F is uniquely specified by U, Λ and V. Now, recalling that $z \sim N(0, I_k)$ implies that $V^T z \sim N(0, I_k)$, one may ignore the superfluous rotation by V^T and reparameterize (2) into the exact same form as (16):

$$x_i = U\Lambda z_i + \mu + \epsilon_i\,, \quad U \sim \text{Uni}_{\mathcal{H}}(\mathcal{O}_{d,k})\,. \tag{17}$$

The collection of $d \times k$ matrices with orthogonal columns, denoted by $\mathcal{O}_{d,k}$, is known as the real *Stiefel manifold*, which is a (compact) Riemannian manifold of dimension $dk - \frac{1}{2}k(k+1)$. In this paper, a uniform prior distribution with respect to the Haar measure $\mathcal{H}$ of $\mathcal{O}_{d,k}$ is specified, but prior information can be incorporated with a matrix Bingham–von Mises–Fisher distributional prior if one so chooses (Hoff, 2009). Posterior inference on model (17) requires embedding geodesic Monte Carlo and extension of HMC that effectively constrains parameters to the relevant manifold (Byrne & Girolami, 2013).

Similar models were considered in Hoff (2007) and Chan et al. (), which model the SVD of the data, instead of the SVD of the factor matrix. In particular, Hoff (2007) assumes that $Z \in \mathcal{O}_{N,k}$, where $Z^T = [z_1, \dots z_N]$. The conditional distributions of each orthonormal column $P(U_j \,|\, U_{-j}, \Lambda, Z, Y)$ and $P(Z_j \,|\, Z_{-j}, \Lambda, U, Y)$ are shown to follow a von Mises–Fisher distribution, making the model amenable to Gibbs sampling. Relaxing this assumption to $Z_1, \dots, Z_N \sim N(0, I_k)$, as we do with (17), is not that different, at least *a priori*. In most situations, we have $N \gg k$, (even if the data dimension $d \approx N$) and the high-dimensional independent Gaussian random variables are orthogonal in prior expectation.

Both equations (16) and (17) are easily extended to their ePCA and supervised ePCA analogs by substituting $F = U\Lambda$. In the following section, the presentation of the hierarchical model is agnostic to the loading matrix model used.

(wileyonlinelibrary.com) DOI: 10.1002/sta4.137

4 The supervised dual-dimensionality reduction model

Let y_i be the in-sequence indicator: $y_i = 1$ if the odour presented for trial i is in sequence and $y_i = 0$ otherwise. Let x_i^S be the vector of spike counts and x_i^L be the LFP time series associated with trial i. For each trial, the joint model takes in a vector of spike counts and a vector of wavelet coefficients, $\tilde{x}_i^L$, derived from the LFP time series of that trial. The purpose of the wavelet transformation is to efficiently discretize the data (making it amenable to PCA) while maintaining local, temporal information that would not be preserved by the Fourier transform.

In turn, both LFP and spike train modules interface with the sequence classification module. The latent, low-dimensional signals, z_i^X and z_i^L, are featured as latent "data" in the Bayesian logistic regression and are learned at the same time as the logistic regression coefficients.

4.1 Local field potential module

The LFP data are modelled using wPCA on the vectors of wavelet coefficients, $\tilde{x}_i^L$. Following Section 3.4, the loading matrix for the wPCA model may be modelled either as standard Gaussian or as uniform orthonormal and is denoted as U^L. The wPCA model reads

$$\begin{aligned}
&\tilde{x}_i^L = U^L \Lambda^L z_i^L + \mu^L + \epsilon_i^L, \\
&\epsilon_i^L \sim N_d(0, \sigma_L^2 I), \qquad \mu^L \sim N_d(0, \tau_L^2 I), \\
&z_i^L \sim N_k(0, I), \qquad \sigma_L^2, \tau_L^2, \lambda_j^L \sim \text{Cauchy}^+(0, 5), \\
&\quad j = 1, \ldots, k, \quad \lambda_j^L > \lambda_{j'}^L, \quad j > j'.
\end{aligned} \tag{18}$$

Although truncated Cauchy priors are specified on the scale parameters, most diffuse priors will do. Ordering the scale parameters λ_j maintains identifiability of the loading matrix and helps the MCMC sampler by reducing multimodality of the density function. As stated earlier, the latent z_i^L is also featured in the sequential classification module of Section 4.3.

4.2 Spike train module

The high-dimensional spike trains are modelled using ePCA with the log link. Carrying over the notation from Section 3.1, $\text{Pois}_\otimes(\cdot)$ takes in vector arguments and indicates a vector of conditionally independent Poisson random variables with means the elements of the vector input. The ePCA spike train model is written thus

$$\begin{aligned}
x_i^S &\sim \text{Pois}_\otimes\big(\exp\{U^S \Lambda^S z_i^S + \mu^S\}\big), \\
\mu^S &\sim N_d(0, \tau_S^2 I), \qquad z_i^S \sim N_k(0, I) \\
\tau_S^2, \lambda_j^S &\sim \text{Cauchy}^+(0, 5), \\
j &= 1, \ldots, k, \quad \lambda_j^S > \lambda_{j'}^S, \quad j > j'.
\end{aligned} \tag{19}$$

The low-dimensional latent variables z_i^S are also featured in the sequential classification module as input "data" for the Bayesian logistic regression.

4.3 Sequential classification module

We model each sequential status, y_i, using Bayesian logistic regression: each y_i is distributed as a Bernoulli random variable with mean given by the inverse logit of the inputs from the LFP and spike train modules multiplied by regression coefficients:

$$\begin{aligned} y_i &\sim \text{Bernoulli}\big(\text{logit}^{-1}(\beta + \beta_S^T z_i^S + \beta_L^T z_i^L)\big) \, , \\ \beta &\sim N(0, 10^2), \quad \beta_S, \beta_L \sim N_k(0, 10^2\, I) \, . \end{aligned} \tag{20}$$

The logit function maps probabilities to their respective log-odds $p \mapsto \log p/(1-p)$; its inverse maps number from $\mathbb{R}$ to the interval $(0, 1)$. For this paper, it has been assumed that $k = k^S = k^L$, that is, that the spike trains and LFP coefficients are given the same latent dimensionality. Of course, the practitioner is in no way restricted to this assumption.

5 Results

5.1 Prediction

The sDDR model outperforms other common methods with respect to prediction. To show this, 10-fold cross-validation is used. We fit two versions of the sDDR model and a number of competing prediction methods to each training set and evaluate "0–1" loss (predicting sequential status) and log pointwise predictive densities on each of the 10 folds. One instance of the sDDR model uses an unconstrained Gaussian loading matrix; the other constrains the matrix U to the Stiefel manifold. Cross-validation is carried out three times for each of the three possible data inputs: LFP alone, spike trains alone and LFP and spikes combined.

The computed log pointwise predictive density ($\widehat{\text{lpd}}$) is an estimate of the log pointwise predictive density (lpd), which is itself a measure of the model's predictive fit to future data (Vehtari et al., 2016). If having observed training data, y, one uses MCMC to obtain a posterior sample $\theta_1, \ldots, \theta_S$, then given hold out data, $y_i, \ldots, y_n$, the lpd and $\widehat{\text{lpd}}$ are given by

$$\begin{aligned} \text{lpd} &= \sum_{i=1}^{n} \log p(y_i|y) = \sum_{i=1}^{n} \log \int p(y_i|\theta) p(\theta|y)\, d\theta \\ &\approx \sum_{i=1}^{n} \log \Big(\frac{1}{S} \sum_{s=1}^{S} p(y_i|\theta_s) \Big) = \widehat{\text{lpd}}. \end{aligned} \tag{21}$$

Results comparing both sDDR models with random forest (Breiman, 2001), PLS-DA (Barker & Rayens, 2003) and logistic regression with lasso (Tibshirani, 1996) are presented in Table I. The left half of Table I compares methods with respect to "0–1" loss, and the right half compares with respect to lpd. Within these halves, the columns show results for LFP input, spike train input and combined input, moving from left to right. The lpd results are presented for all methods except PLS-DA, for which lpd is undefined.

Table I. Predictive fit: 10-fold cross-validation results.

	0–1 Error			$\widehat{\text{lpd}}$		
Method	LFP	Spikes	Joint	LFP	Spikes	Joint
sDDR, Gaussian	0.110	0.064	0.060	−57.98	−33.92	−32.25
sDDR, Stiefel	0.106	0.069	0.064	−56.16	−34.48	−32.34
Logistic lasso	0.106	0.092	0.087	−70.81	−53.78	−49.13
Random forest	0.106	0.096	0.106	−66.56	−48.51	−55.13
PLS-DA	0.106	0.073	0.096	—	—	—

LFP, local field potential; sDDR, supervised dual-dimensionality reduction; PLS-DA, partial least squares-discriminant analysis.

(wileyonlinelibrary.com) DOI: 10.1002/sta4.137

None of the methods performed well when presented with LFP data alone. Most achieve roughly 90% accuracy by predicting in sequence 100% of the time. The Gaussian sDDR model actually performs worse here with respect to "0–1" loss: it predicts out of sequence once but for a trial that is actually in sequence! Despite this, the sDDR model performs best with respect to lpd, registering an $\widehat{\text{lpd}}$ of -57.98 for the Gaussian model and a -56.16 for the Stiefel model. Both the poor performance with respect to "0–1" loss and the strong performance with respect to lpd reinforce the idea that the sDDR model accomplishes more than the trivial fit to this rare event data. For the spike input models, the sDDR model performs better with respect to both criteria: it scores a 0.064 error rate for the Gaussian model and a 0.069 for the Stiefel model, followed by PLS-DA at 0.073; it also achieves the best $\widehat{\text{lpd}}$ at -33.92 and -34.48, followed by the random forest at -48.51. It is reassuring that the sDDR model seems to increase its lead over other methods when the input data have higher predictive capacity, as seems to be the case for the spike data. On the combined data, the sDDR and logistic lasso models continue to improve, unlike the random forest and PLS-DA models. The improvement in performance over the spike-only models is small, as one might expect from the weak performance of the LFP-only models. As with the spike-only models, the hierarchical model performs best: it obtains an error of 0.060 for the Gaussian model and 0.064 for the Stiefel model, compared with the logistic lasso's 0.087, and $\widehat{\text{lpd}}$s of -32.25 and -32.34 compared with the logistic lasso's -49.13.

Both sDDR variants outperform the competition, but their similar performances are worth commenting on. One of the chief motivations for using the Stiefel-constrained loading matrix is that it avoids the over-parameterization (discussed in Section 3.4.2 earlier) of the model caused by the rotational invariance of the standard normal latent variables. Another group of methods that accomplishes the same thing, although with a very different approach, falls under the category of independent component analysis (ICA). ICA models avoid this over-parameterization by modelling the latents as non-Gaussian and therefore no longer rotationally invariant (Murphy, 2012). Indeed, by also including the latents in a logistic regression, the sDDR can be thought of as inadvertently affecting an ICA-like model. To see this, consider the logistic regression module as a prior on the latents that makes them non-Gaussian *a priori*, as well as making them predictive of the outcome y. This insight could explain the similar performance of the two loading matrix models with respect to predictive fit: thanks to the logistic regression module, they are approximately the same model.

5.2 Variable selection and scientific inference

5.2.1 Identifying influential predictors.

In addition to prediction, the hierarchical model can assist in variable selection and meaningful inference. This is carried out through direct interpretation of the model parameters. Recall that both the LFP and spike modules explain the variability of the input variable, x_i, by some low-dimensional variable, z_i. In the case of the spike counts, the relationship is mediated by the log-link function, but it is still useful to consider the log-intensity $U \Lambda z_i \approx \log x_i$. Consider then the covariance of this log-intensity or of the wavelet coefficients, $\tilde{x}_i$, with the log-odds of an odour being in sequence. This covariance is directly assessable for each neuron and each wavelet coefficient in the model. In the case of the wavelet coefficients, this may be written as

$$\text{Cov}\big(\tilde{x}_i,\, \log \frac{P(y_i = 1)}{P(y_i = 0)}\big) = \text{Cov}(U \Lambda z_i,\, \beta^T z_i) = U \Lambda \, \text{Cov}(z_i, z_i)\, \beta = U \Lambda \beta. \qquad (22)$$

Thus, $U \Lambda \beta$ is a vector whose jth element is a measure of the association of each wavelet coefficient (or neuron) with the outcome. MCMC gives the posterior distribution over this vector. Figures 3 and 4 present 95% highest posterior density (HPD) intervals for the elements of this association vector for all neurons and wavelet coefficients. For both figures, orange intervals imply pointwise statistical significance and blue intervals imply a lack thereof. In Figure 3, only a few neurons are shown to be influential. In particular, neurons one and three are associated with an odour being in sequence, while neuron seven is strongly associated with an odour being out of sequence.

(wileyonlinelibrary.com) DOI: 10.1002/sta4.137

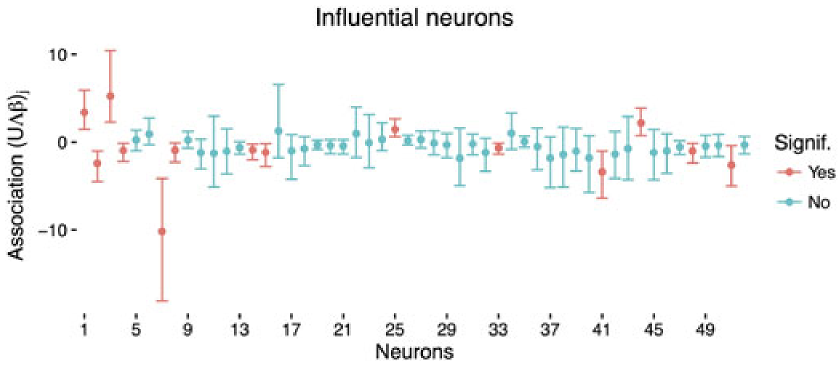

Figure 3. 95% credible intervals for the elements of association vector corresponding to each of the 52 neurons modeled: orange are point-wise statistically significant; blue are not.

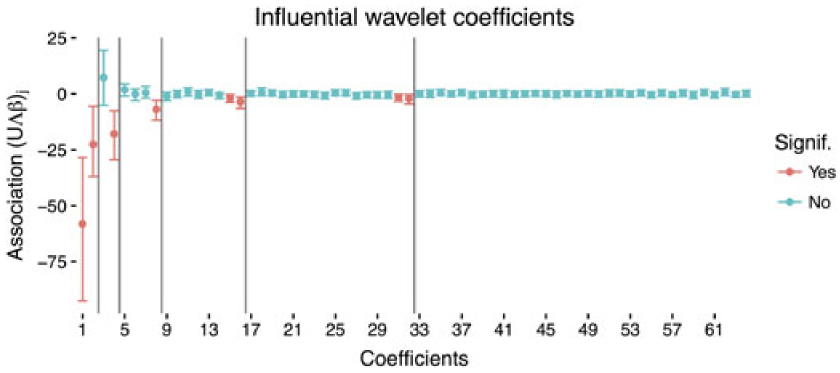

Figure 4. 95% credible intervals for the elements of association vector corresponding to each of the 64 wavelet coefficients modeled: orange are point-wise statistically significant; blue are not. Vertical lines divide coefficients into scale groups moving from coarse to fine from left to right. Within scales, coefficients move from early time series to late from left to right.

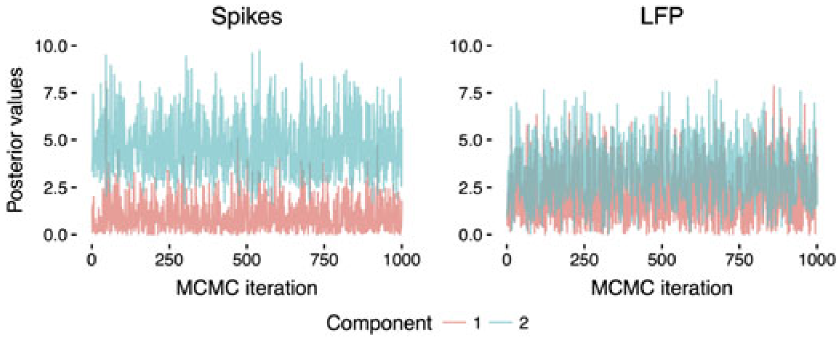

Figure 5. The posterior trace plots of the absolute value of logistic regression parameters associated with the two dimensional latent representations of the spike and LFP data. Non-zero distributions suggest significant association between low-dimensional representations and outcome.

(wileyonlinelibrary.com) DOI: 10.1002/sta4.137

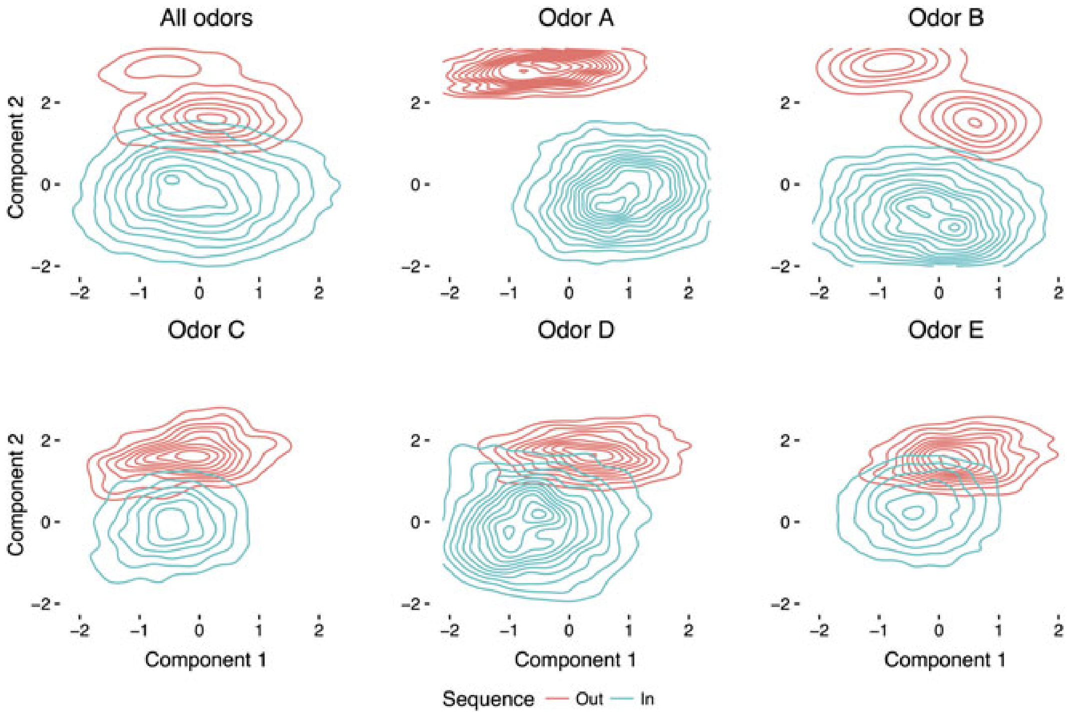

Figure 6. Posterior distribution of low dimensional representations of the spike data: the horizontal axis is the first principal component, the vertical is the second. Blue signifies in-sequence, and orange signifies out-of-sequence trials. Figure divided into plots stratified by odor presented.

Figure 4 shows a clear pattern in terms of the scales and translations associated with the various wavelet coefficients. The plot is divided into scale blocks by vertical lines. The leftmost block features the summation and difference coefficients associated with the half-scale, that is, that pertaining to the time series being divided into halves. The next block is on the quartic scale and so on. In other words, the scale becomes finer and finer moving from block to block towards the right. Note that the low frequencies seem to have the largest association with the outcome: the leftmost block is entirely significant; the next is half significant; the third is a quarter. Apparently, the final block is too fine to be predictive of outcome. Within blocks, another pattern emerges. The rightmost elements of each block (save the last block) are predictive of outcome. These elements pertain to the rightmost translations *in time*. In other words, these are the wavelet coefficients that correspond to the tail of the time series within their respective scales. This suggests that the neural activity occurring right before the rat removes its nose from the port is most informative about the rat's response.

5.2.2 Logistic regression coefficients. The posterior distributions of the regression parameters β and γ reflect the prediction results. Figure 5 presents the absolute values of 1000 draws from the posterior distributions of the logistic coefficients associated with both latent factors from both the LFP and the spike module. For both LFP and spike modules, the second principal component is the significant predictor of sequential status, but the magnitude of the second coefficient corresponding to the spike module is larger than that of that corresponding to the LFP data. This result agrees with the idea that the spike data contain more predictive content than does the LFP data.

The regression coefficients relate low-dimensional patterns in spike counts and LFPs to whether the model believes a sequence is correctly ordered. Their distributions support the hypothesis that the rat hippocampus is a place where sequential learning is performed. Here, *learning* is meant to suggest a global phenomenon, one involving relationships between individual neurons and groups thereof. Figure 5 affirms the hypothesis in a specific sense: if a coefficient has a significantly non-zero posterior distribution, then intensity of the relevant latent variable corresponds to the increased or decreased odds of sequential correctness.

5.2.3 Low-dimensional visualization. The posterior distributions of the low-dimensional representations z_i give insight into the predictive content available in the data modalities. Figure 6 shows contour plots derived from posterior samples of the low-dimensional representations z_i^S of the spike data, coloured by sequential status and separated by odour presented. Combining all odours together, one can see that there are distinct in-sequence and out-of-sequence clusters and that the out-of-sequence cluster has two distinct sub-clusters: one overlaps with the in-sequence cluster, and one is completely separate from the in-sequence cluster. The root of this bimodality can be seen in the odour-specific plots. It is clear that the ability of the rat to detect sequential status varies by the odour presented. Odour A has the clearest distinction in sequential status, odour B has the next biggest distinction, and the other three odours all have poor separation. This result suggests that it is easier for the rat to remember the correct order early on in the sequence. It seems reasonable that the rat should be able to remember the first two odours with which the sequence starts. It is also reasonable that the ability to detect out-of-sequence odours should degrade with progress into the sequence. As one might guess from the poor predictive performance of the LFP-only model, the low-dimensional representations of the LFP data show no such separation.

6 Discussion

We have developed a Bayesian framework for sDDR and used it to perform neural decoding where multiple data modalities were available. In a one-step process, low-dimensional representations of both LFP and spike train data were found and fed into a logistic regression model. The hierarchical model achieved better predictive fit than multiple competitors, as measured by "0–1" loss and lpd. Results were obtained for two different versions of the sDDR, one with a Gaussian loading matrix model, the other with a Stiefel-constrained loading matrix model.

Bayesian inference allowed for the flexible incorporation of the two data modalities in a way that automatically accounted for their predictive capacities. In terms of prediction, the sDDR model was at its best with the combined data. The spike train-only model outperformed the LFP-only model with respect to both predictive measures. The posterior distributions of model parameters such as logistic regression coefficients and low-dimensional representations also reinforced the idea that the spike counts had greater predictive content than the LFP data. That said, the model's ability to exploit the LFP data was completely dependent on the form in which the LFP data were presented to the model, that is, as a vector of wavelet coefficients taken from averaging multiple channels. Transforming time series data into wavelet coefficients involves the throwing away of information (as does averaging over channels), so it is plausible that incorporating the LFP data in another form might provide superior results. Whereas both LFP and spike modules featured extensions of Bayesian PCA, it might be prudent to model the LFP data using functional logistic regression. Preliminary results using Gaussian process regression coefficients have been promising. Whether to perform functional logistic regression on the time series data or on the periodogram is an open question, and its answer may be application dependent. A supervised functional PCA approach, obtained by combining PCA and functional logistic regression, might prove ideal. Possibilities aside, the sDDR model provides a firm basis for future multimodal neural decoding efforts.

Acknowledgements

This work was partially supported by NIH grant R01-AI107034 and NSF grant DMS-1622490. The authors would like to thank Hernando Ombao for helpful discussion.

(wileyonlinelibrary.com) DOI: 10.1002/sta4.137

References

Allen, TA, Salz, DM, McKenzie, S & Fortin, NJ (2016), 'Nonspatial sequence coding in CA1 neurons', *The Journal of Neuroscience*, **36**(5), 1547–1563.

Bakshi, BR (1998), 'Multiscale PCA with application to multivariate statistical process monitoring', *AIChE journal*, **44**(7), 1596–1610.

Barker, M & Rayens, W (2003), 'Partial least squares for discrimination', *Journal of Chemometrics*, **17**(3), 166–173.

Bernardo, JM, Bayarri, MJ, Berger, JO, Dawid, AP, Heckerman, D, Smith, AFM & West, M (2003), 'Bayesian factor regression models in the 'large p, small n' paradigm', *Bayesian Statistics*, **7**, 733–742.

Bishop, CM (1999), 'Bayesian pca', *Advances in Neural Information Processing Systems*, **11**, 382–388.

Breiman, L (2001), 'Random forests', *Machine Learning*, **45**(1), 5–32.

Brown, EN, Frank, LM, Tang, D, Quirk, MC & Wilson, MA (1998), 'A statistical paradigm for neural spike train decoding applied to position prediction from ensemble firing patterns of rat hippocampal place cells', *The Journal of Neuroscience*, **18**(18), 7411–7425.

Byrne, S & Girolami, M (2013), 'Geodesic Monte Carlo on embedded manifolds', *Scandinavian Journal of Statistics*, **40**(4), 825–845.

Chan, Joshua C. C. and Leon-Gonzales, Roberto and Strachan & Rodney, W., *Invariant Inference and Efficient Computation in the Static Factor Model. CAMA Working Paper 32/2013.* Available at SSRN: https://ssrn.com/abstract=2275707orhttp://dx.doi.org/10.2139/ssrn.2275707.

Collins, M, Dasgupta, S & Schapire, RE (2001), *A generalization of principal components analysis to the exponential family*, Vancouver, British Columbia, Canada.

Crouse, MS, Nowak, RD & Baraniuk, RG (1998), 'Wavelet-based statistical signal processing using hidden Markov models', *IEEE Transactions on Signal Processing*, **46**(4), 886–902.

Daubechies, I (1992), *Ten Lectures on Wavelets*, Vol. 61, *Society for Industrial and Applied Mathematics Philadelphia*, PA, USA. ISBN:0-89871-274-2.

Dragoi, G & Buzsáki, G (2006), 'Temporal encoding of place sequences by hippocampal cell assemblies', *Neuron*, **50**(1), 145–157.

Feng, G-C, Yuen, PC & Dai, D-Q (2000), 'Human face recognition using PCA on wavelet subband', *Journal of Electronic Imaging*, **9**(2), 226–233.

Fisher, RA (1936), 'The use of multiple measurements in taxonomic problems', *Annals of Eugenics*, **7**(2), 179–188.

Foster, DJ & Wilson, MA (2006), 'Reverse replay of behavioural sequences in hippocampal place cells during the awake state', *Nature*, **440**(7084), 680–683.

Gupta, AS, van der Meer, MAA, Touretzky, DS & Redish, AD (2012), 'Segmentation of spatial experience by hippocampal theta sequences', *Nature Neuroscience*, **15**(7), 1032–1039.

Gupta, MR & Jacobson, NP (2006), *Wavelet principal component analysis and its application to hyperspectral images*, IEEE, Atlanta, Georgia.

Gustafsson, MG (2001), 'A probabilistic derivation of the partial least-squares algorithm', *Journal of Chemical Information and Computer Sciences*, **41**(2), 288–294.

Hoff, PD (2007), 'Model averaging and dimension selection for the singular value decomposition', *Journal of the American Statistical Association*, **102**(478), 674–685.

Hoff, PD (2009), 'Simulation of the matrix Bingham–von Mises–Fisher distribution, with applications to multivariate and relational data', *Journal of Computational and Graphical Statistics*, **18**(2), 438–456.

Holbrook, A, Vandenberg-Rodes, A & Shahbaba, B (2016), *Bayesian inference on matrix manifolds for linear dimensionality reduction*. arXiv preprint arXiv:1606.04478.

Hotelling, H (1936), 'Relations between two sets of variates', *Biometrika*, **28**(3/4), 321–377.

Johnson, RA & Wichern, DW (1992), *Applied Multivariate Statistical Analysis*, Vol. 4, *Prentice hall Englewood Cliffs, NJ*.

Jolliffe, I (2002), *Principal Component Analysis*. Wiley Online Library.

Jolliffe, IT (1982), 'A note on the use of principal components in regression', *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **31**(3), 300–303.

Mehta, MR, Quirk, MC & Wilson, MA (2000), 'Experience-dependent asymmetric shape of hippocampal receptive fields', *Neuron*, **25**(3), 707–715.

Murphy, KP (2012), *Machine Learning: A Probabilistic Perspective*, *MIT press*, Cambridge, Mass and London, England.

Neal, RM (2011), 'MCMC using Hamiltonian dynamics', *Handbook of Markov Chain Monte Carlo*, **2**, 113–162.

Pearson, K (1901), 'LIII. On lines and planes of closest fit to systems of points in space', *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, **2**(11), 559–572.

Skaggs, WE & McNaughton, BL (1996), 'Replay of neuronal firing sequences in rat hippocampus during sleep following spatial experience', *Science*, **271**(5257), 1870–1873.

Tibshirani, R (1996), 'Regression shrinkage and selection via the lasso', *Journal of the Royal Statistical Society. Series B (Methodological)*, **58**(1), 267–288.

Tipping, ME & Bishop, CM (1999), 'Probabilistic principal component analysis', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **61**(3), 611–622.

Vannucci, M, Sha, N & Brown, PJ (2005), 'NIR and mass spectra classification: Bayesian methods for wavelet-based feature selection', *Chemometrics and Intelligent Laboratory Systems*, **77**(1), 139–148.

Vehtari, A, Gelman, A & Gabry, J (2016), 'Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC', *Statistics and Computing*, 1–20.

Wang, X, Ray, S & Mallick, BK (2007), 'Bayesian curve classification using wavelets', *Journal of the American Statistical Association*, **102**(479), 962–973.

Wold, H. (2006), 'Partial Least Squares', *Encyclopedia of Statistical Sciences*, **9**.

Yu, S, Yu, K, Tresp, V, Kriegel, H-P & Wu, M (2006), *Supervised probabilistic principal component analysis*, *ACM*, Philadelphia, Pittsburgh.