Efficient identification for modeling high-dimensional brain dynamics

Matthew F. Singh, Chong Wang, Michael W. Cole, and ShiNung Ching

***Abstract*— System identification poses a significant bottleneck to characterizing and controlling complex systems. This challenge is greatest when both the system states and parameters are not directly accessible, leading to a dual-estimation problem. Current approaches to such problems are limited in their ability to scale with many-parameter systems, as often occurs in networks. In the current work, we present a new, computationally efficient approach to treat large dual-estimation problems. In this work, we derive analytic back-propagated gradients for the Prediction Error Method which enables efficient and accurate identification of large systems. The PEM approach consists of directly integrating state estimation into a dual-optimization objective, leaving a differentiable cost/error function only in terms of the unknown system parameters, which we solve using numerical gradient/Hessian methods. Intuitively, this approach consists of solving for the parameters that generate the most accurate state estimator (Extended/Cubature Kalman Filter). We demonstrate that this approach is at least as accurate in state and parameter estimation as joint Kalman Filters (Extended/Unscented/Cubature) and Expectation-Maximization, despite lower complexity. We demonstrate the utility of our approach by inverting anatomically-detailed individualized brain models from human magnetoencephalography (MEG) data.**

I. Introduction

Control of complex systems benefits greatly from knowledge of the underlying system and the evolution of its states [1], typically in the form of a dynamical systems model. However, in many real-world examples, obtaining such a model is challenging. Even in situations where a general mathematical form of the underlying dynamics is postulated, a number of unknown parameters typically require specification. The identification of these parameters is often complicated because measurements are opaquely transformed from state variables and/or obfuscated by noise. Substantial progress in system identification research has been made treating these confounds. A wide variety of techniques now enable model-parameterization with well-measured state variables. Conversely, Bayesian methods (e.g., Kalman [3] and Particle filtering [2]) are now well-established for estimating (latent) state variables from measurements with a known system model (state estimation). However, the estimation of both states and parameters (dual-estimation) remains an unmet challenge, especially for large scale problems relevant to applications in the control of biological and/or network-systems

In this work, we are interested in an emerging application of dual estimation for system identification: neurocontrol engineering, i.e., modeling and control of brain networks. Such an application brings forth a number of technical and conceptual challenges. The brain is commonly modeled as containing hundreds of distinct regions (state variables) with thousands of interconnections and heterogeneous local circuitry (unknown parameters). Modern solutions for dual estimation, such as joint-Filtering [4]–[6] and Expectation Maximization [7], perform well for low-dimensional systems, but become computationally cost-prohibitive for systems with a large number of states and unknown parameters. In systems involving many interacting states (e.g., networks), the total number of parameters typically scales quadratically with the state-variables ($\mathcal{O}(n^2)$ parameters), resulting in joint-Filter complexities of $\mathcal{O}(n^4)$ to $\mathcal{O}(n^6)$ in terms of state variables (depending upon how covariances are stored).

Furthermore, identification is limited because it is difficult to causally interact with the brain through exogenous inputs while simultaneously recording activity. This renders traditional active system identification protocols difficult, meaning identification has to proceed from passive output-only measurements. Compounding this challenge is that the activity of brain regions cannot be directly measured in-vivo, due to volume conduction and the fact that noninvasive measurements are taken at far-field (the scalp) which results in signal-mixing. Due to their geometry, many important classes of brain cell (neurons) generate no far-field signals. These features form a bottleneck for modeling and control design as identification entails a high-dimensional dual-estimation problem. However, obviating these challenges would provide significant leverage for analysis and design, since having large-scale brain models may provide new insights into the generate mechanisms underlying overt measurements, and further, suggest new ways of imparting control over the identified dynamics using acute (e.g., brain stimulation) or systemic (e.g. pharmacological) perturbations.

Here, we address the above considerations by building on the Prediction Error Method [8]–[10] for nonlinear system identification and gradient-based training of EKF-coupled neural-networks [11]. In this framework, derivatives with respect to parameter are calculated alongside state-estimation

Corresponding Author: MS is with the Departments of Psychological & Brain Sciences, and Electrical and Systems Engineering at Washington University in St. Louis, USA and the Center for Molecular and Behavioral Neuroscience at Rutgers University, Newark USA. f.singh@wustl.edu

CW is with the Departments of Psychological & Brain Sciences, Washington University in St. Louis, USA. chong.wang@wustl

MC is with the Center for Molecular and Behavioral Neuroscience at Rutgers University, Newark USA, michael.cole@rutgers.edu

SC is with the Department of Electrical and Systems Engineering and Biomedical Engineering, at Washington University in St. Louis, USA, shinung@wustl.edu

MS was funded by NSF-DGE-1143954 from the US National Science Foundation and NIH T32 DA007261-29 from the National Institute on Drug Addiction. Portions of this work were supported by NSF 1653589 and NSF 1835209, from the US National Science Foundation and NIMH Administrative Supplement MH066078-15S1.

to form a coupled filter. While especially accurate (see Sec.III), this method ([10]) has complexity $\mathcal{O}(n^5)$ for large networks which becomes cost-prohibitive. Our primary contribution is to generalize and reformulate the "EKF-output error method" proposed by [10] to enable efficient identification of large-scale network process-models. Our key formulation is to conceptualize the filter-optimization of [10] as analogous to training an artificial neural network, with each 'layer' corresponding to one iteration of the Kalman Filter [11]. We then use the methods typical for training neural networks by deriving the analytic backpropagated gradients and using deep-learning style update rules (gradient clipping [12] and NADAM [13]). Our algorithm reduces complexity from $\mathcal{O}(n^5)$ to $\mathcal{O}(n^3)$ which is dramatic with $n > 100$ and enables us to directly fit biological brain models to human MEG recording. Such model fitting would be computationally intractable under contemporary approaches. We also present exploratory analyses which demonstrate the potential of these models to identify directions for future neuro-control design.

II. Formulation and Methods

A. Dual Estimation Problem

We address the problem of estimating parameters for high-dimensional nonlinear systems in the presence of imperfect state measurement. We consider discrete-time system of state-variables $x_t \in \mathbb{R}^n$ evolving according to the nonlinear dynamics $f_{t+1}(x_t, \theta)$ with process noise $w_t \sim \mathcal{N}(0, Q_t)$. The noise process w_t is independently realized in time (no autocorrelation) and independent of states/measurements. The vector field f is characterized by a set of unknown parameters θ and f is allowed to vary in time, hence known inputs are absorbed in f:

$$x_t = f_t(x_{t-1}, \theta) + w_t \tag{1}$$
$$y_t = H_t x_t + v_t. \tag{2}$$

Here, $y_t \in \mathbb{R}^p$ represents measurements produced by a linear transformation H_t of the state-variables x_t and measurement noise $v_t \sim \mathcal{N}(0, R_t)$ (with the same independence assumptions as process noise). Our task is to estimate the system parameters θ_t and states x_t given knowledge of H_t, Q_t, R_t and the general functional form f_t. Since the parameter θ operates within the latent state-space of x, we first discuss state-estimation.

B. Bayesian Filtering

For a system with known parameterization, the problem of estimating (continuous-valued) states from measurement has been primarily treated using Bayesian Filtering. Common frameworks include the Particle Filter [14], which represents arbitrary distributions via particle-sampling, and the Kalman Filter [3], which classically assumes that all distributions are Gaussian and provides a closed-form recursion for estimating the conditional mean/covariance of states given measurements. While the original Kalman Filter is optimal (in least-squares) for the linear-Gaussian case, multiple nonlinear variants of the Kalman Filter have been proposed to estimate statistics under nonlinear transformations. Generally, the Kalman-based approach evolves estimates of the conditional state mean x_t and covariance P_t given measurements y_t according to:

$$X_{t-1} \sim \mathcal{N}(\hat{x}_{t-1}, P_{t-1}) \tag{3}$$
$$\hat{x}_{t|t-1} = \mathbb{E}[f_t(X_{t-1})] \tag{4}$$
$$P_{t|t-1} = Q_t + cov[f_t(X_{t-1})] \tag{5}$$
$$z_t = y_t - H_t \hat{x}_{t|t-1} \tag{6}$$
$$S_t = H_t P_{t|t-1} H_t^T + R_t, \;\; K_t = P_{t|t-1} H_t^T S_t^{-1} \tag{7}$$
$$\hat{x}_t = \hat{x}_{t|t-1} + K_t z_t \tag{8}$$
$$G_t = I - K_t H_t; \quad P_t = G_t P_{t|t-1} \tag{9}$$

The mean and covariance after the nonlinear transformation f are most commonly estimated using either first-order linearization or sample-based statistics. First-order linearization results in the Extended Kalman Filter [15]):

$$\mathbb{E}[f_t(X)] \approx f_t(\hat{x}_{t-1}); \; cov[f_t(X)] \approx F_t' P_{t-1} F_t'^T \tag{10}$$

with F_t' denoting the Jacobian of f_t which (unless otherwise indicated) is evaluated at x_{t-1}. By contrast, the Unscented [5] and Cubature [6] Kalman Filters represent the prior distribution via a basis $\{\beta^{(i)}\}$ and weighting $\{\kappa^{(i)}\}$ with

$$\chi_t^{(i)} := \hat{x}_{t-1} + \sqrt{P_{t-1}}\beta^{(i)}; \;\; \mathbb{E}[f_t] \approx \sum_i f(\chi_t^{(i)})\kappa_i \tag{11}$$

$$cov[f] \approx \sum_i (f(\chi_t^{(i)}) - \hat{x}_{t|t-1})(f(\chi_t^{(i)}) - \hat{x}_{t|t-1})^T \kappa_i \tag{12}$$

with $\sqrt{P}$ denoting the Cholesky factor of P. The primary distinction between EKF and UKF/CKF is that the former is generally faster to compute, but its accuracy suffers for systems that are not amenable to linearization and it requires differentiability of f. The Unscented [5] and Cubature [6] Kalman Filters only differ in the choice of basis (β) and weights (κ) used to approximate the prior distribution.

C. Parameter Estimation with Latent States

We now consider the 'dual' estimation problem: estimating states and parameters ('dual' in this context simply refers to the pairing of states and parameters, not Lagrange Duality). This problem has been conventionally treating by either augmenting the state space to include parameters with stationary dynamics (i.e. $\hat{\theta}_t = \hat{\theta}_{t-1} + \eta_t$) as in the joint Kalman/Particle Filters, or by alternating between state and parameter estimations as is done in Expectation Maximization [7]. As previously mentioned, these approaches scale poorly to high-dimensional networks. By contrast, the Prediction Error Method (PEM) [8]–[10] takes an alternate path: directly reducing the problem to parameter-estimation which is then optimized using gradients [10]. This method (like other Prediction Error Methods) solves for parameters that produce the most accurate state-estimator as indicated by predicting future measurements. Gradients are computed with respect to parameters for the entire filter/prediction process [8]–[10]. This approach differs from EM in that the former considers state-estimates to be a deterministic function of parameters and measurements (i.e. $x_t = Filter(x_0, P_0, \theta, \mathcal{F}_{t-1}[y])$) in which $\mathcal{F}[y]$ denotes the filtration induced by previous measurements. As a result, differentiating this relationship directly accounts for how changes in parameter affect state-estimates, whereas EM deals with these problems separately.

Thus, PEM treats the dual-estimation of states and parameters using a min-min strategy. Put simply, one solves for parameters that generate the best state-estimator/predictor.

This work has recently been advanced by Sjanic and Skoglund's [10] derivation of analytic gradients for the case of an EKF state-estimator using forward recursion ('EKF output-error approach'). This advance is significant and useful in many identification-for-control scenarios, as it reformulates the gradient estimation by online tracking of the state and covariance Jacobians with respect to parameters. This forward-accumulation of gradients enables online estimation in the form of an additional filter. However, this property also prevents application to problems featuring a very large numbers of parameters as the forward accumulation has complexity $\mathcal{O}(n_x^3 n_\theta)$ for state and parameter dimension, respectively. Thus, while this approach is effective for identification of many systems, it remains computationally arduous for large networks (Sec. III) as the quadratic scaling of parameters with state leads to complexity $\mathcal{O}(n^5)$.

D. Backpropagated Kalman Filtering

We propose to overcome these scaling limitations by instead deriving the analytic back-propagated gradients for the general case of Kalman-Filter (i.e. EKF, UKF, CKF etc.) Whereas the forward accumulation of gradients keeps track of the Jacobians ($\partial P/\partial\theta$, $\partial\hat{x}/\partial\theta$) and is computed simultaneously with state-estimation, back-propagation first evaluates the state-estimation/error function and then calculates gradients recursively through time which avoids accumulating Jacobians with respect to θ. Although analytically equivalent to Sjanic and Skoguland's approach for the EKF case, our algorithm achieves $\mathcal{O}(n^3)$ complexity which powers our application to high-dimensional brain networks that would be otherwise intractable.

We note that this concept has been previously proposed for training convolutional neural networks by Haarnoja and colleagues [11] with great success. However, whereas their approach involved training a neural-network-based Extended Kalman-Filter end-to-end using automatic differentiation, we instead seek to solve grey-box system-identification problems and present the analytic gradients for a generalized definition of the Kalman Filter and arbitrary process models.

Our contributions to the existing PEM frameworks are: 1) we derive the backpropagated analytic gradients of the Kalman Filter; 2) we extend the analytic PEM method to a general class of Kalman Filter (rather than just EKF); 3) we treat cases without initial state/covariance priors. These changes are meant to combat the three pitfalls encountered with nonlinear networks: high-dimensionality; nonlinearity/non-stationarity; and a lack of available prior estimates. Our objective is to minimize prediction-error with respect to a quadratic objective. For an initialization time t and filter-length m, we denote the backwards accumulation of error as

$$\overleftarrow{\mathcal{E}}_t^{(t_0)} := \sum_{k=t_0+t}^{t_0+m} (y_k - H\hat{x}_{k|k-1}^{\{\theta,\mathcal{F}[y]\}})^T M_k (y_k - H\hat{x}_{k|k-1}^{\{\theta,\mathcal{F}[y]\}}) \quad (13)$$

With the objective being to solve for θ which minimizes the total error over all initialization times $\mathbb{E}_{t_0}[\overleftarrow{\mathcal{E}}_1^{(t_0)}]$. From here on, we will omit the start-time notation for $\mathcal{E}$ and the dependency of $\hat{x}$ on θ, $\mathcal{F}[y]$ for brevity's sake. We now derive the backpropagated error-gradients.

$$\nabla\overleftarrow{\mathcal{E}}_t = 2(\partial_\omega z_t)H^T M z_t + \frac{\partial\overleftarrow{\mathcal{E}}_{t+1}}{\partial\hat{x}_t}(\partial_\omega\hat{x}_t) + \frac{\overleftarrow{\mathcal{E}}_{t+1}}{\partial P_t}(\partial_\omega P_t) \quad (14)$$

Evaluating the first term: $\partial_\omega z_t = -H\partial_\omega\hat{x}_{t|t-1}$

$$\partial_\omega\hat{x}_{t+1} = G_{t+1}(\partial_\omega\hat{x}_{t+1|t}) + (\partial_\omega K)z_{t+1} \quad (15)$$

$$\partial_\omega P_{t+1} = G_{t+1}(\partial_\omega P_{t+1|t}) - (\partial_\omega K)HP_{t+1|t} \quad (16)$$

Since H, Q, R are fixed, the Kalman gain is a direct function of $P_{t|t-1}$ (and its influence on S). Using that K_t minimizes $Tr[P_t]$ and the implicit function theorem:

$$-\partial P_{t|t-1}H^T + \partial K_t S + K_t H\partial P_{t|t-1}H^T = 0 \quad (17)$$

$$\partial_\omega K_t = G_t(\partial_\omega P_{t|t-1})H^T S^{-1} \quad (18)$$

$$\partial_\omega\hat{x}_t = G_t[\partial_\omega\hat{x}_{t|t-1} + (\partial_\omega P_{t|t-1})H^T S^{-1} z_t] \quad (19)$$

$$\frac{\partial\overleftarrow{\mathcal{E}}_t}{\partial\hat{x}_{t|t-1}} = G_t^T\frac{\partial\overleftarrow{\mathcal{E}}_t}{\partial\hat{x}_t} - 2H^T M_t z_t \quad (20)$$

$$\partial_\omega P_t = G_t(\partial_\omega P_{t|t-1})G_t^T \quad (21)$$

$$\mathcal{Z}_t := H^T S^{-1} z_t\left[G_t\frac{\partial\overleftarrow{\mathcal{E}}_{t+1}}{\partial\hat{x}_t}\right]^T \quad (22)$$

$$\frac{\partial\overleftarrow{\mathcal{E}}_t}{\partial P_{t|t-1}} = \frac{1}{2}(\mathcal{Z}_t + \mathcal{Z}_t^T) + G_t^T\frac{\partial\overleftarrow{\mathcal{E}}_t}{\partial P_t}G_t \quad (23)$$

Since Kalman-Filter approximations of $\mathbb{E}$ preserve linearity and cov preserve bilinearity, we have, for $\phi := \{\hat{x}_t, \theta_t\}$:

$$\partial_\phi\overleftarrow{\mathcal{E}}_t = \frac{\partial\overleftarrow{\mathcal{E}}_t}{\partial\hat{x}_{t|t-1}}\mathbb{E}[\partial_\phi f] + \frac{\partial\overleftarrow{\mathcal{E}}_t}{\partial P_{t|t-1}}[cov[f,\partial_\phi f] + cov[\partial_\phi f, f]] \quad (24)$$

In particular, for the EKF, we have:

$$\partial_{P_{t-1}}\overleftarrow{\mathcal{E}}_t = F_t'^T\frac{\partial\overleftarrow{\mathcal{E}}_t}{\partial P_{t|t-1}}F_t' \quad (25)$$

$$\frac{\partial\overleftarrow{\mathcal{E}}_t}{\partial\hat{x}_t} = F'^T\frac{\partial\overleftarrow{\mathcal{E}}_t}{\partial\hat{x}_{t|t-1}} + 2\frac{\partial vec[F']}{\partial\hat{x}_t}^T vec\left[\frac{\partial\overleftarrow{\mathcal{E}}_t}{\partial P_{t|t-1}}F'P_t\right] \quad (26)$$

$$\frac{\partial\mathcal{E}}{\partial\theta} = \sum_t\frac{\partial f_t^T}{\partial\theta}\frac{\partial\overleftarrow{\mathcal{E}}_t}{\partial\hat{x}_{t|t-1}} + 2\frac{\partial vec[F']}{\partial\theta}^T vec\left[\frac{\partial\overleftarrow{\mathcal{E}}_t}{\partial P_{t|t-1}}F'P_t\right] \quad (27)$$

For sample-based (e.g. Unscented and Cubature) Kalman Filters, we denote the sampling basis β and the weights for each point as κ_i. The sampled-points are: $\chi_t := \hat{x}_t + \sqrt{P_t}\beta$.

$$\frac{\partial\overleftarrow{\mathcal{E}}_t}{\partial\chi_t^{(i)}} = 2\kappa_i F_{\chi^{(i)}}'^T\frac{\partial\overleftarrow{\mathcal{E}}_t}{\partial P_{t|t-1}}[f(\chi^i) - \hat{x}_{t|t-1}] \quad (28)$$

$$\frac{\partial\mathcal{E}}{\partial\theta} = 2\sum_{t,i}\kappa_i\frac{\partial f(\chi_t^i)^T}{\partial\theta}\frac{\partial\overleftarrow{\mathcal{E}}_t}{\partial P_{t|t-1}}[f(\chi_t^i) - \hat{x}_{t|t-1}] \quad (29)$$

$$\mathcal{Q}_t := \Phi\odot[\sqrt{P_t}\frac{\partial\overleftarrow{\mathcal{E}}_t}{\partial\chi_t^{(i)}}\beta^T];\ \ \Phi_{i,j} := \begin{cases}1/2 & i=j\\ 1 & i>j\\ 0 & i<j\end{cases} \quad (30)$$

$$\frac{\partial\overleftarrow{\mathcal{E}}_t}{\partial P_{t-1}} = \sqrt{P_{t-1}^{-1}}^T(\mathcal{Q}_t + \mathcal{Q}_t^T)\sqrt{P_{t-1}^{-1}} \quad (31)$$

$$\frac{\partial\overleftarrow{\mathcal{E}}_t}{\partial\hat{x}_{t-1}} = \sum_i\frac{\partial\overleftarrow{\mathcal{E}}_t}{\partial\chi_t^{(i)}} \quad (32)$$

with $\odot$ denoting the Hadamard product.

Thus, by deriving the analytic gradients, the error function is decomposed into a direct function of parameters, measurements, and the initial state/error-covariance estimates. In practice, estimates of the initial states and their error covariances are not typically available but, given knowledge of the state distribution's covariance ($cov[X]$ not to be confused with P_t), they can be estimated via a linear filter (L). In order to estimate this covariance, we simulate the model (with process noise) according to current parameter estimates and calculated the covariance (denoted "C"). The

resultant estimates are:

$$L := CH^T(HCH^T + R)^{-1}, \hat{x}_0 = Ly_0, P_0 = (I - LH)C \quad (33)$$

The estimated state-covariance is thus a function of parameters and we analogously back-propagated error-gradients through the initialization and simulation steps. In later analyses, we compare the efficacy of adding this initialization step to the PEM approach, as opposed to guessing initial values.

E. Application to Large Networks

Our approach is most useful for systems in which the number of unknown parameters scales nonlinearly with the number of state variables as occurs in networks/circuits. For benchmarking, we tested algorithms in parameterizing systems represented as recurrent nonlinear-networks. As universal approximators, randomly parameterized nonlinear networks generate a wide range of dynamical phenomena and are thus provide a useful basis for benchmarking. As states are unknown, even this case generates a highly nonlinear $\mathcal{E}$ with respect to θ, while enabling analytic M-steps for comparison with EM [7], [19]. We consider generic recurrent neural network models ([20]), of the form:

$$x_{t+1} = \theta\psi(x_t) + Dx_t + c + w_t \quad (34)$$

with measurements generated by $y_t = Hx_t + v_t$ and the unknown connection matrix $\theta \in \mathbb{R}^n$. In all cases, the vector c and the positive-diagonal matrix D are known and $\psi = tanh$ (applied element-wise). We compared run-time across different network sizes but maintained that there were twice as many state variables as measurement channels (i.e. $y \in \mathbb{R}^{15}$ for $x \in \mathbb{R}^{30}$). The diagonal entries of D were distributed $N(.8, (1/50)^2)$ and entries of c were distributed $N(0, (1/20)^2)$. H was generated by creating a matrix with normal iid. entries and redistributing the singular values by $|N(2, .25)|$ which gave more variety to the spectral distribution between simulations. We considered two values for Q: $I/4$ (low-noise) and I (high-noise) to test the possible interaction of performance with process noise-level. We randomly generated $\sqrt{R}$ according to $N(0, (1/5)^2)$. The connection matrix θ was defined by first generating a base matrix: $\theta_0 \sim N(0,1)/20 + N(0,1)^3/20$. Thus, θ_0 was the sum of terms drawn from a normal and normal-cubed distribution which better reflects the heavy-tailed connection statistics of most networks. The connectivity matrix θ was defined as $\theta = \theta_0 - (\theta_0^T/2)$ for off-diagonal elements and 1 for diagonals to ensure non-trivial dynamics (i.e. promoting complex eigenvalues). Half of off-diagonal entries were then set equal to zero which was treated as a known prior on network structure, resulting in $n(n+1)/2$ non-zero (unknown) parameters.

III. Implementation and Results

We view the primary contribution of our method to be enabling dual state-parameter estimation in very large systems. However, we also tested whether the proposed technique is beneficial within the (relatively) lower-dimensional domains that are applicable to existing methods: joint Filters (Extended, Unscented, and Cubature), Expectation Maximization with cubature smoothing [19], and PEM with forward-propagated gradients ("EKF output error method"). We confirmed that parameters estimated with forward gradients were essentially identical to those using back-propagated-EKF (exactly so when using the same random-number-generator). Joint filters used 60,000 iterations with parameter noise-variance of .0001, which was decreased by 2% every 250 intervals. EM used either the full timeseries (120,000 points; 300 reps) or random batches of 5,000 per interval (600 reps) which we found was more accurate, stable, and efficient (due to failures, performance for EM-Full is not shown). We used 24 randomly-generated 30-node networks for comparing accuracy as well as several simulations for 10 through 60 nodes to benchmark run-time. For 40+ node networks, the runtime was extrapolated from several iterations of each alternative algorithm Gradient-based optimization was performed using the NADAM algorithm [13] with rate .00025 and memory parameters .98, and .95 for the gradients and their squares, respectively. Prior to NADAM, we performed gradient-clipping [12] with the threshold set to twice the mean-square gradient over the first 200 minibatches. For comparison with joint Kalman Filters, each minibatch used a single randomly chosen start-time and evaluated Kalman prediction-errors over the subsequent 20 steps. All initial state estimates were distributed $\mathcal{N}(0, .05^2)$. Analyses were run single-core on Intel Xeon Gold 6226R 2.9GHz.

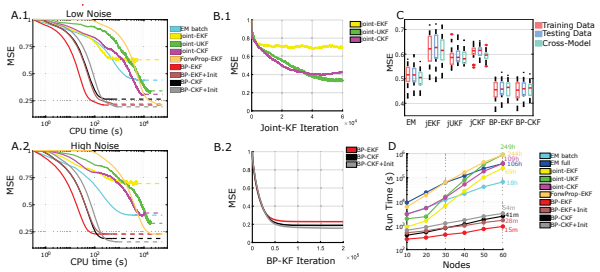

Fig. 1. The proposed technique is accurate and efficient. A) The method quickly converges to a competitively accurate estimate of connectivity for a 30-node network (average over 24 networks). Benefits of using BP-CKF over BP-EKF depend upon the amount of noise (low: A.1, high: A.2) and whether initial statistics are also estimated (i.e. BP-CKF+Init). B) Convergence of joint-KF (B.1) and BP-KF (B.2) in terms of iteration. C) BP-KF generates accurate state-estimates and experiences little loss between training/testing data and cross-validation to new measurement models (new H,R). D) The proposed method scales well to larger network sizes. EM=Expectation Maximiation, jEKF/UKF/CKF= joint Extended/Unscented/Cubature Kalman Filters, BP-E/C=back-propagation technique using EKF/CKF for state-estimation. ForwProp+EKF=using forward-propagated gradients ([10]) Dashed-line=algorithm finished.

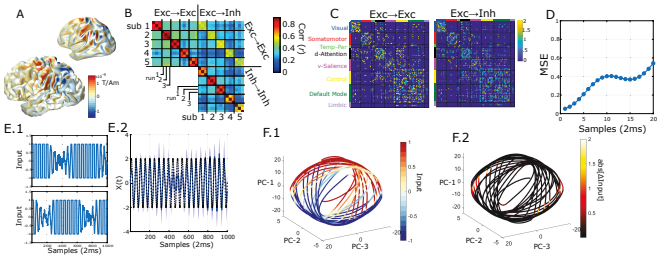

Fig. 2. Application to human brain models. A) Leadfield for a single gradiometer (bottom=true, top="inflated" surface). B) Estimated brain-connectivity matrices are reliable and individualized. C) Group-average brain-connectivity for excitatory (left) and inhibitory(right) targets. D) Group-average deterministic-forecasting (prediction) performance through 20-steps (40ms) into the future. E) Examples of open-loop control solutions (E.1) for tracking a 10.5Hz ("alpha-wave") reference signal (E.2). Shading indicates standard-deviation of the state-distribution. F) Phase-space projections illustrate that the identified control switches brain activity between two sets (E.1) corresponding to the reference signal's peak/trough. E.2) Phase-space projection of switching regions (boundaries between modes).

case (Fig. 1A.2) BP-CKF always performed better. We also note that, whereas joint Kalman Filters still exhibited some instability in parameter estimates (Fig. 1B.1), despite annealing the parameter-variance (their "process-noise"), gradient based-approaches converged smoothly (Fig. 1B.2)

We also tested the accuracy in predicting latent-state values from measurement data. We performed this comparison at three levels: using the training data, new testing data from the same system, or new testing data in which the measurement model (H,R) had changed from that used during training. In the last case, state-estimators (Kalman Filters) combined the new measurement model with state-space models parameterized according to the original data. Across conditions, we found that results favored our approach (Fig. 1C) and there was relatively little loss in accuracy even when applied to the new measurement model.

Results also indicate that the proposed technique is highly scalable (Fig. 1D). In terms of computational complexity, our approach inherits the complexity of the underlying state estimator (E/CKF) but is not significantly affected by the number of parameters. By contrast, joint-filtering approaches scale nonlinearly with both the number of state variables and the number of unknown parameters (itself a quadratic function of network size). We found that our approach was two orders-of-magnitude faster than jEKF/jUKF/jCKF in performing dual estimation for 60-node networks. At all scales, results favored our back-propagated variant of the prediction-error method over the analogous (forward-propagated) "EKF-output Error Method" ([10]) as indicated by drastically reduced run-time for large n.

A. Identification and Control of Human Brain Dynamics

The primary motivation for our methodology is to enable systems identification and control of human brain dynamics for single subjects. We demonstrate our method?s capability in inferring parameters of a high-dimensional brain network model from magnetoencelephagraphy (MEG) recordings and present exploratory analyses indicating the potential of these models to inform neurocontrol design.

The human brain consists of a large number of distinct regions containing neurons (brain cells). Neurons emit electrical impulses at a rate which depends nonlinearly upon their current voltage. The outer-shell (membrane) of each neuron forms an RC circuit with the outside fluid leading to a gradual discharge of accumulated inputs. Each brain area contains two cell types: excitatory (p_t) and inhibitory (r_t). Excitatory cells send positive currents throughout the brain and are oriented normal to the brain surface. Their activity results in electromagnetic dipoles which (after much mixing and volume conduction) are measurable from the scalp. Inhibitory cells decrease the membrane potential of nearby neurons (suppress activity) and are spherical so they do not generate measurable dipoles. They are, however, absolutely essential in modeling brain oscillations. We modeled 100 brain areas [23] each with an excitatory ($p_t \in \mathbb{R}^{100}$) and inhibitory (r_t) population

$$p_{t+1} = W_p\psi_p(p_t) - \beta_p \odot \psi_r(r_t) + \tau_p \odot p_t + w_t^p \quad (35)$$

$$r_{t+1} = W_r\psi_p(p_t) - \beta_r \odot \psi_r(r_t) + \tau_r \odot r_t + w_t^r \quad (36)$$

The non-negative 100×100 matrices W_p, W_r describe excitatory connections throughout the brain (leading to excitatory and inhibitory cells, respectively). Admissible connections were pre-specified at group-level by thresholding fMRI models [16]–[18]. The local inhibitory connection strengths are described by the non-negative vectors $\beta_p\beta_r$ and the discretized time-constants are positive vectors τ_p, τ_r. The baseline input to each brain area is denoted by the vector c_p, c_r. The relative output of each brain region is modeled as the parameterized sigmoid $\psi_p(p_t) = tanh(s_p \odot p_t + v_p)$ with scaling factor $s_p = .25$ and the threshold vector v_p unknown. We similarly modeled ψ_r with $s_r = .375$. We assumed that the process noise w_t^p and w_t^r was iid between brain areas and with variance=1/4. Unknown parameters thus correspond to $W_{p,r}$,$\beta_{p.r}$, and $v_{p,r}$. The sparsity of W was roughly 18% which resulted in 4040 unknown parameters.

We used five subjects' 248-channel publicly-available MEG data from the Human Connectome Project [21] which we downsampled to 500 Hz. Each subject contributed three separate "resting-state" scans lasting 5 minutes each in a single testing session. We used MRI and single-shell boundary element method [22] to estimate the magnetic fields due to dipoles at each brain location (oriented normal the surface and summed within regions) which were condensed into the "lead-field" matrix L describing the relative gain between brain areas and channels (e.g. Fig. 2A) which we rank-reduced to have a maximal matrix-condition number of 100 (resulting in 70-80 measurement dimensions per scan). Thus the measurement model was:$y_t = [L \quad 0]\begin{bmatrix} p_t \\ r_t \end{bmatrix} + v_t$. The covariance (R) for measurement noise v_t was directly estimated from empty-room measurements. We fit models using the back-propagated EKF with 150 random initial conditions per minibatch and 50,000 minibatches which each featured 20-steps of the prediction-error-method and 20 steps of pure forecasting (no Kalman Filtering).

Identified brain models were reliable (Fig. 2B) with

very similar parameters (W_p, W_r) estimated from different scans of the same subject (pairwise-mean $r = .71 \pm .06$, $r = .60 \pm .11$, respectively). By contrast, parameter estimates had far less similarity between subjects ($r = .35 \pm .04$, $r = .24 \pm .04$) indicating the ability to identify individual brain circuitry, as opposed to features which are common among all humans. Group-average brain-connectivity matrices are reported in Fig. 2C. We also note that fitted models were accurate in forecasting future measurements in cross-validated data over the 20-step horizon used in training (Fig. 2D).

We next considered leveraging the identified models to perform exploratory offline analysis for noninvasive neurocontrol (transcranial electrical stimulation/tES). As these analyses are the first of their kind, our objective is to use offline-simulations to discover which control techniques may be particularly promising for future study. A common aim of neurostimulation is increasing the prevalence of phase-locked oscillations in a particular brain area (e.g. prefrontal cortex) and frequency (e.g. "alpha-band":8-12Hz). Our control objective was inducing a target region (parcel 32) to follow a 10.5Hz reference signal. Formally, we considered a stochastic target-tracking control problem with quadratic cost function for errors. We bounded inputs to ± 1 but did not penalize the control energy. For preliminary analyses, we therefore considered open-loop finite-horizon (1000 steps=2s) control separately optimized for different initial conditions (analogous to one prediction-phase of model-predictive control). We directly optimized the expected error (separately for each initial condition) using quasi-Newton methods and single-shooting with distributions approximated by 100 particles (reset each iteration). These methods were chosen to explore potential input strategies with minimal pre-specification; real-world applications will need to be much more efficient. Interestingly, the identified open-loop controls for most subjects, consistently approximated a bang-bang style solution for brief periods (Fig. 2E.1) with intermittent periods of more complex dynamics. This solution was highly accurate in tracking the reference signal (Fig. 2E.2). Further examination suggested that this numerically-identified control was approximating a sliding-mode style solution in which the system switched between two manifolds (F.1) by alternating between +1 and -1 where these manifolds touched (F.2). These analyses and interpretations are, of course, exploratory and should not take the place of rigorous analyses to identify control features (e.g. sliding modes). They do, however, demonstrate the potential of system-identification to identify directions for future, rigorous control analyses and experimentation.

REFERENCES

[1] K. S. Narendra and K. Partbsarathy, "Identification and control of dynamical systems using neural networks," IEEE Trans. Neural Networks, vol 1, no. 1, pp. 4-27, 1990.

[2] T. Schön and F. Gustafsson and P. Nordlund, "Marginalized particle filters for mixed linear/nonlinear state-space models," IEEE Transactions on Signal Processing, vol. 53, no. 7, pp. 2279-2289, 2005.

[3] R. E. Kalman, "A new approach to linear filtering and prediction problems," Trans. ASME J. Basic Eng., vol. 82, pp. 34-45, Mar. 1960.

[4] S. Singhal and L. Wu, "Training Multilayer Perceptrons with the Extended Kalman Algorithm," NIPS, vol. 1, pp. 133-140, 1988.

[5] R. Van der Merwe and E. Wan, "The square-root unscented Kalman filter for state and parameter-estimation," in Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP), vol. 6, 2001, pp. 3461-3464.

[6] I. Arasaratnam and S. Haykin, "Cubature Kalman Filters," IEEE Transactions on Automatic Control, vol. 54, 6, pp. 1254-1269, 2009.

[7] A. P. Dempster and N. M. Laird and D. B. Rubin, D. B. " Maximum likelihood from incomplete data via the EM algorithm," Journal of the Royal Statistical Society: Series B, vol. 39, no. 1, pp. 1-22, 1977.

[8] R. Larsson, Z. Sjanic, M. Enqvist, and L. Ljung, "Direct prediction-error identification of unstable nonlinear systems applied to flight test data," IFAC Proceedings Volumes, vol. 42, no. 10, pp. 144-149, 2009.

[9] M. Kok and T. B. Schön, "Maximum likelihood calibration of a magnetometer using inertial sensors," IFAC Proceedings Volumes, vol. 47, no. 3, pp. 92-97, 2014.

[10] Z. Sjanic and M. A. Skoglund, "Prediction error method estimation for Simultaneous Localisation and Mapping," in 2016 19th International Conference on Information Fusion (FUSION), pp. 927-934, 2016.

[11] T. Haarnoja and A. Ajay and S. Levine and P. Abbeel, "Backprop kf: Learning discriminative deterministic state estimators," Advances in neural information processing systems, 29, 2016.

[12] Razvan Pascanu, Tomas Mikolov, Yoshua Bengio. "On the difficulty of training recurrent neural networks." International conference on machine learning. PMLR, 2013.

[13] T. Dozat, "Incorporating Nesterov momentum into Adam," Proc. of 4th International Conference on Learning Representations, 2016.

[14] J. Kokkala and A. Solin and S. Särkkä. "Expectation maximization based parameter estimation by sigma-point and particle smoothing," 17th International Conference on Information Fusion, pp. 1-8, 2014.

[15] S. J. Julier and J. K. Uhlmann "New extension of the Kalman filter to nonlinear systems", Proc. SPIE 3068, Signal Processing, Sensor Fusion, and Target Recognition VI, 1997, pp. 182-193.

[16] M. F. Singh and T. S. Braver and M. W. Cole and S. Ching. "Estimation and validation of individualized dynamic brain models with resting state fMRI," NeuroImage, 221, 117406, doi:10.1016/j.neuroimage.2020.117046, 2020.

[17] M. F. Singh and A. Wang and T. S. Braver and S. Ching. "Scalable surrogate deconvolution for identification of partially-observable systems and brain modeling," Journal of Neural Engineering, 17, 046025, doi:10.1088/1741-2552/aba07d, 2020.

[18] M. F. Singh and A. Wang and M. Cole and S. Ching and T. S. Braver. "Enhancing task fMRI preprocessing via individualized model-based filtering of intrinsic activity dynamics," NeuroImage, 247, 118836, 2020.

[19] T. Schön and A. Wills and B. Ninness. "System identification of nonlinear state-space models," Automatica, vol. 47, 1, pp. 39-49, 2011.

[20] J.J.Hopfield, "Neurons with graded response have collective, computational properties like those of two-state neuron," PNAS, vo1.81, pp.3088-3092, 1984.

[21] L. J. Larson-Prior and R.Oostenveld and S.Della Penna and G. Michalareas and F. Prior and A. Babajani-Feremi and J.-M. Schoffelen and L.Marzetti and F.de Pasquale and F.Di Pompeo and J. Stout and M.Woolrich and Q.Luo and R.Bucholz and P.Fries and V. Pizzella and G. L. Romani and M. Corbetta and A. Z. Snyder. "Adding dynamics to the Human Connectome Project with MEG," NeuroImage, 80, pp. 190-201, 2013.

[22] G. Nolte, "The magnetic lead field theorem in the quasi-static approximation and its use for magnetoenchephalography forward calculation in realistic volume conductors," Physics in Medicine and Biology, 48 22, pp. 3637-3652, 2003.

[23] A. Schaefer, R. Kong, E. M. Gordon, T. O. Laumann, X.-N. Zuo, A. J. Holmes, S. B. Eickhoff and B. T. T. Yeo, "Local-Global Parcellation of the Human Cerebral Cortex from Intrinsic Functional Connectivity MRI," Cerebral Cortex, pp. 1-20, 2017.