Journal of Neural Engineering

PAPER

Scalable surrogate deconvolution for identification of partially-observable systems and brain modeling

To cite this article: Matthew F Singh *et al* 2020 *J. Neural Eng.* **17** 046025

Recent citations

- Estimation and validation of individualized dynamic brain models with resting state fMRI
  Matthew F. Singh *et al*

IOP Publishing J. Neural Eng. 17 (2020) 046025 https://doi.org/10.1088/1741-2552/aba07d

Journal of Neural Engineering

Scalable surrogate deconvolution for identification of partially-observable systems and brain modeling

RECEIVED
31 January 2020

REVISED
31 May 2020

ACCEPTED FOR PUBLICATION
26 June 2020

PUBLISHED
4 August 2020

Matthew F Singh[1,2,4], **Anxu Wang**[1,2], **Todd S Braver**[2] **and ShiNung Ching**[1,3]

[1] Department of Electrical and Systems Engineering, Washington University in St. Louis, St. Louis, MO, United States of America
[2] Department of Psychological and Brain Sciences, Washington University in St. Louis, St. Louis, MO, United States of America
[3] Department of Biomedical Engineering, Washington University in St. Louis, St. Louis, MO, United States of America
[4] Author to whom any correspondence should be addressed.

E-mail: f.singh@wustl.edu

Keywords: model fitting, deconvolution, fMRI, dual estimation

Abstract

Objective. For many biophysical systems, direct measurement of all state-variables, $in - vivo$ is not feasible. Thus, a key challenge in biological modeling and signal processing is to reconstruct the activity and structure of interesting biological systems from indirect measurements. These measurements are often generated by approximately linear time-invariant dynamical interactions with the hidden system and may therefore be described as a convolution of hidden state-variables with an unknown kernel. *Approach.* In the current work, we present an approach termed surrogate deconvolution, to directly identify such coupled systems (i.e. parameterize models). Surrogate deconvolution reframes certain non linear partially-observable identification problems, which are common in neuroscience/biology, as analytical objectives that are compatible with almost any user-chosen optimization procedure. *Main results.* We show that the proposed technique is highly scalable, low in computational complexity, and performs competitively with the current gold-standard in partially-observable system estimation: the joint Kalman Filters (Unscented and Extended). We show the benefits of surrogate deconvolution for model identification when applied to simulations of the Local Field Potential and blood oxygen level dependent (BOLD) signal. Lastly, we demonstrate the empirical stability of Hemodynamic Response Function (HRF) kernel estimates for Mesoscale Individualized NeuroDynamic (MINDy) models of individual human brains. The recovered HRF parameters demonstrate reliable individual variation as well as a stereotyped spatial distribution, on average. *Significance.* These results demonstrate that surrogate deconvolution promises to enhance brain-modeling approaches by simultaneously and rapidly fitting large-scale models of brain networks and the physiological processes which generate neuroscientific measurements (e.g. hemodynamics for BOLD fMRI).

1. Introduction

A key challenge in neural engineering pertains to estimating neural model parameters from indirect observations that are temporally convolved from source measurements. For example, many imaging modalities reflect convolution of neural activity with temporal kernels associated with slower physiological processes such as blood flow (figure 1(A)), or molecular concentrations (table 1). Often, these kernels are not known, necessitating so-called 'dual estimation' of both the latent neural activity and the neural model (including convolutional kernels) at the same time. Our paper presents a computational framework for addressing this problem. Specifically, we assume that the system can be described in the following form (or its discrete-time equivalent):

$$\dot{x} = f(\theta, x, \hat{z}) + \varepsilon(t) \qquad (1)$$

$$\hat{z}_i(t) = [h_i(\eta_i) * x_i](t) \qquad (2)$$

$$z_i(t) = \hat{z}_i + \nu_i(t) = [h_i(\eta_i) * x_i](t) + \nu_i(t) \qquad (3)$$

Here, $x \in \mathbb{R}^n$ are the hidden non-convolutional state variables and $\hat{z}$ are the physiological variables generated by convolution. We denote unknown parameters for the non-convolutional plant as $\theta \in \mathbb{R}^q$ and for the convolutional plant as $\eta_i \in \mathbb{R}^{r_i}$. Each parameterized kernel (h_i) represents the process generating the corresponding measurable variable z_i via convolution (denoted $*$). This formulation requires the assumption that these processes may be well-approximated by a finite-impulse response function and that structural priors may be placed on each kernel (i.e. h_i is known up to a small number of parameters: η_i; see section 5.2 for discussion). We denote process noise in the hidden state variables by $\varepsilon(t) \in \mathbb{R}^n$) and denote measurement noise $\nu_i(t)$, both of which we assume to be drawn from stationary distributions, independently realized at each time step (noise is not auto-correlated). In the current context, x represents latent neural state-variables. The measurements z_i are multi-dimensional recordings of neural data and $\hat{z}$ are the corresponding physiological sources. These sources can either feed back into the latent system (e.g. Ca^{2+} concentration) or be modeled as purely downstream (as is typical for BOLD). Formally, we seek to estimate the convolutional kernel parameters $\{\eta\}$ and the neural model parameters θ using the measurements z (i.e. the 'dual' estimation). This problem formulation is highly relevant to neuroscience and neural engineering since it would enable inferences regarding brain activity via indirect and uncertain dynamical transformations.

1.1. Relevance to neuroscience and neural engineering

Whereas many neural models emphasize membrane potentials, channel conductances, and/or firing rate as state variables, high-coverage measurements often consist of the extracellular ('local') field potential, concentration of signaling molecules (e.g. Ca^{2+}), blood-oxygenation (and the derived BOLD-fMRI contrast) or radionuclide concentrations (e.g. PET). In all of these cases, the measurements reflect downstream, temporally extended consequences of neuronal firing (table 1). Thus, in the context of neuroscience, the dual-problem consists of simultaneously estimating the parameters of neural systems, while inverting measured signals into their unmeasured neural generators (the state-variables specified by a given model framework). Often this linkage (from generator to measurement) is assumed to be a linear time-invariant (LTI) system so that the relationship can be described via convolution with parameterized kernels. For, example, post-synaptic currents are often modeled via synaptic 'kernels' (e.g. 'alpha-synapse', [1]), kernels for molecular concentrations (e.g. Ca^{2+}, [2, 3]) are derived from Markovian kinetic-schemes [2, 9], and the neurovascular coupling kernel (linking BOLD-fMRI and neural 'activity') is described by a Hemodynamic-Response Function (HRF, [4]; figures 1(A) and (B)). If these functions are assumed fixed, it may be possible to estimate the neural state-variables via deconvolution, in which case, conventional modeling approaches are feasible. However, in many cases only the general form of the kernel function is known (e.g. up to a small number of unknown parameters). This underspecification results in computationally difficult dual estimation problems (estimating the neural states and the model parameters). The current work aims to treat such dual problems in a computationally efficient, and highly scalable manner.

1.2. Previous work

Currently, there are several methods to deal with dual-identification for small systems and these approaches may be grouped into black-box and grey box models. However, whereas black-box modeling encompasses diverse approaches such as neural networks [10], Volterra Expansion [11], and Nonlinear Autoregressive Moving Average (NARMA) models [12]; grey-box identification (model parameterization) has largely centered upon the dual/joint Kalman-Filters (linear, extended, unscented etc [13–16]) and related Bayesian methods. Under these approaches, the convolutional component is converted into the equivalent linear time-invariant (LTI) system format and free-parameters are modeled as additional state-variables. Thus, joint state-space techniques re-frame the dual-estimation problem as conventional state-estimation with a fully-determined model.

However, none of these methods are well situated to perform dual-identification for large (grey-box) systems due to the high computational complexity and data-intensive nature of Bayesian dual-estimation. These features are particularly limiting to neuroscience applications which typically feature a large number of connectivity parameters and potentially few sampling times (e.g. fMRI). These approaches also increase in complexity with the number of additional state variables necessary to represent complex kernels such as the hemodynamic response function (figure 1(B)).

Neural systems present two challenges to the current status quo: the dimensionality of the neural system/parameters and the complexity of the convolutional kernel. Neurobiological recordings are often high-dimensional, containing dozens or hundreds of neurons/neural populations. Moreover, the number of free parameters often scales nonlinearly with the number of populations (e.g. quadratically for the number of connectivity parameters). Current dual-estimation approaches such as joint-Kalman-Filtering are computationally limited in these settings due to their high computational complexity in terms of both the number of state variables and the number of parameters estimated.

Previous approaches are also limited in terms of kernel complexity. Since joint-Kalman-Filtering employs a state-space representation, convolutional

Table 1. Common neuroimaging measures subject to convolution.

Modality	Physiology	Popular Kernels	Interpretation
fMRI	BOLD Signal	Di-Gamma [4]	Hemodynamic Response
PET	Radionucleotide Concentration	Multi-Exponential [5, 6]	Exchange, Radio decay
Ca^{2+} Imaging	Ca^{2+} Concentration	Multi-Exponential [2, 3]	Diffusion +Kinetics
LFP (low freq.)	Membrane Potential	Multi-Exponential/ Alpha [7]	K^+ Leak +Kinetics
Dendritic Recording	Post-Synaptic Potential	Multi-Exponential/ Alpha [1, 8]	K^+ Leak +Kinetics

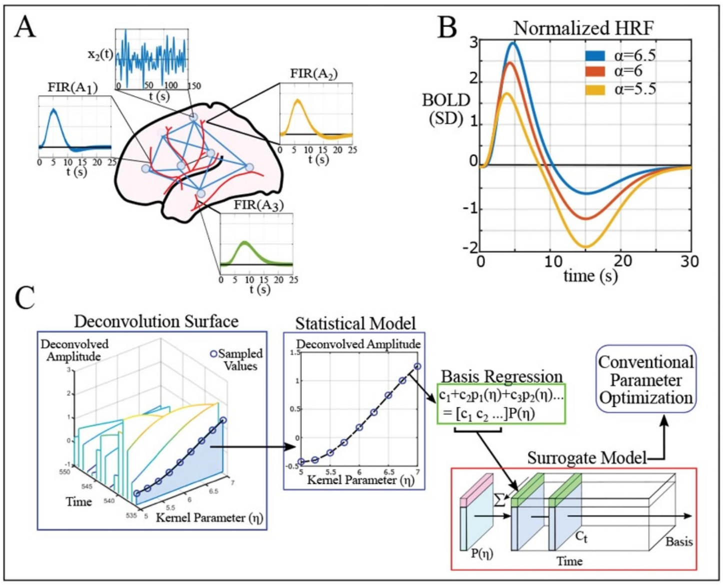

Figure 1. General Approach. (A) The brain is an example of a convolutional system when viewed through BOLD fMRI. Dynamics among brain regions are highly non-linear and usually cannot be directly observed. (B) Measurements made using fMRI reflect latent brain activity passing through a hemodynamic response function (HRF). (C) Surrogate Deconvolution workflow: a deconvolution surface is estimated by sampling the time-series deconvolution across a variety of kernel parameters (left). A separate surrogate model is formed for each time-point using basis regression to approximate this surface (middle). The combined surrogate models then represent the deconvolution process during parameter estimation.

variables are implicitly generated via linear time-invariant (LTI) systems. This issue is not inherently problematic, as many neural models contain simple exponential kernels which are easily converted to an additional LTI variable (e.g. table 1). However, specific domains feature higher-order kernels such as the Hemodynamic Response Function (HRF) that relates latent neural activity and observed BOLD signal in fMRI. Approximating the HRF through a linear time-invariant (LTI) system requires multiple additional layers of state-variables which greatly increases the difficulty of estimating neural activity and also increases the overall computational burden.

2. Approach

We propose to treat this problem by directly performing optimization within the latent state-space using Surrogate Models to replace the state-estimation step (figure 1(C)). Surrogate functions comprise a means to approximate computationally intensive functions, typically through a linear combination of *a priori* specified non-linear bases (e.g. polynomial families, radial-basis functions etc). In the current case, we propose using surrogate models to explicitly estimate latent variables by deconvolving the measured time-series by the current estimate of the convolutional kernel at each iteration. Deconvolution is typically performed either by iterative algorithms such as the Richardson-Lucy algorithm [17, 18], Alternating Direction Method of Multipliers (e.g. [19]; ADMM) or explicit transformations in the Fourier domain. The proposed surrogate techniques are compatible with any combination of deconvolution algorithm and additional signal processing that are smooth with respect to the kernel parameters. In a later example with empirical fMRI data, we use the Wiener-deconvolution [20] coupled with variance normalization in the time-domain:

$$x_i(t) \approx \frac{w(z_i(t), h_i(\eta_i), K_i)}{\sigma(w(z_i(t), h_i(\eta_i), K_i))} \tag{4}$$

$$w(z_i(t), h_i(\eta_i), K_i) := \mathcal{F}^{-1}\left[\frac{\mathcal{F}^*[h_i(\eta_i)]\mathcal{F}[z_i(t)]}{|\mathcal{F}[h_i(\eta_i)]|^2 + K_i}\right] \tag{5}$$

with $w(z_i, h_i(\eta_i), K_i)$ denoting the Wiener deconvolution of z_i with respect to kernel $h_i(\eta_i)$ and noise-factor K_i equal the mean power-spectral density of the measurement noise $\nu_i(t)$ divided by the mean power spectral density of z_i. We denote standard deviation by σ and $\mathcal{F}, \mathcal{F}^*$ denote the Fourier transform and its complex conjugate, respectively. Through deconvolution, we reduce the dual-estimation problem to conventional system identification with the convolutional kernel as an additional free parameter. Using surrogate models we reduce deconvolution-algorithms into simple, differentiable functions of the kernel parameters (figure 2(A)). Thus rather than solving the dual estimation problem:

$$\arg\min_{\hat{\theta}, \hat{\eta}, \hat{x}_t}\left(J(\hat{\theta}, \hat{\eta}, \hat{x}_t, z_t)\right) \tag{6}$$

for some loss function J, we solve the parameter-estimation problem:

$$\arg\min_{\hat{\theta}, \hat{\eta}}\left(J(\hat{\theta}, \hat{\eta}, S(t, \hat{\eta}), z_t)\right); \quad S(t, \hat{\eta}) \approx h(\hat{\eta}) *^{-1} z_t \tag{7}$$

for which S denotes the Surrogate Deconvolution model and $*^{-1}$ is a user-defined deconvolution algorithm, potentially incorporating priors on the distribution of $\nu_i(t)$ and further signal processing (e.g. normalization or additional filtering). In later experiments, we set J as the mean-squared error of 1-step predictions, e.g.

$$J = E_{t \in T}\left[||z_{t+1} - f(\theta, S_t(\{\eta\}), z_t)||_2^2\right] \tag{8}$$

with T denoting the set of initialization times during training. The surrogate model S is a linear combination of smooth, non-linear bases and is therefore smooth for both iterative deconvolution algorithms, such as Richardson-Lucy, and for explicit transformations. Thus, algorithms which are natively nonsmooth due to randomization or stopping criteria (e.g. Richardson-Lucy) are converted to an accurate, but differentiable form via the Surrogate representation (e.g. figure 2(B)). The remaining, (surrogate-assisted) fitting problem is thus amenable to highly scalable techniques such as gradient-based optimization.

2.1. Contributions

Our contribution in this regard is generating surrogate models to explicitly approximate the deconvolution process in a computationally-efficient closed form. To our knowledge, previous approaches have not sought to estimate non linear models using parameterized deconvolution. We do so in a two-step process. First, we build a surrogate model of the deconvolution process (deconvolving $z_i(t)$ by $h_i(\eta_i)$ as a direct function of the kernel parameters η_i). We fit one surrogate function per measurement in the deconvolved space: the value of a deconvolved channel evaluated at a specific time. For a fixed basis, this representation may be fit rapidly at scale. For example, the empirical brain data treated later requires nearly two million surrogate functions per subject (419 brain regions × 4444 time points), all of which can be fit in seconds as the only computation of non linear complexity is shared across time points (equation (9)). In the second step, we directly integrate the surrogate model into optimization algorithms. By doing so, the time-course of each latent state-variable is expressed as a direct, easily differentiable, function of the kernel parameters (η).

We present these results as follows. First we introduce surrogate methods and the proposed technique, Surrogate Deconvolution, in which surrogate models of the latent variable are directly integrated into the optimization procedure. In the next section we consider the special case of gradient-based optimization and demonstrate how error-gradients are efficiently back-propagated through the Surrogate Model. We then test Surrogate Deconvolution in two sets of experiments. First, we consider a low-dimensional case (a small LFP model) in which existing techniques for dual-estimation remain tractable. This simplified setting allows us to benchmark Surrogate Deconvolution's accuracy in parameter/state estimation relative the joint-Extended Kalman Filter and the joint-Unscented Kalman Filter. Results demonstrate that Surrogate Deconvolution is competitive even within the Kalman Filter's operating domain. We then consider more complicated fMRI models in which current dual-estimation techniques are not applicable due to high-dimensionality and kernel complexity. We demonstrate that Surrogate Deconvolution is robust to spatial variation in the HRF kernel in contrast to state-of-the-art non-dual approaches. Lastly, we demonstrate the approach's feasibility to empirical fMRI data. Thus, Surrogate Deconvolution performs competitively within the scope of current dual-estimation approaches and enables robust dual-estimation for a much larger set of problems than previously considered.

3. Surrogate deconvolution

Our procedure contains two parts. First, we construct a surrogate function for each channel and time-point,

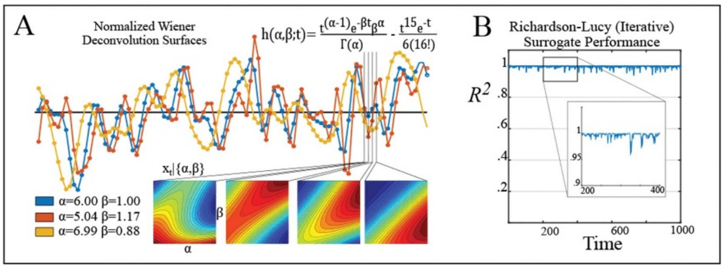

Figure 2. Surrogate surfaces for performing deconvolution. (A) Example fMRI time-series deconvolved by hemodynamic models with different values for the kernel parameters (α, β). The inset equation is the hemodynamic response function (kernel). Subplots show the surrogate surfaces for four sequential time-points (e.g. the deconvolved signal amplitude at that time as a function of kernel parameters). Note that the surrogate surfaces at a fixed time-point are smooth, whereas the variation in deconvolved signal across time is much less regular. (B) Performance in reconstructing the iterative Richardson-Lucy deconvolution of the same signal across an 81x81 grid (same range as (A) using third-order bivariate polynomials (appendix section A.4). Inset shows a representative stretch of 200 points (144 s).

a process which can be massively parallelized, if necessary. We use the surrogate construction to express the estimation of unobserved state variables as a direct function of η. The surrogate function then replaces unobserved variables in a user-chosen identification-algorithm for fully observable systems. This process is advantageous as it enables direct calculation of how changing parameters η influence the final estimate of unobserved state variables (for the current set of parameters) as opposed to techniques such as the dual Kalman Filter which do not 'look-ahead' to see how changing downstream parameters will affect state estimates (since $\nabla_\eta f = 0$ without a surrogate model).

The key insight underlying surrogate deconvolution regards the effect of varying a kernel parameter. As demonstrated in figure 2, changing a kernel parameter produces intricate effects upon deconvolved estimates when viewed from the time-domain. Even when these effects can be expressed analytically (as in the Wiener deconvolution) they are not readily reduced to a temporally-local calculation using first-principles. However, when the kernel is lower frequency than the signal (as usually happens in biology), the effect of kernel variation on a single estimate is often quite smooth with respect to the kernel parameter. Thus, the effect of kernel variation on a single deconvolved estimate is very well-approximated by simple functions of the kernel parameter. Together, these functions comprise the surrogate model.

3.1. Building surrogate representations

We efficiently define and evaluate surrogate models by storing coefficients in tensor format. For a vector of m_i stacked basis functions $P_i(\eta_i) : \mathbb{R}^{v_i} \to \mathbb{R}^{m_i}$ we define the 3-tensor C defined for each channel ('i') and a prior distribution on η:

$$C_{i,t,:} = E_{\eta_i}[w_i(t)P_i^T]E_{\eta_i}[P_iP_i^T]^{-1}. \tag{9}$$

Thus, C stores the coefficients of regressing the basis functions P_i on the deconvolved time series w_i (one of P_i's bases should be $[P_i]_j = 1, \forall \eta_i$ to provide the intercept). For clarity of presentation, we have reduced the input arguments of w_i to time alone. By E_{η_i} we denote the expectation taken over some prior distribution on η_i. In practice, the choice of prior is not usually impactful, as an arbitrarily fine sampling of the response surface can be quickly computed in parallel and the surrogate goodness-of-fit can be similarly increased by adding additional (linearly independent) basis functions. In all later examples, we simply assume a uniform distribution over reasonable bounds. The tensor C holds coefficients of each time-point's surrogate model with $C_{i,t,j}$ denoting the coefficient of basis j in predicting the deconvolution of channel i at time t in the deconvolved time-domain (which is shifted from the measurement times). We evaluate the surrogate functions in parallel by defining the following product between 3-tensor C and a 2-tensor-valued function $[P(\{\eta\})]_{i,j} := [P_i(\eta_i)]_j$:

$$[P(\{\eta\}) \star C]_{i,t} = \sum_j [P_i(\eta_i)]_j C_{i,j,t} \approx x_i(t)|\eta_i \tag{10}$$

with the right-hand side denoting the optimal estimate of $x_i(t)$ (e.g. in the least-squares sense for Wiener deconvolution) given $\eta_i, z_i(t)$ and any fixed priors used to define the chosen deconvolution. In principle, this technique could be used for system identification objectives in which errors are defined in terms of predicting x_t or z_t or both. In practice, however, we have found that including x_t predictions within the objective function leads to a moving-target problem in which identification algorithms enter periods of attempting to maximize auto-covariance (by changing η). Therefore, we assume that objectives are given of the form:

$$J = \sum_{k \in \hat{\mathbf{k}}} \left(J_k \left([z_{t+k}]_{Actual}, [\bar{\mathbf{z}}_{t+k}|\theta, z_t, \{\eta\}] \right) \right). \tag{11}$$

The final cost function J is a sum of the sub-costs J_k evaluated at the time-steps $k \in \hat{\mathbf{k}}$. Here, $\hat{\mathbf{k}}$ denotes the user-determined time steps at which to evaluate the cost function J which potentially varies by timestep (e.g. choosing to weight temporally distant predictions less). The right-hand side denotes the current estimate ($\bar{\mathbf{z}}$) of z_{t+k} given θ, $\{\eta\}$, and z_t. Thus, the new cost function incorporates the actual measurements and their prediction. However, unlike conventional dual approaches, the predictions are a direct, explicit function of previous measurements, rather than in terms of both measurements and an iteratively estimated latent variable.

3.2. Deploying surrogate models

To evaluate the cost function, we make forward predictions in the latent-variable (deconvolved) domain and then convolve those predictions forward in time to evaluate error in terms of observations. For k-step predictions and kernel length τ, this corresponds to:

$$\bar{\mathbf{z}}_{t_0+k|t_0} := h * [f^k(x_{t_0-\tau}, z|\eta)\ \ f^k(x_{t_0-\tau+1}, z|\eta)\ldots] \tag{12}$$

$$= \sum_{k=1}^{\tau}\left(h_{1+\tau-k}\circ f^{t-t_0}(P(\{\eta\})\star C_{t_0+k-\tau}, z)\right) \tag{13}$$

We use $\bar{z}_{t_0+k|t_0}$ to denote the estimate of z_{t+k} using initial conditions for the convolutional variable (z) and latent variable (x) prior to t_0. The operator ° denotes the Hadamard product (element-wise multiplication). In the latter equation, we have condensed notation for the effect of z on f as follows:

$$f^{k+1}(x_t, z) := f(f^k(x_t, z), z_{t+k}) \tag{14}$$

with $f(x_t, z) := f(x_t, z_t)$. Thus, f^k is not a proper iterated composition when it accepts both x_t and z_t as arguments, since only one variable (x_{t+1}) is output. However, we abuse this notation for clarity of presentation. For brevity, we also use $\hat{*}$ to indicate convolution over initial conditions as indicated in the variable indices. Hence the earlier equation (equation (13)) condenses to:

$$\bar{\mathbf{z}}_{t|t_0} := h\hat{*}f^{t-t_0}(P(\{\eta\})\star C_{[t_0-\tau,t_0]}, z) \tag{15}$$

As a general technique for re-representing dual estimation problems, Surrogate Deconvolution is compatible with most estimation techniques. However, the approach is particularly advantageous for gradient-based estimation as the deconvolution process is replaced with an easily differentiable surrogate form. For single-step prediction, the resulting error gradients for the non linear plant's parameters (θ) and the convolution kernel parameters ($\{\eta\}$) are as follows:

$$\frac{\partial J}{\partial\theta} = \frac{\partial J}{\partial\bar{\mathbf{z}}}\left(h(\{\eta\})\hat{*}\frac{\partial f(\theta, P\star C_t, z_t)}{\partial\theta}\right) \tag{16}$$

$$\frac{\partial J}{\partial\{\eta\}} = \frac{\partial J}{\partial\bar{\mathbf{z}}}\left[\frac{\partial h}{\partial\{\eta\}}\hat{*}f + h\hat{*}\left(\frac{\partial f}{\partial x}\left[\frac{\partial P}{\partial\{\eta\}}\star C_t\right]\right)\right]. \tag{17}$$

Thus, surrogate deconvolution re-frames dual-estimation problems into conventional parameter-estimation problems which are amenable to gradient-based approaches. The analogous gradients for multi-step prediction are derived by augmenting the one-step prediction gradients with back-propagation through time. We demonstrate the power of surrogate deconvolution by reconstructing large brain network models from either simulated data or empirical recordings.

4. Data-driven model identification

We present two applications to brain discovery to illustrate the advantage of surrogate deconvolution-enhanced methods for both state-estimation (Kalman-Filtering) and grey-box parameter identification. Both examples are dual-estimation problems (state and parameter), but we assess their performance in the state and parameter components separately to make comparisons with existing work which may be particularly designed for either domain. For instance, dual-estimation using the joint unscented Kalman Filter has been particularly successful in parameterizing black-box models for filtering (e.g. [15]), but requires further modification in some more complicated grey-box models. To demonstrate the adaptability of surrogate deconvolution we consider two different simulated system identification /estimation problems and one empirical application.

4.1. Modeling and isolating local activity from the LFP

Our first example compares performance across methodologies designed for state-estimation. This simulated problem consists of identifying the wiring of a neural system and subsequently reconstructing cellular activity from simulated extracellular recording of the 'local' field potential (LFP). This signal is primarily generated by the combined currents entering into the local population of nerve cells as opposed to the currents directly generated by neural firing. Thus, the measured LFP reflects the temporally extended effects of input into a brain region rather than the current population activity (figure 3(A)). To describe this process, we use a three-level discretized-model combining 10 coupled neural-mass models ($n_{pop}=10$) with passive integration of post-synaptic currents:

$$x_{t+1} = a\zeta(by_t) + \frac{x_t}{\tau} + c + \epsilon_t \tag{18}$$

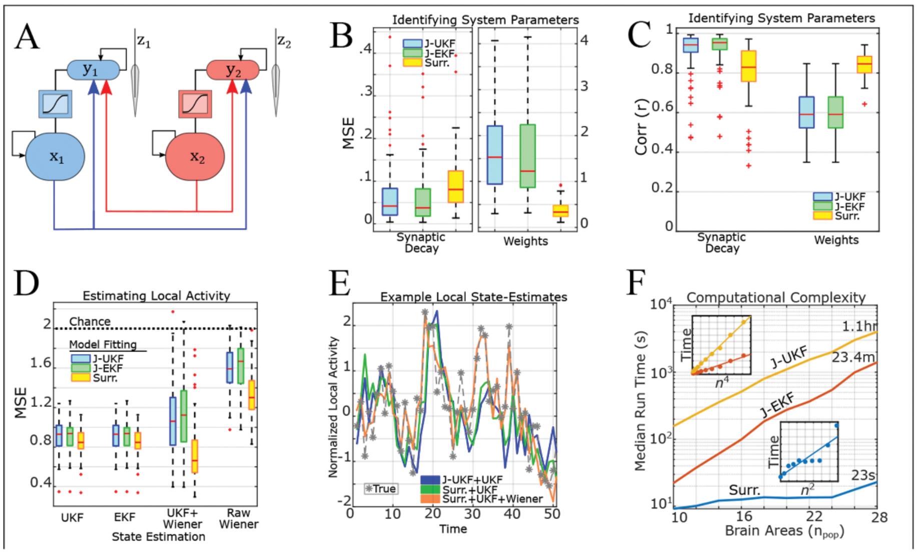

Figure 3. Surrogate deconvolution's performance in inverting a neural-mass model of Local Field Potential. (A) Model schematic: output signals from each population arrive at post-synaptic terminals. Electrode measurements primarily reflect the post-synaptic potentials generated from synaptic activity. (B) Bench-marking total error in identifying the synaptic time-constants and network connectivity. Surrogate deconvolution is compared to the current gold-standard: joint-Kalman Filters (Unscented and Extended). (C) Same as (B), but displaying performance in terms of correlation rather than mean-square-error. (D) Performance in reconstructing local spiking activity from electrode measurements using the identified system models for a variety of state-estimation techniques. (E) State estimation performance for a representative case (the simulation with median jUKF+UKF performance). (F) Computational complexity of system identification methods in terms of the number of brain regions considered. The top inset shows the run-time and least-squares fit on rescaled x-axes (n_{pop}^4) to demonstrate the $\mathcal{O}(n_{pop}^4)$ complexity of Kalman-Filtering in terms of the number of brain regions (n_{pop}). The bottom inset is the same, but for surrogate deconvolution and $\mathcal{O}(n_{pop}^2)$. Calculations were performed single-core on an Intel Xeon $10^{-5}-2630v3$ CPU.

$$y_{t+1} = S \circ y_t + Wx_t + \omega_t \tag{19}$$

$$z_t = y_t + \nu_t \tag{20}$$

Here, x is the synaptic-gating variable which describes neural activity. The sigmoidal transfer-function is denoted $\zeta(x) := (1+\exp[-x/5])^{-1}$ with scaling coefficient $a = 3$. The time constant of x is denoted τ and baseline drive to x is denoted c. The parameters $a, \tau \in \mathbb{R}$ and $c \in \mathbb{R}^{n_{pop}}$ are assumed known as are the covariances of process noise ϵ_t, ω_t and measurement noise ν_t (see appendix). Thus, the unknown parameters are the connections between neural populations (W) and the synaptic time-constants S. We considered two general approaches to system identification: either using the current gold-standard (joint Kalman estimation) or using surrogate deconvolution for least-squares optimization. The joint Kalman filter and associated variants operate identically to the original Kalman filter, except that the state-space model is augmented with unknown parameters being treated as additional state-variables with trivial dynamics (e.g. $W_{t+1} = W_t + \hat{\epsilon}_t$ and similarly for S). The 'noise' terms $\hat{\epsilon}_t$ for parameter state-variables are assumed i.i.d. and with a user-defined variance that determines the learning rate. Based upon numerical exploration, we found that the best performance for both EKF and UKF was with an initial prior on parameter variance $var[\hat{\epsilon}] = .001$. Every 50 time-steps we decreased the variance prior by 5% of its current value.

For comparison with existing techniques we used both the joint-Extended Kalman Filter (jEKF) which linearizes the non-linear portion of dynamics and the joint-Unscented Kalman Filter (jUKF) which directly propagates noise distributions through nonlinearities using the Unscented Transformation [14]. We compare these traditional methods with system identification through surrogate deconvolution. The benefit of surrogate deconvolution is the ability to apply a wide variety of optimization techniques to partially-observable identification problems which can decrease computation time over conventional techniques (figure 3(F)) and expand the scope of problems which may be tackled. For this first example, we have chosen a relatively simple case (low-dimensional, single-exponential kernels) so that conventional methods (Kalman Filtering) apply. Therefore, the goal of this test is not to demonstrate an overwhelming advantage of surrogate deconvolution over Kalman Filtering, but to show that the proposed technique can perform competitively in cases for which Kalman-Filtering is applicable, but imperfect. Subsequent examples will consider cases in which Kalman Filtering is not tenable.

To implement Surrogate Deconvolution, we first reformulate this problem as a convolutional equation through the change of variable $r_t := Wx_t$:

$$r_{t+1} = aW\zeta(y_t) + \frac{r_t}{\tau} + Wc + W\epsilon_t \tag{21}$$

$$y_{t+1} = S \circ y_t + r_t + \omega_t \tag{22}$$

or, equivalently,

$$r_{t+1} = \frac{r_t}{\tau} + W[a\zeta([(r+\omega) * \mathcal{P}(S)]_t) + c + \epsilon_t] \tag{23}$$

$$z_t = [(r+\omega) * \mathcal{P}(S)]_t + \nu_t \tag{24}$$

with $\mathcal{P}(S)$ denoting the discrete-time kernel formed from polynomials of S to a suitably long length $[0\ 1\ S\ S^2\ S^3 \ldots]$ analogous to exponential decay for continuous-time systems. In this form, the parameters can be estimated using traditional least-squares methods, optimizing over W and S. However, by leveraging the tensor representation of surrogate models, this equation can be reduced into a single equation in S by representing the optimal choice of W for a given S as a direct function of S. To do so we define the matrix

$$F_t := a\zeta(z_t) + c \tag{25}$$

and the associated 3-tensor

$$M_{i,j,p} = (E_t[(C_{t+1,p} - C_{t,p}\tau^{-1})F_t^T]E_t[F_tF_t^T]^{-1})_{i,j}. \tag{26}$$

Each $n \times n$ page of this tensor (e.g. the matrix formed by holding p constant) stores the coefficients of the least squares solution for W in predicting $C_{t+1,p} - C_{t,p}\tau^{-1}$ from F_t for the p^{th} basis function. Since $r_t^*(S) = C_t \star P(S)$, for a given synaptic decay term S we use the notation $r^*(S)$ to denote the estimate of r produced through Surrogate Deconvolution of measurements z with the kernel $\mathcal{P}(S)$. This produces the least-squares estimate for W as a direct function of S:

$$\arg\min_W ||r^*_{t+1}(S) - (WF_t + r^*_t(S)\tau^{-1})||_F^2 = \mathcal{P}(S) \star M, \tag{27}$$

$$z_{t+2} \approx Sz_{t+1} + [\mathcal{P}(S) \star M]F_t + [\mathcal{P}(S) \star C_t]\tau^{-1} \tag{28}$$

Thus, in this case, surrogate deconvolution enables the approximation of $2n_{pop}$ difference equations containing $n_{pop}(n_{pop}+1)$ unknown parameters (W and S) using only n_{pop} difference equations with n_{pop} unknown parameters (only S). The resultant model (from equation (28)) is also compatible with a wide variety of optimization techniques. For simplicity, we fit the parameters S through ordinary least-squares optimization in terms of predicting z_{t+2} as in equation (28). Optimization was performed using Nesterov-Accelerated Adaptive Moment Estimation (NADAM; [26]) with both NADAM memory parameters set equal to .95, and the NADAM regularization parameter set to .001. Training consisted of 15 000 iterations with each minibatch containing 1000 time points. The step size (learning rate) of updates was .0001.

All methods were able to retrieve accurate estimates of the synaptic decay term S (figures 3(B) and (C)). The best-performing method varied by simulation (e.g. for different true values of W, S), but the mean error was greater for surrogate deconvolution than Kalman Filtering methods (Unscented and Extended) which performed near-identically. By contrast, surrogate deconvolution always provided a more accurate estimate of the connectivity weight parameter (W) and the advantage relative Kalman-Filtering was substantial (figures 3(B) and (C)). The poor performance of the Kalman Filter for identification in this case is not surprising as the Kalman Filter is known to be non-robust for large systems [24] and the W parameter adds 100 additional latent state-variables to the joint Kalman model as opposed to the 10 state-variables added by S.

4.2. Reconstructing firing-rate from LFP

We next examined the ability of each method to recover the time series of neural activity x_t using the previously generated state-space models. During this stage, models produced during the previous identification step were used to estimate the latent state variable x_t (figures 3(D) and (E)). It is important to distinguish between state-estimation techniques (e.g. UKF) which we used to estimate x_t from previously-fit models and the techniques used to fit those initial models (e.g. jUKF) as these steps need not 'match' (e.g. UKF-based state-estimation from a surrogate-identified model). Measurements consisted of simulated extracellular voltages z_t generated by resimulating the same ground-truth model (i.e. the same parameters, but new initial conditions and noise realizations). As in the identification stage, we considered two general approaches to recovering the latent variable x_t: either through deconvolution or Kalman Filtering (unscented and extended). Kalman filtering in this setting produces direct estimates of x_t and y_t. By contrast, deconvolving $z_t \approx y_t$ produces an estimate of r_t, so deconvolution estimates of x_t were produced by premultiplying the deconvolved time series with W_{est}^{-1} (the inverse estimated connectivity parameter). We considered deconvolution applied either directly to the raw measurements (z_t) or to the estimates of y_t produced by Kalman filtering z_t with the estimated models (both unscented and extended Kalman filters were considered). Noise covariance estimates for Kalman filtering at this stage were the same as those assumed in the initial stage: a value close to the mean tendency, rather than the true values which were randomly selected for each simulation.

We found that the type of Kalman Filter used for state-estimation had no appreciable effect upon accuracy (figure 3(D)). Likewise, the technique used for system identification (surrogate deconvolution vs. EKF/UKF) had little effect, although surrogate deconvolution was slightly more accurate on average. However, model performance differed greatly for deconvolution-based state-estimation (using $x_t \approx W^{-1}[\mathcal{P}(S) *^{-1} y_{est}]_t$). Models estimated using joint-Kalman Filtering (jEKF/jUKF) performed worse using deconvolution-based estimation of x_t than Kalman-based estimation (figure 3(D)). This result is unsurprising as the deconvolution-based estimate additionally requires the inverse weight parameter W^{-1} and both jUKF and jEKF poorly estimated W. Interestingly, however, estimation accuracy for surrogate-identified models decreased when using deconvolution of the raw, unfiltered measurements, but increased for the UKF+decconvolution hybrid. The former result is not surprising as pure deconvolution is clearly suboptimal in not considering the noise covariance. The latter result was unexpected and it suggests the possible benefit of using a two-stage estimation procedure in which Kalman-Filtering first dampens measurement noise and improves estimates of measurable state-variables. Then, subsequent deconvolution might improve the estimate of latent state-variables by considering the impact of estimates across time, rather than just the directly subsequent measurement. However, these benefits are likely situation-dependent and therefore require more study. In any case, results indicate that state-estimates from models produced by surrogate-deconvolution are at least as accurate as those produced by Kalman-Filtering and potentially more so depending upon the state-estimation procedure (figures 3(D) and (E)).

4.3. Computational efficiency

Surrogate deconvolution is also computationally efficient as it scales linearly with the number of measurement channels ($\mathcal{O}(n)$) in both forming and evaluating surrogate functions which is also parallelizable across channels. However, since Surrogate deconvolution is not a system identification procedure in and of itself, time-savings depend upon how the technique is used (e.g. which optimization scheme it is coupled to). The advantage of surrogate deconvolution is that it can be combined with a wide-variety of optimization techniques which are otherwise ill-suited to partially-observable problems. In this first simulation, for instance, the number of parameters scale with $n_{pop}(n_{pop}+1)$ and the number of state variables (in the native space) scale with $2n_{pop}$. Thus, the dominant complexity of joint-UKF and joint-EKF is greater than $\mathcal{O}(n_{pop}^4)$ as joint-UKF/EKF are $\mathcal{O}(n^2)$ in the number of parameters and at least $\mathcal{O}(n^2)$ in the number of native (non-parameter) state-variables. By contrast, the gradient approaches applied with surrogate deconvolution have approximately $\mathcal{O}(n_{pop}^2)$ complexity (figure 3(F)). However, surrogate deconvolution is not limited to gradient-based approaches. The main effect is to simplify error functions to a direct equation in the measurable variables so surrogate deconvolution is compatible with a wide variety of non-gradient techniques, as well (e.g. heuristic-based or Bayesian). As such, surrogate deconvolution presents the opportunity to identify significantly larger partially-observable systems than previously considered.

4.4. Reconstructing Connectivity and Hemodynamics in Simulated fMRI

For our second example, we considered the ability to correctly parameterize large-scale brain models from simulated fMRI data. Brain regions were modeled through the continuous-valued asymmetric Hopfield model [25] and simulated fMRI signals were produced by convolving the simulated brain activity with randomly parameterized Hemodynamic Response Function (HRF) kernels [4]:

$$x_{t+\Delta t} = W[\tanh(b_0 \circ x_t)]\Delta t + (1-\Delta t)Dx + \epsilon_t \tag{29}$$

$$z_t = [x * h(\alpha,\beta)]_t \tag{30}$$

$$h_i(\alpha,\beta,t) = \frac{t^{\alpha_i-1}e^{-\beta_i t}\beta_i^{\alpha_i}}{\Gamma(\alpha_i)} - \frac{t^{15}e^{-t}}{6(16!)} \tag{31}$$

Parameter distributions for simulation are detailed in the appendix. Simulations were integrated at $\Delta t = 100$ ms and sampled every 700 ms (mirroring the Human Connectome Project's scanning TR of 720 ms [27]). Simulated HRF's (h_i) were independently parameterized for each brain region according to the distributions $\alpha_i \sim \mathcal{N}(6,\sigma^2)$ and $\beta_i \sim \mathcal{N}(1,(\sigma/6)^2)$ in which the variability term σ was systematically varied. Each HRF can be well approximated by a finite-length kernel and therefore can be represented as a discrete-time linear plant. However, doing so, in this case, requires multiple hidden state-variables per region which impairs Kalman-based dual-estimation procedures. Instead, most current procedures to deal with fMRI-based systems identification at scale ignore inter-regional variability in h_i and instead seek to retrieve x_t by fixing HRF parameters (e.g. [30, 31]) to the so-called 'canonical HRF' (e.g. $\alpha = 6, \beta = 1$). In this example, we demonstrate the potential pitfalls of this assumption and the benefits accrued by efficiently fitting hemodynamics through Surrogate Deconvolution. To do so, we attempted to reconstruct W using Mesoscale Individualized NeuroDynamics (MINDy) in either its base form (which assumes a canonical HRF) or in an augmented form in which the predictions are

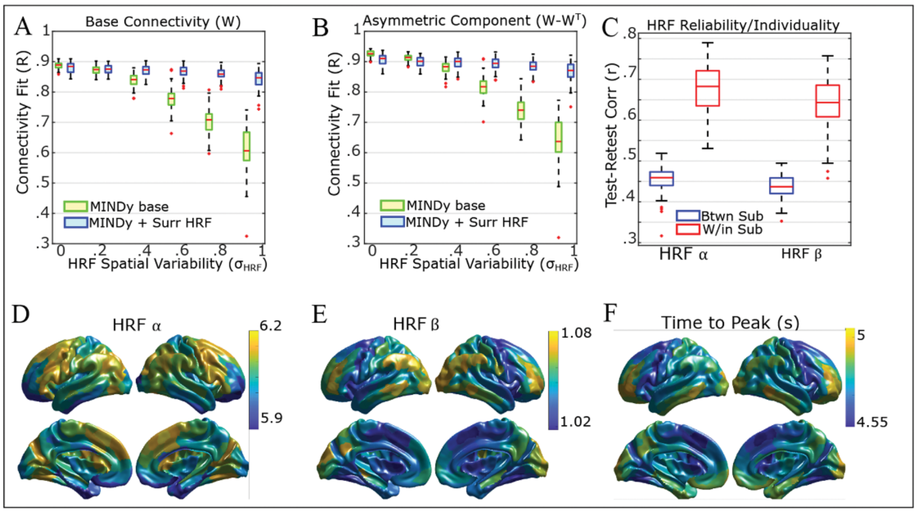

Figure 4. Incorporating HRF surrogate-deconvolution into MINDy. (A) Without HRF modeling, connectivity estimates degrade with spatial variability in the neurovascular coupling. Fitting the HRF through surrogate deconvolution preserves performance. (B) Same as (A) but for the asymmetric component of connectivity. (C) HRF parameter estimates from HCP data are reliable across scanning days and subject-specific. (D) Spatial map of the mean α parameter estimate across subjects. (E) Same as (D), but for the second HRF parameter (β). (F) Spatial map of the mean time-to-peak in the fitted HRF's.

calculated as in equation (13). MINDy model fitting consists of using NADAM-enhanced gradient updates [26] to minimize the following cost function:

$$J = \frac{1}{2}E_T[\|(x_{T+\Delta t} - x_T) - [(W_S + W_L)\psi_\gamma(x_t) - Dx_T]\|_2^2] + \lambda_1\|W_S\|_1 + \lambda_2 \mathrm{Tr}(|W_S|) + \lambda_3(\|W_1\|_1 + \|W_2\|_1) + \frac{\lambda_4}{2}\|W_L\|_2^2 \tag{32}$$

in which the estimated weight matrix $\hat{W}$ is decomposed into the sum of estimated sparse (W_S) and low-rank (W_L) components satisfying:

$$\hat{W} = W_S + W_L = W_S + W_1 W_2^T \tag{33}$$

for some $W_S \in \mathbb{M}^{n\times n}$ and $W_1, W_2 \in \mathbb{M}^{n\times k}$. The hyperparameter $k < n$ determines the rank of the low-rank component W_L and the regularization hyperparameters $\{\lambda_i\}$ define statistical priors on each of the weight matrix components (Laplace prior for W_S, W_1, W_2 and a normal prior for $W_L := W_1 W_2^T$). This decomposition has been shown useful to estimating large brain networks [30]. The non linear function ψ is parameterized by the parameter vector $\gamma \in \mathbb{R}^n$ with

$$\psi_\gamma(x) = \sqrt{\gamma^2 + (bx_t + .5)^2} - \sqrt{\gamma^2 + (bx_t - .5)^2} \tag{34}$$

For the Surrogate-Deconvolution, however, these analyses are performed in the original space to prevent the afore-mentioned moving target problem. Hyper-parameter determination and simulation parameters are detailed in the appendix.

Results demonstrated a clear benefit for additional modeling of the local hemodynamic response (figures 4(A) and (B)). When hemodynamics differed only slightly between simulated brain regions both methods produced highly accurate estimates of the connectivity parameter W. However, past $\sigma = .4$ (the SD of spatial variation in one of the HRF parameters), the accuracy of estimated connectivity rapidly decreased for conventional methods, while only slightly decreasing for surrogate deconvolution. In addition, the hemodynamic parameter estimates also became increasingly accurate. Thus, surrogate deconvolution enables accurate system (brain) identification when the downstream plants (hemodynamics) are variable and unknown.

4.5. Empirical dual estimation with the human connectome

Lastly, we tested the effects of using Surrogate Deconvolution in fitting MINDy models to data from the Human Connectome Project [27]. By using empirical data, this analysis demonstrates that human hemodynamics are spatially variable and that accounting for this variability produces more nuanced and reliable brain models. Data consists of one hour of resting-state fMRI per subject spread across two days (30 minutes each). Data were processed according to the recommendations of Siegel and colleagues [29] and divided into 419 brain regions [28] +19 subcortical. We then fit MINDy models either with or without surrogate deconvolution to this data using the same fitting procedure and hyperparameters [30] as before. Results indicated the the parameters which describe the hemodynamic response function are reliable across scanning days and reliably differ between individuals (figure 4(C)). Each of the two HRF parameters had a stereotypical spatial distribution (figures 4(D) and (E)) as did the time-to-peak of the recovered HRF kernels (generated by substituting in the recovered kernel parameters). Time to peak was slowest for anterior prefrontal cortex, particularly in the right hemisphere (figure 4(F)). Because current knowledge of the 'true' hemodynamic response is limited, future study establishing ground-truths for HRF variation across human cortex is needed to facilitate more rigorous empirical validation.

5. Discussion and conclusion

5.1. Generalizability of the problem framework

The methods that we propose are dependent upon the problem satisfying two criteria: 1) unmeasured variables can be related to measured ones via convolution and 2) the form of the convolutional kernels are known up to a small number of parameters per kernel. These assumptions are satisfied in many areas of neuroscience (see table 1) in which measurements have high spatial precision relative to the underlying models. These kernels can also be derived by analytically reducing large state-space models to a smaller convolutional form (see section A.5). In state-space formulation, these problems all contain more state-variables than recording channels, but they all still contain one channel per anatomical unit (region, population, cell etc depending upon the model). In other words, the inverse-problem from these scenarios results from mechanistic undersampling (i.e. only measuring one type of variable) rather than spatial undersampling.

The relationship between the measured and unmeasured variables can be either unidirectional (e.g. neural activity influencing BOLD but not vice-versa) or bidirectional (e.g. neural activity and synaptic currents influencing each other) and this formulation covers a large number of empirically relevant scenarios. However, the assumption of full spatial precision (relative the model) also limits our approach to specific modalities (see table 1). As presented, our technique is not applicable to sensor-level EEG or MEG recordings since each channel's signal reflects a (weighted) summation of activity at many anatomical locations. By contrast, other techniques such as the joint Kalman Filters (with which we compare our method) are applicable to these scenarios and simultaneously perform spatial-inversion and model parameterization. Thus, the proposed techniques only cover specific classes of modeling scenarios which are but a subset of problems in which the joint Kalman Filters are applicable. However, as we have demonstrated in the results, our approach scales much better with model size. Thus, our technique is generalizable in terms of model scale, whereas the Kalman Filter is more general with respect to model type.

5.2. The Role of priors in deconvolution

A second assumption of our technique is that the convolutional kernels are known up to a relatively small number of parameters each, thus constituting semi-blind deconvolution. This assumption holds in a wide variety of scenarios in which prior empirical evidence suggests an approximate functional form (e.g. the double-gamma HRF [4]). However, there remain cases in which the general form of the kernel is unknown, or the form contains many unknown parameters (e.g. an unknown kinetic scheme with many conformations). Fortunately, several statistical approaches to blind deconvolution exist, many of which require few assumptions regarding the kernel's form (e.g. [32, 33]). The Richardson-Lucy algorithm [17, 18, 32] is one popular example for both blind and semi-blind deconvolution when the noise statistics are Poisson. These approaches can also be applied to unknown kernels which span both time and space, whereas our technique only considers convolution in the temporal domain. For these reasons, blind-deconvolution algorithms have been previously applied to a variety of neuroscience domains (e.g. [34, 35]). The primary drawback of statistical blind-deconvolution algorithms, however, is that solutions are at most unique up to an unknown lag for each channel so it may not be possible to discern the order of latent events between channels. By contrast, the proposed method considers the dynamical relationship between channels. As a result, solutions identify the relative timing of latent neural events across channels.

5.3. Conclusion

Data-driven modeling remains one of the key challenges to neuroengineering and computational neuroscience. Although a wealth of theoretical model forms have been produced, the state-variables of these models (e.g. neuronal firing rate) are often difficult to directly measure $in-vivo$ which complicates system-identification (model-parameterization). Instead, many clinical and experimental contexts record proxy variables which reflect the physiologically downstream effects of neuronal activity (e.g. on blood oxygenation, signaling molecules, and synaptic currents). In the current work, we aimed to parameterize conventional neural models using indirect measurements of neural activity. This problem involved simultaneously estimating the generative neural model as well as the latent neural activity thus comprising a dual-estimation problem. Through surrogate models, we approximated the state-estimation step as a parameterized deconvolution,, thus reducing computationally challenging dual-estimation problems to a closed-form, conventional identification problem. The primary advantage of this approach is speed/scalability.

Current approaches to model-based dual-estimation emphasize the joint/dual Kalman Filters and related Bayesian approaches (e.g. [16]). These approaches suffer, however, in terms of scalability and data quantity. As demonstrated in numerical simulations, the computational complexity of Kalman-Filtering limits application to relatively small models (figure 3(F)), whereas Surrogate Deconvolution enables optimization techniques that scale well with the number of parameters (figure 3(F)). However, despite requiring orders of magnitude fewer computations, Surrogate Deconvolution performs competitively with Kalman Filtering in estimating system parameters (figures 3(B) and (C)) as well as estimating states (latent neural activity; figures 3(D) and (E)). Thus, the computational advantages of Surrogate Deconvolution do not compromise accuracy.

Scalability is particularly salient in empirical neuroimaging, as several recent approaches have eschewed detailed modeling of physiological measurements (e.g. [30, 31]) in order to increase the spatial coverage of models. However, ground-truth simulations indicate that these reductions potentially compromise accuracy (figures 4(A) and (B)). By contrast, methods augmented with Surrogate Deconvolution maintained high levels of performance (accuracy) even for extreme spatial variation in physiological signals. Interestingly, hemodynamic variation appeared to be a reliable feature in empirical data with consistent differences across individuals (figure 4(C)) and brain regions (figures 4(D)–(F)) which can potentially lead to systematic biases (as opposed to random error) when these features are not modeled. Thus, for neuroimaging in particular, it may be critical to parameterize both the generative neural model and the measurement models to account for these biases. Surrogate Deconvolution provides a means to parameterize such models without compromising the detail of either component.

Acknowledgments

MS was funded by NSF-DGE-1143954 from the US National Science Foundation. TB acknowledges R37 MH066078 from the US National Institute of Health. SC holds a Career Award at the Scientific Interface from the Burroughs-Wellcome Fund. Portions of this work were supported by AFOSR 15RT0189, NSF ECCS 1509342 and NSF CMMI 1537015, NSF NCS-FO 1835209 and NIMH Administrative Supplement MH066078-15S1 from the US Air Force Office of Scientific Research, US National Science Foundation, and US National Institute of Mental Health, respectively.

Appendix A. ‘Local’ Field Potential Simulations

The ‘local’ field potential recordings from section 4.1 were simulated using the discrete-time neural mass model:

$$x_{t+1} = a\zeta(by_t) + \frac{x_t}{\tau} + c + \epsilon_t \tag{A1}$$

$$y_{t+1} = S \circ y_t + Wx_t + \omega_t \tag{A2}$$

$$z_t = y_t + \nu_t \tag{A3}$$

The paramaters $a = 3$, $b = 1/5$, and $\tau = 2$ were fixed. For each simulation, the remaining neural-mass model parameters were redrawn from fixed, independent distributions: $c \in \mathbb{R}^{n_{pop}} \sim \mathcal{N}(-1, .25^2)$ and $S \sim \mathcal{N}(.5, .2^2) \cap [.2, .8]$. The connectivity parameter W was sampled using a two-step procedure:

$$W_0 \sim \mathcal{N}(0, .1^2) \quad W = W_0 + 2(W_0 - W_0^T). \tag{A4}$$

This exaggerated asymmetry serves to ensure solutions have nontrivial dynamics in the absence of noise. The noise processes ε_t, ω_t, and ν_t were all independent, white Gaussian processes with the same variance for each population. For each simulation the standard deviations of ε_t, ω_t, and ν_t were drawn from $.25 + .5|\mathcal{N}(0,1)|$, $.05 + .1|\mathcal{N}(0,1)|$, and $.1 + .2|\mathcal{N}(0,1)|$, respectively. The variances assumed by Kalman Filtering were .5, .1, and .2 for ε_t, ω_t, and ν_t, respectively.

Appendix B. Randomized Networks and MINDy Hyperparameters for simulated fMRI

Ground-truth simulations for BOLD fMRI (section 4.4) were produced by a 40 node Hopfield-type [25] recurrent neural network with asymmetric connectivity:

$$x_{t+\Delta t} = W[\tanh(b_0 \circ x_t)]\Delta t + (1 - \Delta t)Dx + \epsilon_t. \tag{A5}$$

Here, the timescale of integration was $\Delta t = .1$ s and measurement occurred every 700 *ms*. The process noise ε_t was Gaussian ($\sigma^2 = .625$) and independent between channels. The simulation parameters and generic MINDy fitting hyperparameters were generally identical to those in the original 40-network MINDy simulations [30]. Ground-truth connectivity parameters (W) for the simulations were generated by a hyperdistribution characterized by four hyperparameters which scale the reduced-rank magnitude (σ_1), sparseness (σ_2), degree of asymmetry (σ_a), and degree of population clustering ($\hat{p}$). These hyperparameters are distributed $\sigma_1, \sigma_a \sim \mathcal{N}(4, .1^2)$ and $\sigma_2 \sim \mathcal{N}(3, .1^2)$. The hyperparameter $\hat{p}$ is either 1 or 2 with equal probability. These parameters were used to generate three matrices (M_1, M_2, M_3) distributed as follows:

$$M_1 \sim [\mathcal{N}(0, 1/\sigma_1^2) + \mathcal{N}(0, 1/\sigma_1^2)^3]_{40/\hat{p} \times 40/\hat{p}} \tag{A6}$$

$$M_2 \sim [\mathcal{N}(0,1/\sigma_2^2)^3]_{40\times 40} \tag{A7}$$

$$M_3 \sim [\mathcal{N}(0,1/\sigma_1^2)+\mathcal{N}(0,1/\sigma_1^2)^3]_{40\times 5}\times \ldots \\ [\mathcal{N}(0,1/\sigma_1^2)+\mathcal{N}(0,1/\sigma_1^2)^3]_{40\times 5} \tag{A8}$$

To generate population clustering we use the ones matrix $1_{\hat{p}\times\hat{p}}$ and define $\hat{M}_1 := 1_{\hat{p}\times\hat{p}} \otimes M_1$ in which $\otimes$ denotes the Kronecker product. The final connectivity matrix (W) for each simulation is formed as follows:

$$Q := \hat{M}_1 + M_2 + M_3 \quad W = Q + (Q - Q^T)/\sigma_1. \tag{A9}$$

The slope vector $b_0 \in \mathbb{R}^{40}$ is distributed $b_0 \sim \mathcal{N}(6,(.5)^2)$ and the diagonal decay matrix D has (diagonal) elements i.i.d. distributed $D_{i,i} \sim \mathcal{N}(.4,.1^2)\cap[.2,\infty]$. Deconvolved time series were z-scored. Base MINDy regularization parameters for the 40-node simulation were generated by rescaling the empirical fMRI regularization parameters $(\hat{\lambda}_1 = .075, \hat{\lambda}_2 = .2, \hat{\lambda}_3 = .05, \hat{\lambda}_4 = .05)$ by $1/r_n, 1/r_n, 1/\sqrt{r_n}, and 1/r_n^2$, respectively with $r_n = 10$ is the approximate ratio between the number of empirical brain regions (419) and those used in the simulation (40) [30] which used the method described below (section A.3). The maximum-rank of the low rank component W_L was 15. Initial values for MINDy parameters were distributed as in [30]. The NADAM update rates for the HRF parameters α and β were 5×10^{-4} and 2.5×10^{-4}, respectively for the 40-node simulation. Surrogate deconvolution used the third-order bivariate polynomial basis $\{1,\alpha,\beta,\alpha^2,\alpha\beta,\beta^2,\alpha^3,\alpha^2\beta,\alpha\beta^2,\beta^3\}$ which was fit to the z-scored deconvolution surfaces.

Appendix C. Empirical Selection of MINDy Hyperparameters

The MINDy hyperparameters we used were previously determined [30] by pseudo-optimization of empirical Human Connectome Project [27] fMRI data. In the former study, values were chosen to maximize cross-validated goodness of fit while retaining a test-retest correlation (reliability) of at least .7 for each type of estimated parameter (W,α,D). In brief, values were sampled from a grid over the 4-dimensional space and used to fit models to a set of 10 left-out subjects with test-retest data (none of these subjects were reused in our analyses). The gridded fits determined the likely vicinity of local minima and the final values were chosen based upon iterated coordinate-descent with a fixed resolution (.005). More sophisticated approaches for hyperparameter selection also exist [23] and may be more efficient in future applications.

Appendix D. HCP Data for Surrogate-HRF MINDy

For the empirical data, MINDy used the original regularization parameters $(\hat{\lambda}_i)$. NADAM update rates were 2.5×10^{-4} for α and 2.5×10^{-5} for β. Resting-state fMRI from the Human Connectome Project (HCP; [27]) was preprocessed according to Siegel and colleagues [29] and smoothed via nearest-neighbor. Deconvolution was performed using Wiener's method with noise-signal-ratio $=$ .002. On each minibatch, next-step predictions were made for 250 sequential frames using an HRF kernel length of 30 TRs (21.6 s) and parameter updates were performed using NADAM for 6000 minibatches. As before, surrogate deconvolution used the third-order bivariate polynomial basis $\{1,\alpha,\beta,\alpha^2,\alpha\beta,\beta^2,\alpha^3,\alpha^2\beta,\alpha\beta^2,\beta^3\}$ to approximate the z-scored deconvolved time-series. For fitting surrogate coefficients, α was assumed uniform on $[5,7]$ and β was assumed uniform on $[.5,1.5]$. Expected values were taken by sampling this two-dimensional space along an evenly-spaced 10×10 rectangular grid.

Appendix E. Derivation of Kernels from Partially-Observable State-Space Models

Convolutional representation can reduce differential/difference equation models of large, hierarchical systems into much smaller (integro-differential) forms. These systems are hierarchical in the sense that they contain a small set of nonlocal (potentially non-linear) state-variables $(x_t \in \mathbb{R}^n)$ with an equal number of recording channels $(z_t \in \mathbb{R}^n)$. Each of these interconnected state-variables $(x_t^{(i)} \in \mathbb{R})$, however, can have several coupled local state-variables which produce linear intrinsic dynamics $(y^{(i)} \in \mathbb{R}^{k_i})$. In neuroscience applications, this scenario typically corresponds to one channel per brain area with each area defined by multiple state-variables (e.g. physiological mechanisms):

$$x_{t+1} = f(x_t, y_t) + \eta_t, \tag{A10}$$

$$y_{t+1}^{(i)} = A_i y_{t+1}^{(i)} + b_i x_t^{(i)}. \tag{A11}$$

Thus, the state-variables $y^{(i)}$ evolve according to the matrix $A_{k_i\times k_i}$. We assume that the A matrix is stable in the discrete-time sense (eigenvalues have absolute values strictly less than one) which prevents 'exploding' solutions and guarantees the existence of an equivalent convolutional form. We note that the local state-variables $(y^{(i)})$ do not need to be the same size for each $x^{(i)}$ (e.g. brain area) and they are only defined to be local in terms of input: $y^{(i)}$ can directly influence $x^{(j\neq i)}$, but not vice-versa. The measurement from each channel $z_t^{(i)}$ is a noisy linear summation of

k_i+1 state-variables: $x_t^{(i)} \in \mathbb{R}$ and $y_t^{(i)} \in \mathbb{R}^{k_i+1}$. Thus, at each instance n channels measure a system with $n + \sum k_i$ (partially) coupled state-variables.

$$z_{t+1}^{(i)} = c_i^T y_{t+1}^{(i)} + a_i x_t^{(i)} + \nu_{t+1}. \tag{A12}$$

Due to the linear intrinsic dynamics of $y_t^{(i)}$, measurements can be re-written in convolutional form:

$$z_t^{(i)} = [h_i * x^{(i)}]_t + \nu_t \tag{A13}$$

$$h_i = [a_i \quad c_i^T b_i \quad c_i^T A_i b_i \quad c_i^T A_i^2 b_i \cdots]. \tag{A14}$$

The unknown kernel parameters can factor into any of the local terms (b_i, A_i, c_i, or a_i). When little is known regarding these parameters a-priori, the mapping from state-space parameters onto the kernel (h_i) is not always invertible, so there are cases in which the parameterization problem is well-posed in convolutional form but not in state-space form (e.g. if A_i is symmetric and both c_i, b_i are unknown). Analogous results hold for the continuous-time case:

$$\dot{y}_i(t) = A y_i(t) + b_i x_i(t) \tag{A15}$$

$$h_i(\tau) = a_i \delta(\tau) + c_i^T e^{A_i \tau} b_i \tag{A16}$$

with A now required to be Hurwitz-stable (all eigenvalues have negative real-part), δ denoting the Dirac function, and $e^{A_i\tau}$ denoting the matrix-exponential of A_i multiplied by the lag τ.

ORCID iD

Matthew F Singh https://orcid.org/0000-0003-0051-336X

References

[1] Rall W 1967 Distinguishing theoretical synaptic potentials computed for different soma-dendritic distributions of synaptic input *J. Neurophysiol.* **30** 1138–68

[2] Vogelstein J T, Packer A M, Machado T A, Sippy T, Babadi B, Yuste R and Paninski L 2010 Fast nonnegative deconvolution for spike train inference from population calcium imaging *J. Neurophysiol.* **104** 3691–704

[3] Theis L, Berens P, Froudarakis E, Reimer J, Román Rosón M, Bader T, Euler T, Tolias A S and Bethge M 2016 Benchmarking spike rate inference in population calcium imaging *Neuron* **90** 471–82

[4] Friston K J, Fletcher P, Josephs O, Holmes A, Rugg M D and Turner R 1998 Event-related fMRI: characterizing differential responses *NeuroImage* **7** 30–40

[5] Lammertsma A A and Hume S P 1996 Simplified reference tissue model for PET receptor studies *Neuroimage* **4** 153–8

[6] Watabe H, Ikoma Y, Kimura Y, Naganawa M and Shidahara M 2006 PET kinetic analysis–compartmental model *Ann. Nuclear Med.* **20** 583–8

[7] Einevoll G T, Kayser C, Logothetis N K and Panzeri S 2013 Modelling and analysis of local field potentials for studying the function of cortical circuits *Nat. Rev. Neurosci.* **13** 770–85

[8] Hodgkin A L and Huxley A F 1952 A quantitative description of membrane current and its application to conduction and excitation in nerve *J. Physiology* **117** 500–44

[9] Destexhe A, Mainen Z and Sejnowski T J 1994 An efficient method for computing synaptic conductances based on a kinetic model of receptor binding, *Neural Comput.* **6** 14–18

[10] Narendra K S and Partbsarathy K 1990 Identification and control of dynamical systems using neural networks *IEEE Trans. Neural Networks* **1** 4–27

[11] Ogunfunmi T 2007 *Adaptive Nonlinear System Identification: The Volterra and Wiener Based Approaches* (New York: Springer)

[12] Chen S and Billings S A 1989 Representation of nonlinear systems: The NARMAX model *Int. J. Contr.* **49** 1013–32

[13] Kalman R E 1960 A new approach to linear filtering and prediction problems, *Trans. ASME J. Basic Eng.* **82** 34–45

[14] Julier S J and Uhlmann J K 1997 New extension of the Kalman filter to nonlinear systems *Proc. SPIE* **3068** 182–93

[15] Van der Merwe R and Wan E 2001 The square-root unscented Kalman filter for state and parameter-estimation *Proc. IEEE Int. Conf. Acoustics, Speech and Sig. Proc.* **6** 3461–4

[16] Sun X, Jin L and Xiong M 2008 Extended kalman filter for estimation of parameters in nonlinear state-space models of biochemical networks *PLoS One* **3** e3758

[17] Richardson W H 1972 Bayesian-based iterative method of image restoration *JOSA* **62** 55–9

[18] Lucy L B 1974 An iterative technique for the rectification of observed distributions *The Astronomical Journal* **79** 745–55

[19] Afonso M, Bioucas-Dias J and Figueiredo M 2011 An augmented Lagrangian approach to the constrained optimization formulation of imaging inverse problems *IEEE Trans. Image Process.* **20** 681–95

[20] Wiener N 1949 *Extrapolation, Interpolation and Smoothing of Stationary Time Series 2* (Cambridge: MIT Press)

[21] Vogelstein J T, Watson B O, Packer A M, Yuste R, Jedynak B and Paninski L 2009 Spike inference from calcium imaging using sequential Monte Carlo methods *Biophys.* J. **97** 636–55

[22] Buckner R L 1998 Event-related fMRI and the hemodynamic response *Human Brain Mapping* **6** 373–7

[23] Jones D R, Schonlau M and Welch W J 1998 Efficient global optimization of expensive black-box functions *J. Global Optim.* **13** 455–92

[24] Khan U A and Moura J M F 2008 Distributing the Kalman filter for large-scale systems *IEEE Trans. Signal Processing* **56** 4919–35

[25] Hopfield J J 1984 Neurons with graded response have collective, computational properties like those of two-state neuron *PNAS* **81** 3088–92

[26] Dozat T 2016 Incorporating Nesterov momentum into Adam, *Proc. of 4th Int. Conf. on Learning Representations, Workshop Track* No. 107

[27] Van Essen D C, Smith S M, Barch D M, Behrens T E J, Yacoub E and Ugurbil K 2013 The WU-Minn human connectome project: an overview *NeuroImage* **80** 62–79

[28] Schaefer A, Kong R, Gordon E M, Laumann T O, Zuo X-N, Holmes A J, Eickhoff S B and Yeo B T T 2017 Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI *Cerebral Cortex* **28** 3095–114

[29] Siegel J S, Mitra A, Laumann T O, Seitzman B A, Raichle M, Corbetta M and Snyder A Z 2017 Data quality influences observed links between functional connectivity and behavior *Cerebral Cortex* **27** 4492–502

[30] Singh M F, Braver T S, Cole M W and Ching S 2020 Estimation and validation of individualized dynamic brain models with resting state fMRI *NeuroImage* 117046

[31] Frässle S, Lomakina E I, Kasper L, Manjaly Z M, Leff A, Pruessmann K P, Buhmann J M and Stephan K E 2018 A generative model of whole-brain effective connectivity *NeuroImage* **179** 505–29

[32] Fish D A, Brinicombe A M, Pike E R and Walker J G 1995 Blind deconvolution by means of the Richardson–Lucy algorithm *J. Opt. Soc. Am.* A **12** 58–65

[33] Bell A J and Sejnowski T J 2008 An information-maximization approach to blind separation and blind deconvolution *Neural Comput.* **7** 1129–59

[34] Park I J, Bobkov Y V, Ache B W and Principe J C 2013 Quantifying bursting neuron activity from calcium signals using blind deconvolution *J. Neurosci. Methods* **218** 196–205

[35] Ekanadham C, Tranchina D and Simoncelli E P 2011 A blind sparse deconvolution method for neural spike identification *Adv. Neural Inform. Process. Syst.* **24** 1440–8