Fast state-space methods for inferring dendritic synaptic connectivity

Ari Pakman* , Jonathan Huggins†, Carl Smith* and Liam Paninski*

*Department of Statistics
Center for Theoretical Neuroscience
Grossman Center for the Statistics of Mind
Columbia University
New York, NY, 10027

†Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, MA 02139

August 8, 2013

Abstract

We present fast methods for filtering voltage measurements and performing optimal inference of the location and strength of synaptic connections in large dendritic trees. Given noisy, subsampled voltage observations we develop fast l_1-penalized regression methods for Kalman state-space models of the neuron voltage dynamics. The value of the l_1-penalty parameter is chosen using cross-validation or, for low signal-to-noise ratio, a Mallows' C_p-like criterion. Using low-rank approximations, we reduce the inference runtime from cubic to linear in the number of dendritic compartments. We also present an alternative, fully Bayesian approach to the inference problem using a spike-and-slab prior. We illustrate our results with simulations on toy and real neuronal geometries. We consider observation schemes that either scan the dendritic geometry uniformly or measure linear combinations of voltages across several locations with random coefficients. For the latter, we show how to choose the coefficients to offset the correlation between successive measurements imposed by the neuron dynamics. This results in a "compressed sensing" observation scheme, with an important reduction in the number of measurements required to infer the synaptic weights.

1 Introduction

Understanding the synaptic organization of local neural circuits remains a central challenge in neuroscience. To make progress towards this aim it would be of great value to measure the full synaptic connectivity on the dendritic tree. In particular, we would like to quantify not just which neurons are connected to a given cell, but also where these synaptic inputs are on the postsynaptic dendritic tree, and with what strength (Fig. 1). Such a technique would help in addressing a variety of open questions on the localization and maintenance of synaptic plasticity (Sjostrom et al., 2008), and would facilitate the study of nonlinear dendritic computations.

To achieve this goal, we can combine the ability to stimulate individual presynaptic neurons with high temporal resolution (either electrically, via intracellular stimulation, or optically (Packer et al., 2012) and to simultaneously image postsynaptic neurons at subcellular spatial resolution. In particular, we can use two available, complementary types of data to obtain the best possible estimates:

1. Anatomical measurements of the postsynaptic neuron's shape and dendritic arborization. This provides a backbone on which we can build a dynamical model of the postsynaptic cell.
2. Voltage-sensitive fluorescence, observed at subcellular resolution. Modern imaging methods can access small dendritic structures and allow rapid observations from many spatial locations (Reddy and Saggau, 2005; Iyer et al., 2006; Vucinic and Sejnowski, 2007; Kralj et al., 2011). This provides access to the key dynamical variable of interest, the spatiotemporal subthreshold voltage.

Since current voltage imaging technologies have relatively low signal-to-noise ratio (SNR) (Djurisic et al., 2004; Dombeck et al., 2004; Sacconi et al., 2006; Nuriya et al., 2006; Canepari et al., 2007; Milojkovic et al., 2007; Fisher et al., 2008; Djurisic et al., 2008; Canepari et al., 2008; Peterka et al., 2011), we have to apply optimal filtering methods to exploit these measurements fully.

In this paper we present fast methods to optimally filter these voltage measurements and infer the location and strength of synaptic connections in the dendritic tree. The problem is formulated in a state-space model framework and builds on fast Bayesian methods that we have previously developed (Huys et al., 2006; Huys and Paninski, 2009; Paninski and Ferreira, 2008; Paninski, 2010; Huggins and Paninski, 2012; Pnevmatikakis, Paninski, Rad and Huggins, 2012), for performing optimal inference of subthreshold voltage given noisy and incomplete observations. A key contribution of this work is to note that these fast filtering methods can be combined with fast optimization methods from the sparse Bayesian literature (Efron et al., 2004) to obtain a fast solution to this synaptic estimation problem.

We also present a fully Bayesian approach to the inference problem, using a spike-and-slab prior (Mitchell and Beauchamp, 1988). Although computationally more intensive, this approach provides confidence intervals for the estimates of the synaptic weights.

An additional contribution of this paper is in the area of experimental design. There has been much interest recently in experimental implementations of the compressed sensing paradigm (Nikolenko et al., 2008; Studer et al., 2012), which allows one to reconstruct a signal that is sparse in some basis from a small number of measurements (Candès and Wakin, 2008). In our case, the measurements are performed on voltages with a temporal dynamics dictated by the cable equation of the neuron. We show how to compensate for this dynamics in the voltage observations in such a way that the compressed sensing results apply to our case.

The paper is organized as follows. Section 2 presents the basic ideas of the inference and observation methods, with the mathematical details presented in the Appendices. Section 3 illustrates our results with simulated data on a toy neuron and on a real large dendrite tree. We conclude in Section 4 with some possible extensions to our model.

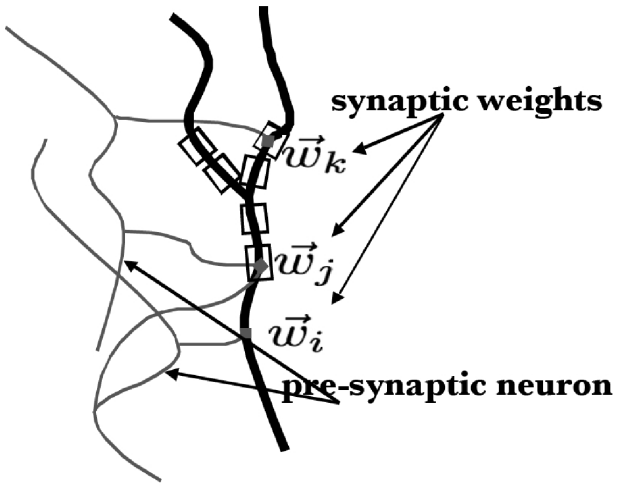

Figure 1: Schematic of proposed method. By observing a noisy, subsampled spatiotemporal voltage signal on the dendritic tree, simultaneously with the presynaptic neuron's spike train, we can infer the strength of a given presynaptic cell's inputs at each location on the postsynaptic cell's dendritic tree.

2 The dynamical model

The dynamical model

For concreteness, we begin with the following model. We assume that observations are available from a neuron with N compartments in which the passive cable dynamics and the observation equations are

$$V_{t+dt} = AV_t + WU_t + \epsilon_t, \qquad \epsilon_t \sim \mathcal{N}(0, \sigma^2 dt I) \qquad t = 0, \ldots, T-1 \tag{2.1}$$

$$y_t = B_t V_t + \eta_t, \qquad \eta_t \sim \mathcal{N}(0, C_y I) \qquad t = 1, \ldots, T\,. \tag{2.2}$$

In the first equation, V_t is an unobserved N-dimensional vector of compartment voltages at time t that evolves according to a discretized cable equation with timestep dt, perturbed by a Gaussian noise source ϵ_t; $\mathcal{N}(\mu, C)$ denotes a Gaussian density of mean μ and covariance C and I is the identity matrix of the appropiate dimension. Assuming we can stimulate K presynaptic neurons in a controlled manner, U_t represents a K-dimensional vector of known presynaptic signals (the presynaptic spike times filtered by some fixed synaptic current filter). Finally, W is the $N \times K$ matrix of synaptic weights that we want to estimate.

We assume an experimental setting in which we simultaneously perform S voltage observations at each discrete time t. (S could vary with time, but to keep the notation manageable we will assume that S is fixed here.) In the second equation, (2.2), y_t is an $S-$dimensional vector of observations related instantaneously to V_t by the $S \times N$ matrix B_t that specifies how the observations are performed. We will discuss below several forms for B_t. C_y is the noise covariance of the observations, which depends on the imaging apparatus used in each experiment. We assume this covariance is proportional to the identity for simplicity, i.e. $C_y \propto I$, but this condition can also be easily relaxed.

The inverse of the cable dynamics matrix $A \in \mathbb{R}^{N \times N}$ is sparse and symmetric. It encodes the membrane leak at each compartment as well as coupling currents between adjacent compartments. The sparseness of A^{-1} follows from its "tree-tridiagonal" form: the off-diagonal elements $A^{-1}_{n_1 n_2}$ are non-zero only if compartments n_1 and n_2 are first neighbors in the dendritic tree (see Hines (1984) and Paninski (2010) for details). Note that we are taking explicit advantage of the fact that the

anatomy of the imaged cell is known (or at least may be reconstructed post hoc); thus we know a priori which compartments are adjacent, and specifying the matrix A comes down to setting a few resistivity coefficients, if the diameter of each compartment can be estimated reliably (see Huys et al. (2006) for further discussion.) In general, we can potentially recover A and σ^2 via an Expectation-Maximization approach (Huys and Paninski, 2009), by observing the neuron's response to varied subthreshold current injections, before any presynaptic spikes are elicited.

This linear Gaussian model, with a passive (i.e. voltage independent) dynamic matrix A, can be a valid description for regimes with low network firing rate, so that the postsynaptic dendritic tree is in a subthreshold state. Furthermore, we assume that synaptic inputs are sufficiently small that the postsynaptic response may be treated using linear dynamics. (In an experimental setting we may enforce the subthreshold condition pharmacologically, by blocking voltage-gated sodium channels in the post-synaptic cell.)

On the other hand, real neural systems are known to depart from this linear, passive Gaussian regime. The noise can be non-Gaussian and strongly correlated (due, e.g., to unobserved spikes in the circuit), and the dynamics equation becomes non-linear when voltage dependent conductances and driving forces are taken into account. Also, for some measurement techniques, the observation equation may depart from the form (2.2). We discuss some of these generalizations in Section 4.

The likelihood function and the sparsity penalty

We assume that in equations (2.1)-(2.2) the variables and parameters are as follows:

- Known/Observed: $A, U_t, \sigma^2, dt, B_t, y_t, C_y$.
- Unknown/Unobserved: V_t, W.

If the system evolves during T time units, we can collect all the voltages V_t into the NT-vector V and all the observations y_t into the ST-vector Y. The complete log-likelihood for the combined V and Y variables is (Durbin et al., 2001)

$$\begin{aligned}
\log p(Y,V|W) &= \log p(Y|V) + \log p(V|W) && (2.3)\\
&= \sum_{t=1}^{T} \log p(y_t|V_t) + \sum_{t=2}^{T} \log p(V_t|V_{t-1},W) + \log p(V_1) && (2.4)\\
&= -\frac{1}{2}\sum_{t=1}^{T} (y_t - B_t V_t)^T C_y^{-1} (y_t - B_t V_t) && (2.5)\\
&\quad -\frac{1}{2}\sum_{t=2}^{T} (V_t - AV_{t-1} - WU_{t-1})^T C_V^{-1} (V_t - AV_{t-1} - WU_{t-1})\\
&\quad -\frac{1}{2} V_1^T C_0^{-1} V_1 + const.\,,
\end{aligned}$$

where $C_V = \sigma^2 dt I$, by eq. (2.1). In equation (2.3), $\log p(Y,V|W)$ abbreviates $\log p(Y,V|W,U,A,\sigma,B,C_y)$. Equation (2.3) uses the factorization $p(Y,V|W) = p(Y|V)p(V|W)$ and the first sum in (2.4) reflects the independence of the measurements y_t for each t. The second sum in (2.4) follows from the fact that the probability distribution of V_t only depends on V_{t-1} and WU_t. Finally, equation (2.5) reflects the Gaussian nature of each term in (2.4), as follows from (2.1)-(2.2). In the last term we assumed $\mathrm{E}(V_1) = 0$ (having parameterized $V \to V - V_{rest}$ to simplify the dynamics equation) and for the initial state covariance $C_0 = cov(V_1)$ we choose a convenient initial stationary condition on which we elaborate in Appendix C.

The log-likelihood (2.3) cannot be evaluated because it involves the unobserved voltages V_t, so it is useful to consider $p(Y|W)$, obtained by marginalizing the voltages as

$$p(Y|W) = \int p(Y,V|W)\, dV\,. \qquad (2.6)$$

Since $p(Y, V|W)$ is Gaussian in (Y, V) with mean linearly dependent on W, the marginal $p(Y|W)$ has the same property, and therefore $\log p(Y|W)$ is quadratic in W,

$$\log p(Y|W) = \sum_{i,j} r_{ij} W^{ij} + \frac{1}{2} \sum_{i,i',j,j'} W^{ij} M_{ij,i'j'} W^{i'j'} + const. \tag{2.7}$$

where $i = 1 \dots N\,, j = 1 \dots K$. We compute the coefficients r_{ij} and $M_{ij,i'j'}$ in Appendix A. In particular, $M_{ij,i'j'}$ is negative semidefinite and symmetric.

Our goal is to estimate the $N \times K$ synaptic weights W as

$$\hat{W}(\lambda) = \arg\max_W \log p(W|Y, \lambda) \tag{2.8}$$

$$= \arg\max_W \left[\log p(Y|W) + \log p(W|\lambda)\right]\,, \tag{2.9}$$

where we take for W a log-prior with an l_1 penalty that enforces a sparse solution,

$$\log p(W|\lambda) = -\lambda \sum_{i,j} |W^{i,j}| + const. \tag{2.10}$$

The prior (2.10) is referred to as the *lasso* prior (for 'least absolute shrinkage and selection operator') in the statistics literature; its effect is to sparsen the estimated $\hat{W}$ in (2.9), i.e., to make its components exactly 0 if they do not have a strong measurable influence on the observed data Y (Tibshirani, 1996). In our case, we introduce this prior because we expect the synaptic contact with the presynaptic neuron to occur only at a relatively small number of compartments in the dendritic tree.

The value of λ in (2.10) controls the sparsity of $\hat{W}(\lambda)$: when λ is large, the number of non-zero components in $\hat{W}(\lambda)$ is small, and vice versa. The motivation to introduce this prior here is that we expect the number of non-zero synaptic weights for each presynaptic neuron to be much smaller than the number of compartments N. While this prior turns out to be extremely convenient here (as we discuss next), more involved priors are possible; see section 4 for some further discussion.

Eq.(2.9) is a concave problem. We would like to solve it for a range of values of λ and then select a particular λ according to some criterion. If $\log p(Y|W)$ were the quadratic error of a linear regression with Gaussian noise, the solution to (2.9) for all λ could be obtained by the Least Angle Regression (LARS) algorithm introduced in Efron et al. (2004).

In our case the quadratic expression (2.7) contains contributions from the integrated unobserved voltages V. For this reason, we reformulate in Appendix A.1 the LARS algorithm of Efron et al. (2004) for our quadratic function of W, $\log p(Y|W)$. Moreover, synaptic connections are either excitatory or inhibitory ("Dale's law"), so the non-zero elements of each of the K columns of the true synaptic matrix W must have a definite sign. We consider in Section A.2 a slight modification of the LARS algorithm, LARS+, that imposes this sign condition and avoids inferring weights with the wrong sign due to poor observations or high noise.

As we show in Appendix A, the l_1-penalized form of our objective function in (2.9) implies that the solution for $\hat{W}(\lambda)$ is a piecewise linear function of λ. For $\lambda = \infty$, the solution to (2.9) is $W = 0$. As λ becomes smaller, more and more components become non-zero at the breakpoints of $\hat{W}(\lambda)$, although at some breakpoints non-zero components can also become zero. The form of the solution is thus

$$\hat{W}(\lambda) = \begin{cases} 0 & \text{for} \quad \lambda_1 < \lambda \\ \hat{W}(\lambda_i) + \mathbf{a}_i(\lambda_i - \lambda) & \text{for} \quad \lambda_{i+1} < \lambda < \lambda_i \qquad i = 1 \dots R\,, \end{cases} \tag{2.11}$$

where $\mathbf{a}_i$ are $N \times K$ matrices and the number R of breakpoints until $\lambda_{R+1} = 0$ depends on the data. The LARS/LARS+ algorithm proceeds by successively computing the pairs $(\lambda_i, \mathbf{a}_i)$ until $\lambda = 0$ is reached.

An important byproduct of the algorithm is that it provides us with an estimate of the unobserved voltages,

$$\hat{V}(\lambda) = E[V|Y, \hat{W}(\lambda)] \tag{2.12}$$
$$= \arg\max_V p(V|Y, \hat{W}(\lambda)), \tag{2.13}$$

where the second line is equal to the first because $p(V|Y, \hat{W}(\lambda))$ is a Gaussian distribution. The function $\hat{V}(\lambda)$ will be important below to select the best value for λ.

The value of $\hat{W}(\lambda)$ at $\lambda = 0$, the end point of the path, is the solution to the unpenalized maximum likelihood problem

$$\hat{W} = \arg\max_W \log p(Y|W), \tag{2.14}$$

which is the optimal least-squares (OLS) linear solution, due to the quadratic nature of $\log p(Y|W)$. This $\hat{W}$ is the linear combination of the observations y_t that minimizes the log-likelihood (2.7). For LARS+, the end point of the path is the minimum of (2.7), with the restriction that each of the K columns of the inferred synaptic matrix $\hat{W}$ has a definite sign.

Computational cost

The major computational challenge in obtaining $W^{ij}(\lambda)$ lies in explicitly computing the coefficients r_{ij} and $M_{ij,i'j'}$. As shown in Appendix A, computing r_{ij} or a row of $M_{ij,i'j'}$ requires that we solve a linear equation involving the $NT \times NT$ Hessian matrix

$$H_{VV} = \frac{\partial^2 \log p(Y, V|W)}{\partial V \partial V}. \tag{2.15}$$

Using the block tri-diagonal structure of H_{VV} (see Appendix C), this matrix solve can be performed in $O(TN^3)$ time using standard methods. However, as we show in Appendix C, if $S \ll N$ (i.e., only a minority of compartments are imaged directly, as is typically the case in these experiments) we can perform this operation approximately much more quickly, in $O(TNS^2)$ instead of $O(TN^3)$ time, using low-rank perturbation techniques similar to those introduced in Paninski (2010); Huggins and Paninski (2012); Pnevmatikakis, Paninski, Rad and Huggins (2012); Pnevmatikakis and Paninski (2012).

To run the LARS/LARS+ algorithm we need the coefficients r_{ij}, and at each breakpoint in which a new weight W^{ij} becomes non-zero, the ij-th row of $M_{ij,i'j'}$ must be computed. In general we will need to compute $W^{ij}(\lambda)$ up to $k \ll NK$ breakpoints (see below), leading to a total computational cost from acting with H_{VV} of $O(kTNS^2)$. The LARS/LARS+ algorithm also performs some smaller matrix computations at each breakpoint; including these, the total computational cost is $O(kTNS^2 + k^3)$.

Model Selection

The next task is to select the point along the LARS path $\hat{W}^{ij}(\lambda)$ that yields the "best" inferred weights $\hat{W}^{ij}$. Two major methods of model selection for our case include cross-validation and minimization of C_p-like criteria (see e.g. Efron (2004)). In both cases, the selected model is not that which minimizes the squared error on the training dataset. This solution (corresponding to (2.14)) would be computationally costly and typically greatly overfits the data, unless the data are very informative (i.e., high-SNR and $T \gg N$).

We will consider Mallows' C_p criterion (Mallows, 1973) in the low signal-to-noise ratio (SNR) limit, when the stochastic term in (2.1) can be ignored. As we elaborate in Appendix B, in this limit

and for $ST > NK$, an estimate of the generalization error of our model is given by

$$C_p(d) = \sum_{t=1}^{T} \parallel y_t - B_t \hat{V}_t(\lambda_d) \parallel^2 + 2dC_y \qquad d = 1, 2 \ldots NK, \tag{2.16}$$

where λ_d is the smallest value of λ at which there are d non-zero weights in the path (2.11). The value of d is an estimate of the degrees of freedom of the fitted model and we select the value λ_d (or equivalently d) that minimizes (2.16). This gives the best compromise between the fit to the data, represented by the first term (which decreases with lower λ) and the complexity of the model, represented by the factor d in the second term (which increases with lower λ).

As we discussed above, we expect the number of non-zero synaptic weights to be much smaller than NK. Thus one can stop the LARS algorithm if after $k \ll NK$ steps one believes that the minimum of $C_p(d)$ was attained at some $d \leq k$. This is often possible in practice (as we will see below), though the C_p curve is typically non-convex in d or λ.

The cross-validation approach is conceptually somewhat simpler. In 2-fold cross validation, we split the interval T into 2 segments and compute the $W^{ij}(d)$ weights using data from one of the two segments. With these weights we evaluate the log-likelihood in (2.7) with coefficients $r_{ij}, M_{ij,i'j'}$ computed from the left-out segment. We repeat the procedure (interchanging the training and test segments) and compute an average likelihood curve, $\bar{Q}(d)$, as the mean of the two test log-likelihoods. We select the model d at the minimum of this curve. We finally run LARS/LARS+ on the whole interval T, and estimate the value $\hat{W}^{ij}$ for d active weights.

For n-fold cross validation with $n > 2$, the data can be split into n segments. As above for $n = 2$, we successively leave out each of the n observation subsets and use the remaining observations to compute the synaptic weights $W^{ij}(d)$. In this case the held-out test set will lead to an observed training dataset Y with a gap in time (where the test set was removed). The likelihood coefficients $r_{ij}, M_{ij,i'j'}$ for this case can be obtained by a straightforward application of the method developed in Appendix A, but for simplicity in this paper we only consider the 2-fold case.

Note that there is a trade-off between these two methods. The C_p criterion is computationally fast because we only have to run the LARS/LARS+ algorithm once in order to compute the $C_p(d)$ values in (2.16), but the derivation of equation (2.16) assumes low SNR. On the other hand, the cross validation method is computationally more intensive, but its valid for any SNR.

A fully Bayesian approach

The methods presented above provide us with point estimates of the synaptic weights, but in some situations we may be interested in confidence intervals. In such cases we can adopt a fully Bayesian approach, and consider the posterior distribution of the synaptic weights (not just the maximum) under a sparsity-inducing prior. Among several possibilities for the latter, we will consider here the spike-and-slab prior (Mitchell and Beauchamp, 1988) and restrict ourselves to one presynaptic signal ($K = 1$) for simplicity. The idea is to augment each synaptic weight W_i with a binary variable s_i and consider the prior distribution

$$p(W, s|a, \tau^2) = \prod_{i=1}^{N} p(W_i|s_i, \tau^2) p(s_i|a)\,, \tag{2.17}$$

where each pair (W_i, s_i) is sampled from

$$s_i|a \;=\; \begin{cases} 1 & \text{with prob. } a\,, \\ 0 & \text{with prob. } 1-a\,, \end{cases} \tag{2.18}$$

$$W_i|s_i, \tau^2 \;\sim\; \begin{cases} \delta(W_i) & \text{for} \quad s_i = 0\,, \\ \frac{1}{\sqrt{2\pi\tau^2}} e^{-\frac{W_i^2}{2\tau^2}} & \text{for} \quad s_i = 1\,. \end{cases} \tag{2.19}$$

Note that the sparsity is achieved by assigning a finite probability a to the event $s_i = W_i = 0$. On the other hand, when $s_i = 1$, W_i is sampled from the distribution in (2.19), which is a Gaussian with zero mean and variance τ^2. In general, if we have prior information about the synaptic weights, a and τ could depend on the location of each weight, but we do not pursue this idea here.

For the hyperparameters a and τ^2 we will use conjugate hyperpriors (Gelman et al., 2004): a Beta distribution for a and an Inverse Gamma for τ^2,

$$p(a|\alpha_a, \beta_a) \quad \propto \quad a^{\alpha_a}(1-a)^{\beta_a}\,, \tag{2.20}$$

$$p(\tau^2|\alpha_\tau, \beta_\tau) \quad \propto \quad \tau^{-2\alpha_\tau - 2} e^{-\frac{\beta_\tau}{\tau^2}}\,. \tag{2.21}$$

The presence of *two* hyperparameters makes the spike-and-slab similar to the elastic net (Zou and Hastie, 2005), which combines both lasso and ridge penalties. The sparsity parameter a corresponds roughly to the lasso λ parameter, while τ^{-2} is the coefficient of the ridge penalty. The latter is particularly important for large neurons with synaptic weights localized in *several* nearby locations: small values of τ^2, which can be favored by an appropriate hyperprior, lead to similar values for these correlated weights and avoid incorrectly inferring one big and many small weights (Zou and Hastie, 2005).

Given the observations Y, we combine the prior distribution with the data likelihood,

$$p(Y|W) \quad \propto \quad e^{\frac{1}{2}\sum_{i,j} W_i M_{ij} W_j + \sum_i r_i W_i}\,, \tag{2.22}$$

and consider the joint posterior distribution

$$p(W, s, a, \tau^2|Y) \quad \propto \quad p(Y|W)p(W, s|a, \tau^2)p(a|\alpha_a, \beta_a)p(\tau^2|\alpha_\tau, \beta_\tau)\,, \tag{2.23}$$

which we can sample from using Markov Chain Monte Carlo (MCMC) methods. In particular we use a Gibbs sampler that cyclically samples a, τ^2 and (W, s). For the latter, we use an exact Hamiltonian Monte Carlo sampler based on the method of (Pakman and Paninski, 2013). The sign constraint from Dale's law can be imposed by simply restricting (2.23) to be non-zero only for $W_i \geq 0$ or $W_i \leq 0$. Note that this fully Bayesian approach is computationally more intensive than the LARS method, not only due to the computational cost of the MCMC sampling, but because we need to pre-compute the full M matrix in (2.22).

Note that in (2.19) we assumed a (truncated) Gaussian prior distribution for the non-zero synaptic weights. This choice simplifies the sampling from the posterior (2.23), but it is known that heavy-tailed distributions, such as log-Normal, are a more realistic choice (Song et al., 2005; Barbour et al., 2007). Although we will not consider them here, such priors can also be studied using appropriate MCMC techniques, see (Smith, 2013) for details.

Observation Schemes

An observation scheme is a particular selection of the matrices B_t appearing in the observation equation (2.2). The simplest such matrix would be the identity, i.e., all compartments are observed directly. Since it is currently not experimentally feasible to observe the voltages on a full large dendritic tree in three dimensions with high temporal resolution, we will consider the following two cases in which B_t are fat rectangular matrices, i.e., the number of observations S per time step is less than the number of compartments N.

- **Scan observation**.
  In this scheme, the $S \times N$ observation matrices B_t are

$$(B_t)_{ij} = \begin{cases} 1 & \text{for } j = p * i + t \mod N \\ 0 & \text{otherwise.} \end{cases} \qquad i = 1 \ldots S, \quad t = 1 \ldots T, \quad p \in \mathbb{N}^+ \tag{2.24}$$

  In other words, we observe S compartments at each time, at changing locations. Each row has a 1 at columns separated by p entries, that move cyclically at each observation. Variations of this scheme are common in many electro-physiological experiments.

- **Compressed sensing**.
  This relatively new paradigm asserts that the number of measurements needed to reconstruct a signal is much smaller for sparse than for non-sparse signals. For a review and references to the literature see, e.g., (Candès and Wakin, 2008). To see how this applies to our case, let us consider the case with no noise in the voltage dynamics, i.e., $\sigma = 0$. As shown in eqs.(B.3)-(B.6), in this limit the observations y_t are related to the synaptic weights as

$$y_t = D_t W + \eta_t \tag{2.25}$$

  where

$$\begin{aligned} D_t &= B_t F_t \qquad \in \mathbb{R}^{S\times N} && (2.26)\\ F_t &= A^{t-2}U_1 + A^{t-1}U_2 + \cdots + U_{t-1} && (2.27) \end{aligned}$$

  The matrix D_t is an "effective" observation matrix. Note that the total number of measurements in an experiment lasting T time steps is ST, and suppose that W has K non-zero entries. If the entries of D_t are chosen randomly from a zero-mean Gaussian and the total number of measurements obeys

$$ST \geq c_1 K \log(N/K) \tag{2.28}$$

  for some constant c_1, then, with overwhelming probability, the quadratic error between W and its lasso estimate $\hat{W}$ is bounded as

$$||\hat{W} - W||^2 \leq c_2 C_y \,, \tag{2.29}$$

  for some constant c_2 (Candes et al., 2006). The bound (2.28) stands in contrast to non-sparse signals, which require $O(N)$ measurements for a low-error reconstruction. The experimental realization of this observation scheme is presently a very active area of research, see e.g. (Nikolenko et al., 2008; Studer et al., 2012).

  In our case, we can implement the compressed sensing scheme by choosing, at each t, a matrix D_t whose entries are i.i.d. samples from a positive Gaussian distribution, and an observation matrix

$$B_t = D_t F_t^{-1} \,. \tag{2.30}$$

  A potential numerical problem arises for an extended set of times without external stimulus U_t. Suppose $U_t = 0$ for $t = t_1 + 1, \ldots, t_2$. Then, as follows from (2.27),

$$F_{t_2} = A^{t_2 - t_1} F_{t_1} \,, \tag{2.31}$$

  and since the matrix A is stable (i.e., its eigenvalues are < 1), the matrix F_{t_2} is ill-conditioned, leading to a numerical instability in the computation of (2.30). So this observation scheme is better applied to measurements performed at those times t in which a stimulus $U_t \neq 0$ is present.

3 Results

We illustrate our methods using simulated data in a toy model and in a real neuronal geometry. In each case, we started with a known matrix A (based on the known dendritic geometery), and chose values for σ^2, dt, C_y, U_t and B_t. We sampled values for the dynamic noise ϵ_t and obtained values for V_t by simulating eq.(2.1) forward in time. We next sampled values for the observation noise η_t and used eq.(2.2) to obtain y_t. In all the cases we initialized the V_t dynamics so that V_t was a time-stationary process, ensuring the validity of the approximations discussed in Appendix C. All the algorithms were implemented in MATLAB.

Toy neuron

The toy model neuron, shown in Figure 2, has $N = 35$ compartments and one branching point, and we assumed three positive synaptic weights, indicated by circles in Figure 2. Figures 3 to 9 show results corresponding to one presynaptic input ($K = 1$), with a stimulus U_t shown in the upper panel of Figure 4. For the scan observation scheme, we used B_t as in (2.24), with $S = 7$ observations at each instant and column spacing $p = 5$.

Figure 3 illustrates the inferred weights as a function of λ for both the LARS and LARS+ algorithms with scan observation, for a simulation with low noise covariance $C_y = 0.05$, where in a slight abuse of notation we abbreviate $C_y = [C_y]_{11}$ (recall that the observation noise C_y is proportional to the identity here). As described above, the weights are zero for $\lambda = \infty$ and become active at breakpoints as λ becomes smaller. In this Figure (as in Figures 5, 6, 8 and 9), the colors of the weights correspond to their location in Figure 2. In the upper panel, corresponding to the LARS algorithm, all the weights are active at $\lambda = 0$. On the other hand, for LARS+ in the lower panel, some weights never become active because that would violate the $W \geq 0$ restriction. The vertical lines show the weights selected by the C_p and 2-fold cross validation criteria.

An estimate of the signal power for V can be obtained as

$$P_s = \text{Mean}_i \left(\text{Var}_t V_t(i) \right) \simeq 0.012, \tag{3.1}$$

where $V_t(i)$ is the voltage at compartment i and time t. Using this value we can estimate the signal-to-noise ratio (SNR) for Figure 3 as

$$\text{SNR} = P_s / C_y \simeq 0.24 \tag{3.2}$$

Figure 4 shows, along with the presynaptic signal, the true, observed and estimated voltages for a similar simulation with scan observations, but higher noise covariance $C_y = 8$ and SNR estimated as

$$\text{SNR} = P_s / C_y \simeq 0.0015\,. \tag{3.3}$$

Figure 4 also shows the voltages estimated at the end of the LARS path, the OLS point. These estimates, shown in the fifth panel, are of poor quality compared with those at the point selected by the C_p criterion, shown in the last panel. This highlights the importance of the l_1 prior in the model. The higher observation noise in this case translates into a solution path $\hat{W}^{ij}(\lambda)$ in Figure 5 in which the weights at locations with zero true weights grow as $\lambda \to 0$ to values comparable to the non-zero true weights (i.e., overfitting).

The weights inferred at the LARS+/C_p point are shown in more detail in Figure 6. An interesting phenomenon in this noisier case is that the inferred weights are locally spread around their true locations. This is due to the matrix A in the dynamical equation (2.1), whose inverse, as mentioned in Section 2, has a non-zero off-diagonal entry $A^{-1}_{n_1 n_2}$ if compartments n_1 and n_2 are first neighbors.

Figure 7 shows C_p and cross-validation statistics for these data as a function of the degrees of freedom d, for LARS. Note that the minima for both model selection curves are at close-by points.

Figure 8 presents a comparison of scan observations and compressed sensing. Across 20 simulations, the synaptic weights were C_p-estimated as a function of the experiment time. For each neuron compartments the median and .25/.75 quantiles of the estimated weights are indicated. In each simulation, the observations for both observation schemes were made on the same data. The figure shows, as expected, that the compressed sensing results are superior, having a smaller dispersion and converging faster to the true values.

In Figure 9, we examine a population summary of the results of 100 simulations with the same parameters as in Figures 4 and 5. As expected, the LARS/LARS+ results are (downward) biased and have low variance, and the OLS results are unbiased but have high variance. This illustrates the bias-variance trade-off involved in the l_1 regularization (Geman et al., 1992). Note that for LARS+ the values above the median are less dispersed than for LARS.

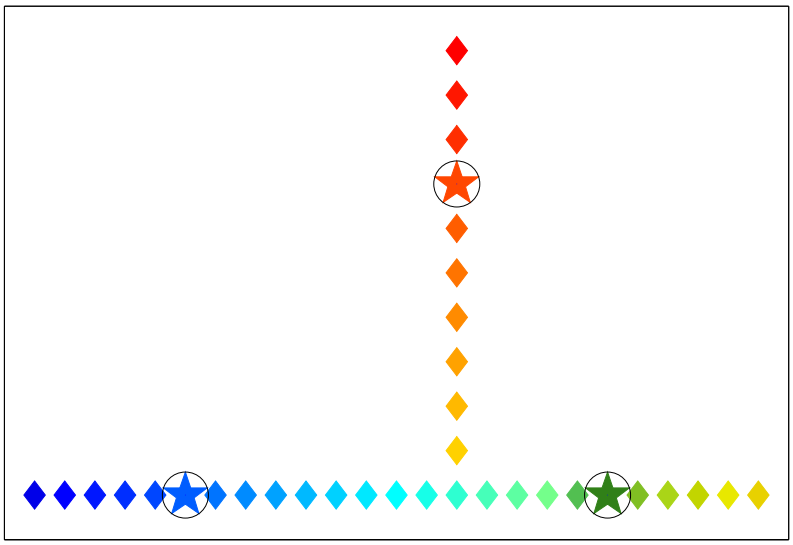

Figure 2: Toy neuron with 35 compartments used for the data of Figures 3-10. The three compartments with non-zero synaptic weights for the simulations with one presynaptic signal are indicated by a circle. We colored the compartments to ease the visualization of the corresponding inferred weights in Figures 3, 5, 6, 8 and 9.

Figure 10 shows the stimuli and voltages for a simulation in the toy neuron with $C_y = 8$ and two presynaptic signals ($K = 2$). There were three non-zero weights for each presynaptic signal, located at different compartments. The signal-to-noise ratio is estimated as

$$\text{SNR} = P_s/C_y \simeq 0.0016 \tag{3.4}$$

While the quality of the voltages estimated in Figure 10 is good, note that in general we expect the inferred weights and voltages to lose accuracy as K grows, since we have to estimate more weights, KN, given the same number of observations y_t. Thus our focus here is on modest values of K.

Real neuron geometry

Figures 11 to 13 show simulated results on a real neuronal geometry with $N = 2133$ compartments, taken from the THINSTAR file[1], corresponding to a rabbit starburst amacrine cell (Bloomfield and Miller, 1986). In all the cases we considered one presynaptic input signal ($K = 1$) and estimated the weights with the C_p criterion. We chose 28 compartments with non-zero weights (Fig. 12, top left panel). In all the figures we had

$$\text{SNR} = P_s/C_y \simeq 0.0034\,, \tag{3.5}$$

and the observation matrices B_t had $S = 40$ rows and $N = 2133$ columns. Inference required < 10 minutes to process 700 timesteps of data, with $k = 140$, on a laptop (Linux; Core 2 Duo processor, 2.53GHz).

Figure 11 shows clearly that with the noisy and subsampled voltages of the third panel ('Noisy Observations'), obtained with scan observations, we can reconstruct with good quality the full spatiotemporal voltages in the last panel.

Figure 12 shows the true and inferred synaptic weights in the THINSTAR neuron geometry for 20 simulations of scan observations. The lower left panel, showing the median of the inferred

[1] Available at `http://neuromorpho.org`.

weights, shows that our method is able to infer the locations of almost all the synaptic locations, with a strength slightly biased toward lower values. To measure the variability of the results across the 20 simulations, we computed for each compartment, the quartiles $w_{.75}$ and $w_{.25}$ of the .75 and .25 percentiles, respectively, of the weights inferred at each location over the 20 simulations. The dispersion, shown in the lower right panel, is the difference $\Delta = w_{.75} - w_{.25}$, computed at each compartment. Comparing the dispersion pattern with the true weights shows that there is some variability across the 20 simulations in the *strength* of the inferred weights, but minimal variability in the *location* of the inferred synaptic connections.

Finally, Figure 13 compares the median of the inferred weights for 10 simulations, in scan observations and compressed sensing, for both a short experiment of $T = 200$ ms. and a long experiment of $T = 720$ ms. Again, the compressed sensing scheme gives better results, already in the short experiment, while the scan observations results, while not optimal, improve from the short to the long experiment.

Some additional comments are in order. Firstly, in situations with high noise, some small non-zero weights tend to grow along the LARS path, as is clear in Figure 5. In these cases, an additional thresholding of the final inferred weights is recommended. Secondly, the geometry of the neuron, encoded in the matrix A in the dynamic equation (2.1), is not always known with full precision. In our simulations, we have noted that the imposition of the positivity constraint in the LARS+ algorithm improves significantly the robustness of the inferred weights under perturbations of the neuron geometry. Finally, note that our derivation of the compressed sensing observation matrices assumed zero noise in the hidden sector, but our results show the superiority of this observation scheme even when some amount of noise is present.

Bayesian approach

For the fully Bayesian approach with a spike-and-slab prior, Figures 14 and 15 show results from samples of the posterior distribution (2.23). We considered an experiment in the THINSTAR neuron with scan observations and $T = 400$. The rest of the parameters were similar to the $LARS/C_p$ results reported above. We used the hyper-priors (2.20) and (2.21), with parameters $\alpha_a = 5, \beta_a = 30, \alpha_\tau = 20, \beta_\tau = 0$. Figures 14 and 15 show results from 1200 samples, after discarding 300 as burn-in.

A point estimate for the synaptic weights can be obtained from the mean of the posterior samples of W_i, which allows comparison with the $LARS/C_p$ results for the same dataset. In particular, we found that the mean squared error (MSE) (compared with the true weights) was 8.48 for the Bayesian mean and 9.89 for the $LARS/C_p$ estimates.

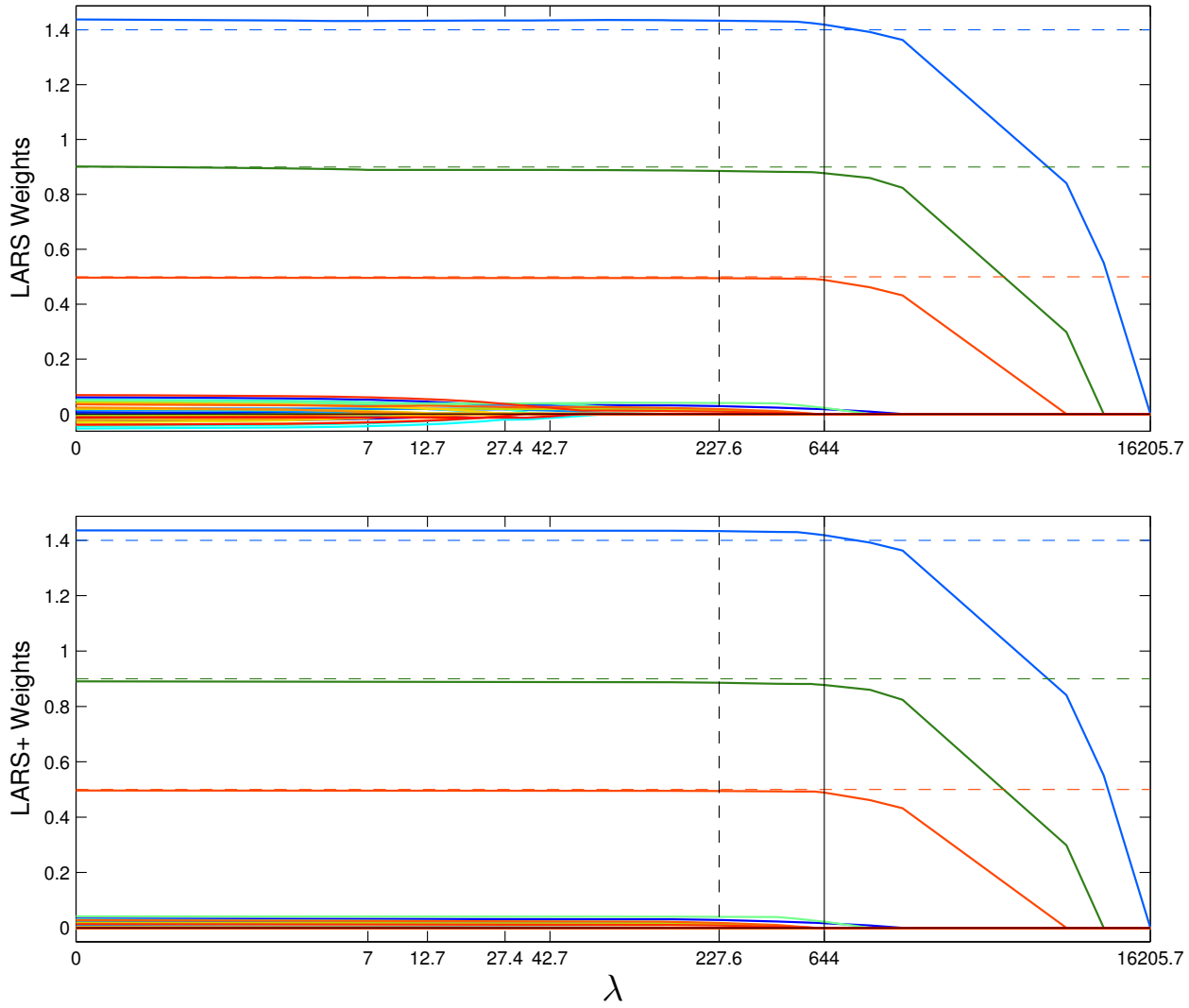

Figure 3: Toy neuron with low noise. Active weights as a function of λ along the LARS and LARS+ paths for the toy neuron model of Figure 2 with SNR $\simeq 0.24$, 7−dimensional observations and one pre-synaptic signal ($K = 1$). The simulation was run for $T = 500$ ms; The λ axis has logarithmic scale and the three true non-zero weights are indicated by horizontal dashed lines. The colors of the weights correspond to their location in Figure 2. The models selected by 2-fold cross validation are indicated by a vertical dashed line and by the C_p criterion by a straight line. The ticks in the horizontal axis indicate the value of λ at every five successive breakpoints. The high quality of the inferred weights is due to the relatively high SNR.

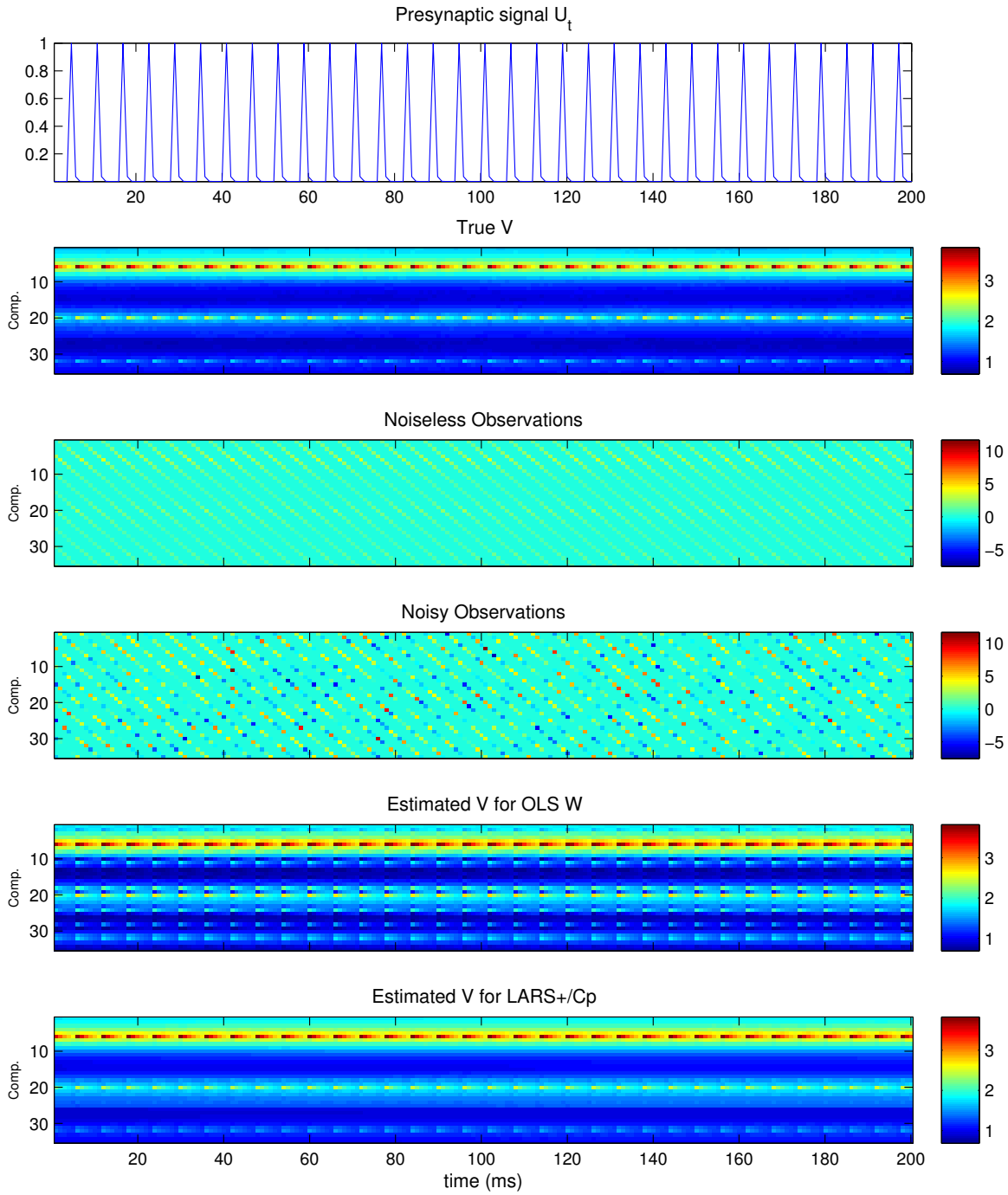

Figure 4: Toy neuron with high noise. Tracking voltages and observations for the toy neuron model of Figure 2 with SNR $\simeq 0.0015$, 7−dimensional observations and one pre-synaptic signal ($K = 1$). The simulation was run for $T = 500$ ms; only the last 200 ms are displayed. Top panel: presynaptic signal U_t, formed by exponentially filtering the (periodic) presynaptic spike trains. Second panel: true voltages evolving according to the cable equation (2.1). Third panel: noiseless observations, given by $B_t V_t$, that would have been observed had there been no noise term η_t in the observation equation (2.2). Compartments where no observations are taken at a given time t are left at zero. Fourth panel: true, noisy observations from the observation equation (2.2). Fifth panel: voltage estimates at the end of the LARS path, the OLS point. Bottom panel: inferred voltages $\hat{V}(\lambda)$ (see eq. 2.12) estimated using the sparse LARS+ weights selected by the C_p criterion. The poor quality of the OLS voltage estimates highlights the importance of the l_1 prior in the model.

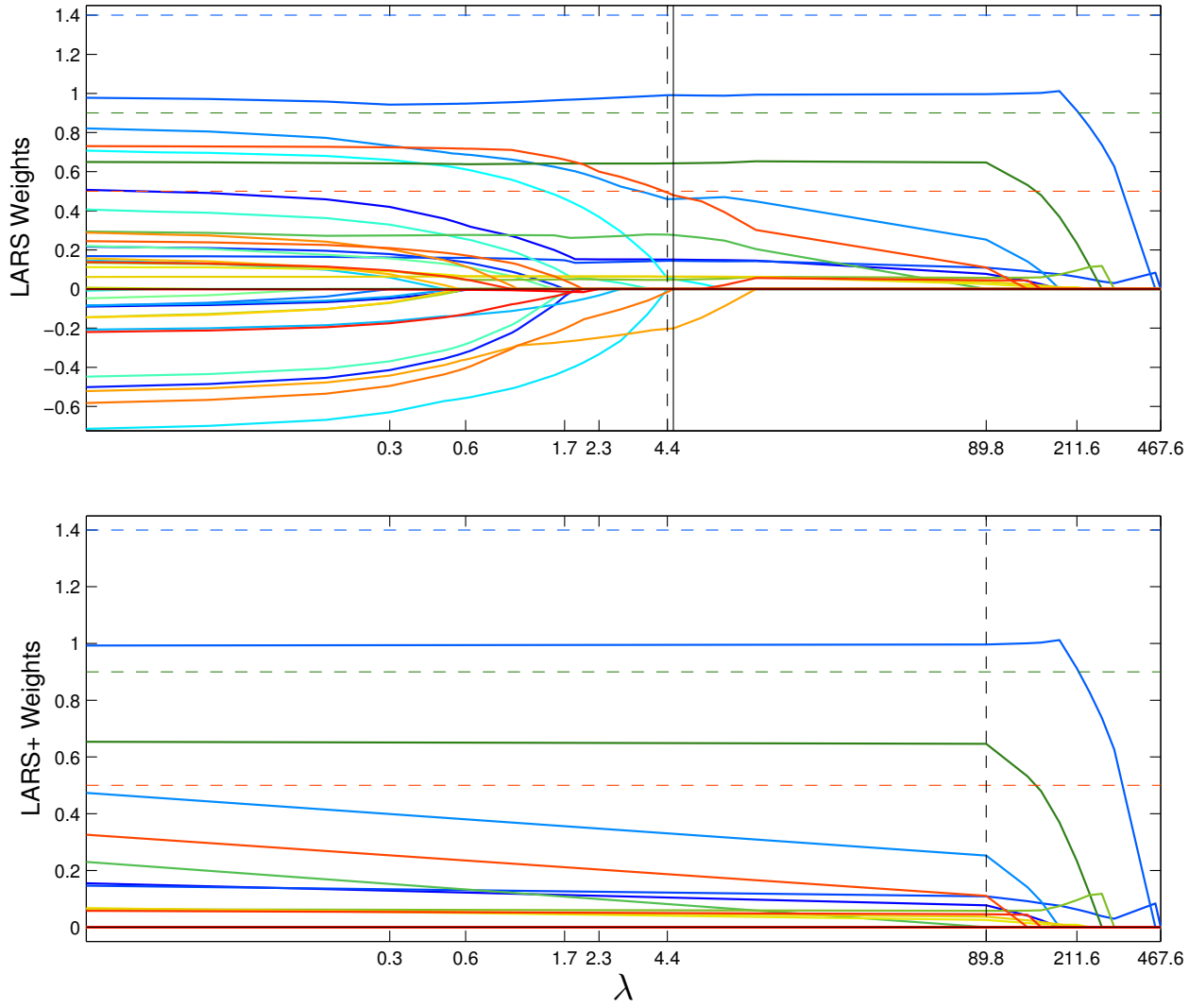

Figure 5: Toy neuron with high noise. Active weights as a function of λ along the LARS and LARS+ paths for the data shown in Figure 4 for the toy model neuron of Figure 2. Conventions as in Fig. 3. Note that inference is significantly more challenging here. The LARS solutions select weights which are biased downwards and somewhat locally spread around the corresponding true weights, as indicated by active weights with similar colors. Note that the OLS solution, at the $\lambda = 0$ point of the upper panel, performs relatively badly here.

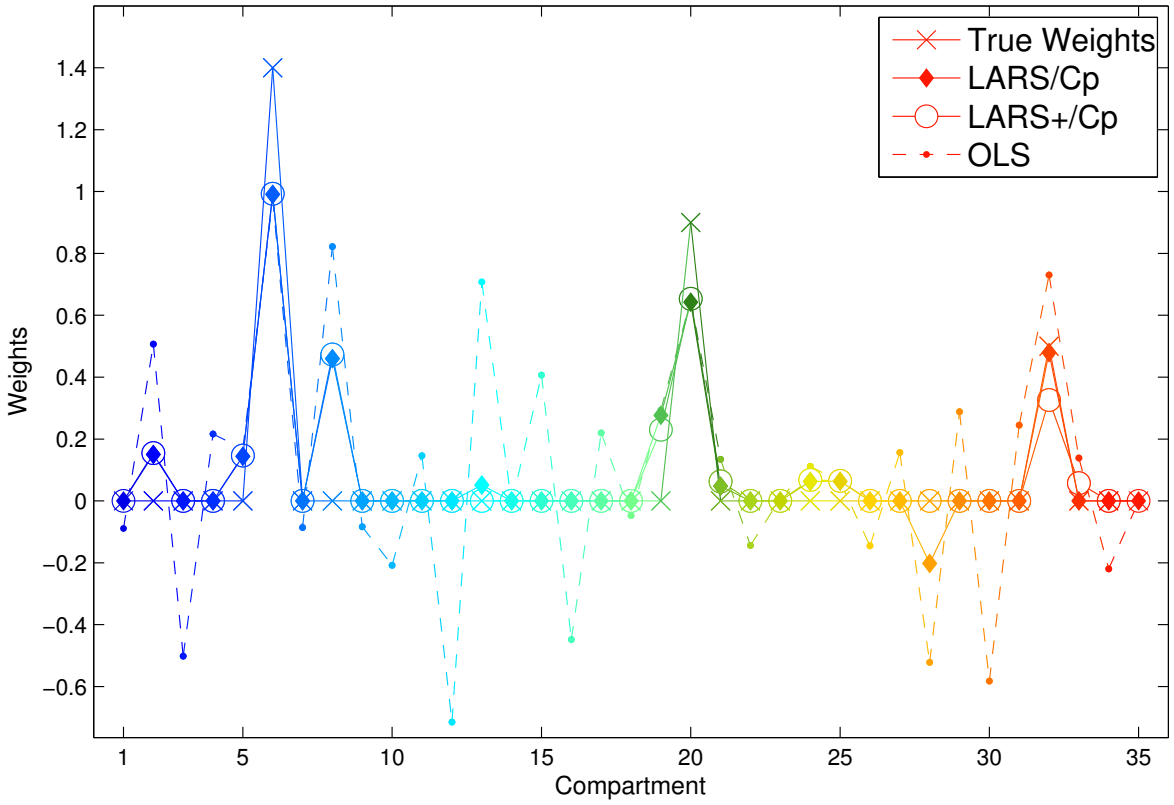

Figure 6: Toy neuron with high noise. True and C_p-selected inferred weights for the data described in Figures 4 and 5. Due to the local nature of the dynamic matrix A, there are non-zero inferred weights in the vicinity of the original weights. Note the noisy nature of the OLS results compared with the penalized results.

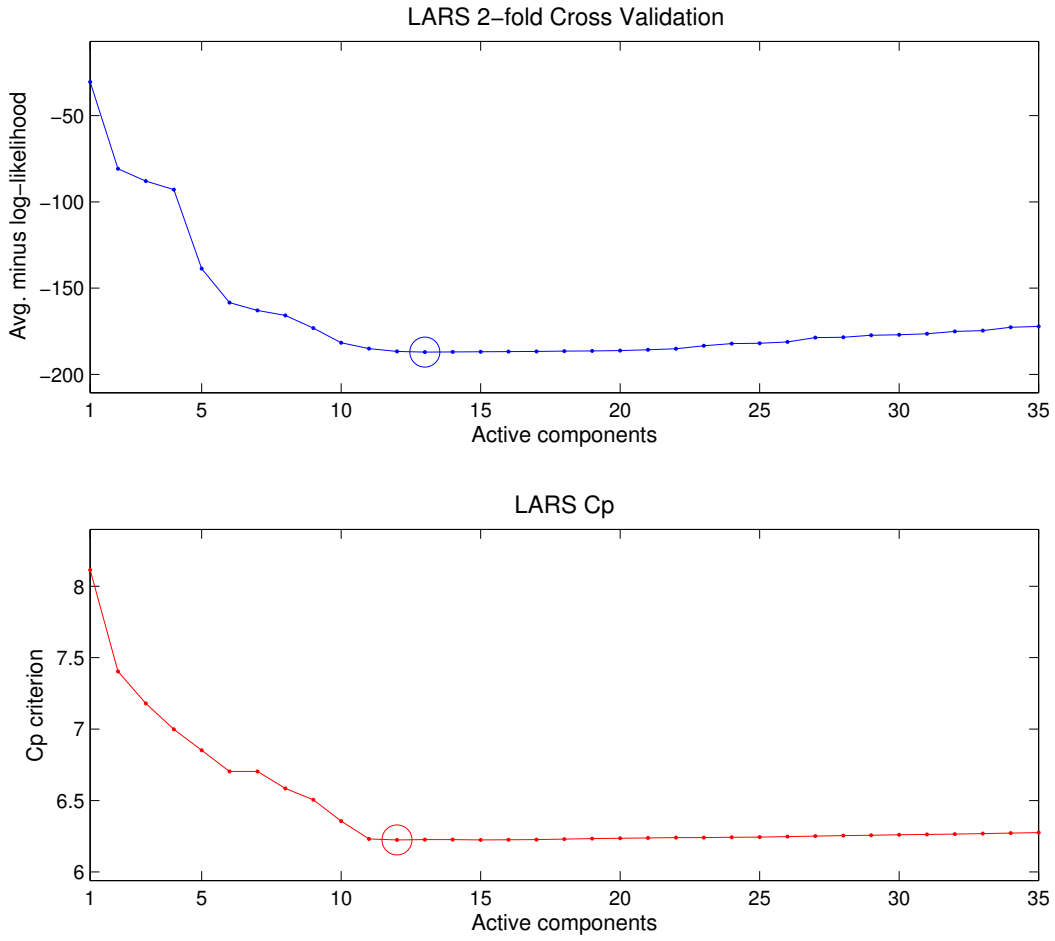

Figure 7: Cross-validation and C_p curves. Model selection curves for the low SNR data described in Figures 4 and 5 using LARS inference. The upper panel shows the average negative log-likelihood of the held-out data as a function of the number of active weights. The lower panel shows the values of the C_p criterion (eq.(2.16)) as a function of the number of active weights. The C_p criterion estimates the out-of-sample error and expresses a trade-off between a better fit to the observed data and the simplicity of the model. The minima for both model selection curves are indicated by circles.

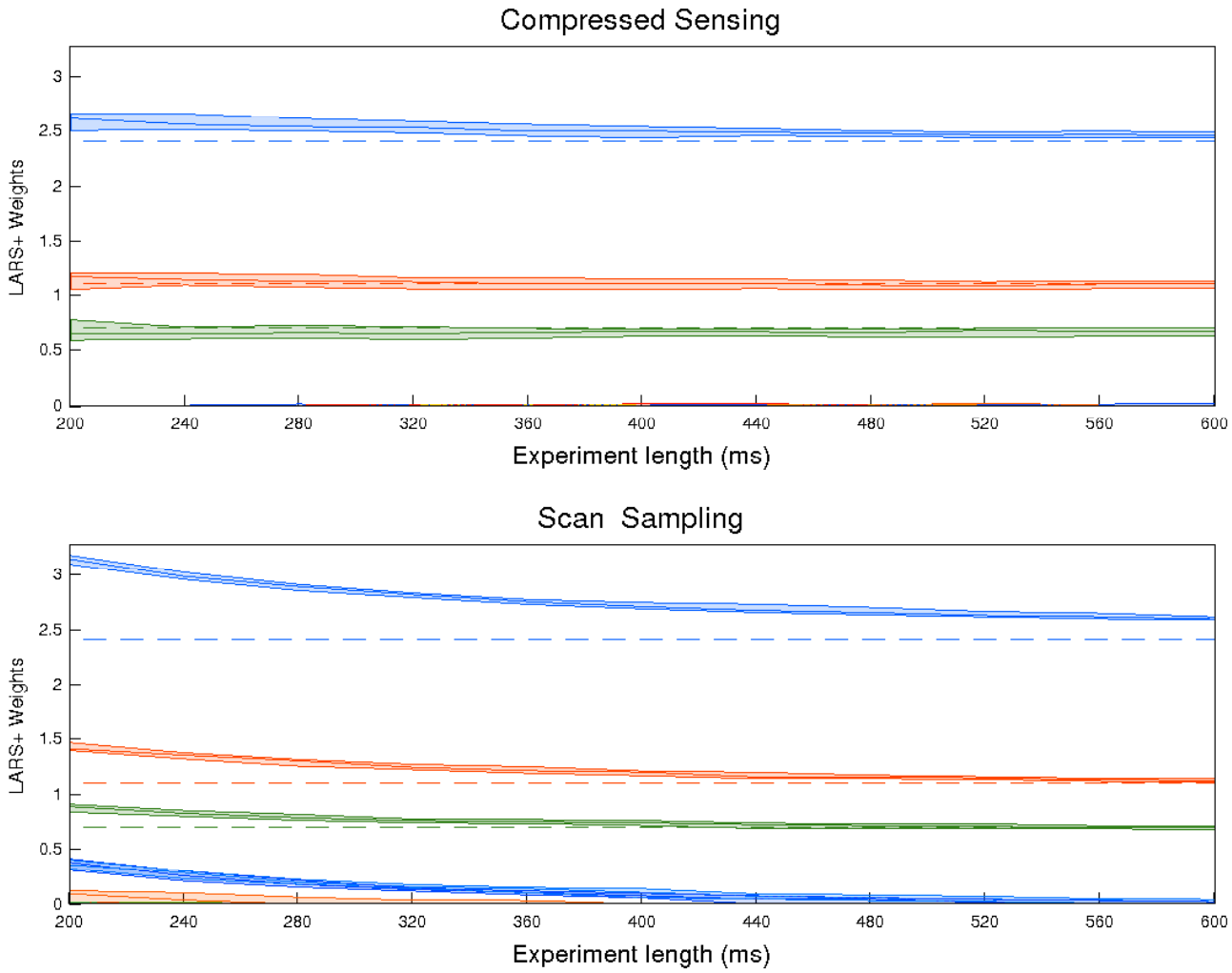

Figure 8: Compressed Sensing vs. Scan Observations. LARS+/C_p-estimated synaptic weights in 20 simulations of the toy neuron as a function of the experiment time. For each of the 35 compartments the median and .25/.75 quantiles are indicated. The average SNR was 2.18 and the dashed lines indicate the true weights. In each simulation, the observations for both observation schemes were made on the same data and at the same times. Note that the compressed sensing estimates reach the true weight values in shorter experiments, as expected.

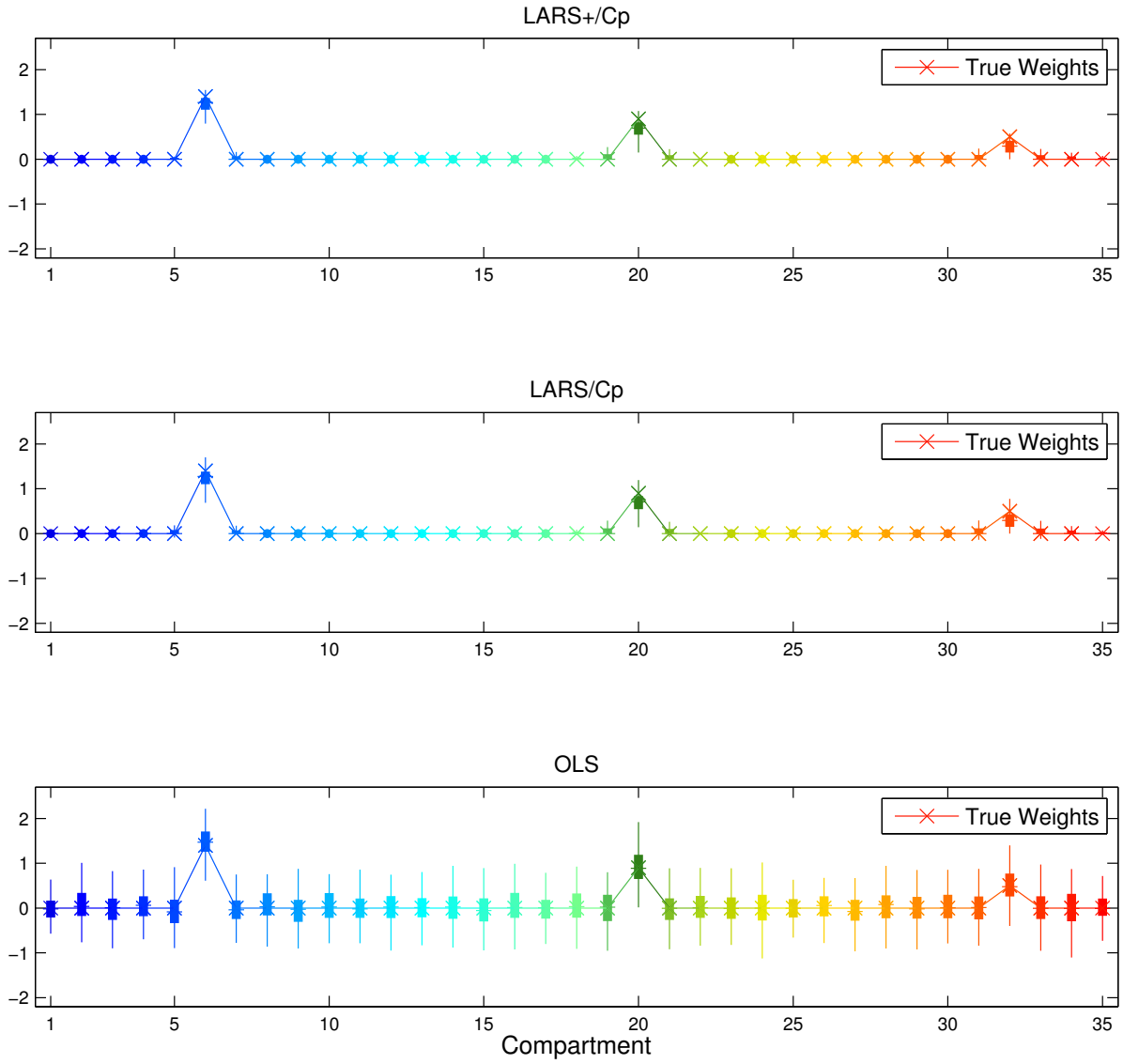

Figure 9: Toy neuron with high noise. Distribution of LARS/Cp and LARS+/Cp inferred weights for 100 simulations of the toy neuron described in 4 and 5. The filled rectangles extend from the 25th to the 75th percentile and the horizontal lines denote the median. Both the LARS and the LARS+ results are downward biased and have low variance, and the OLS results are unbiased but have high variance. Note that for LARS+ the values above the median are slightly less dispersed than for LARS.

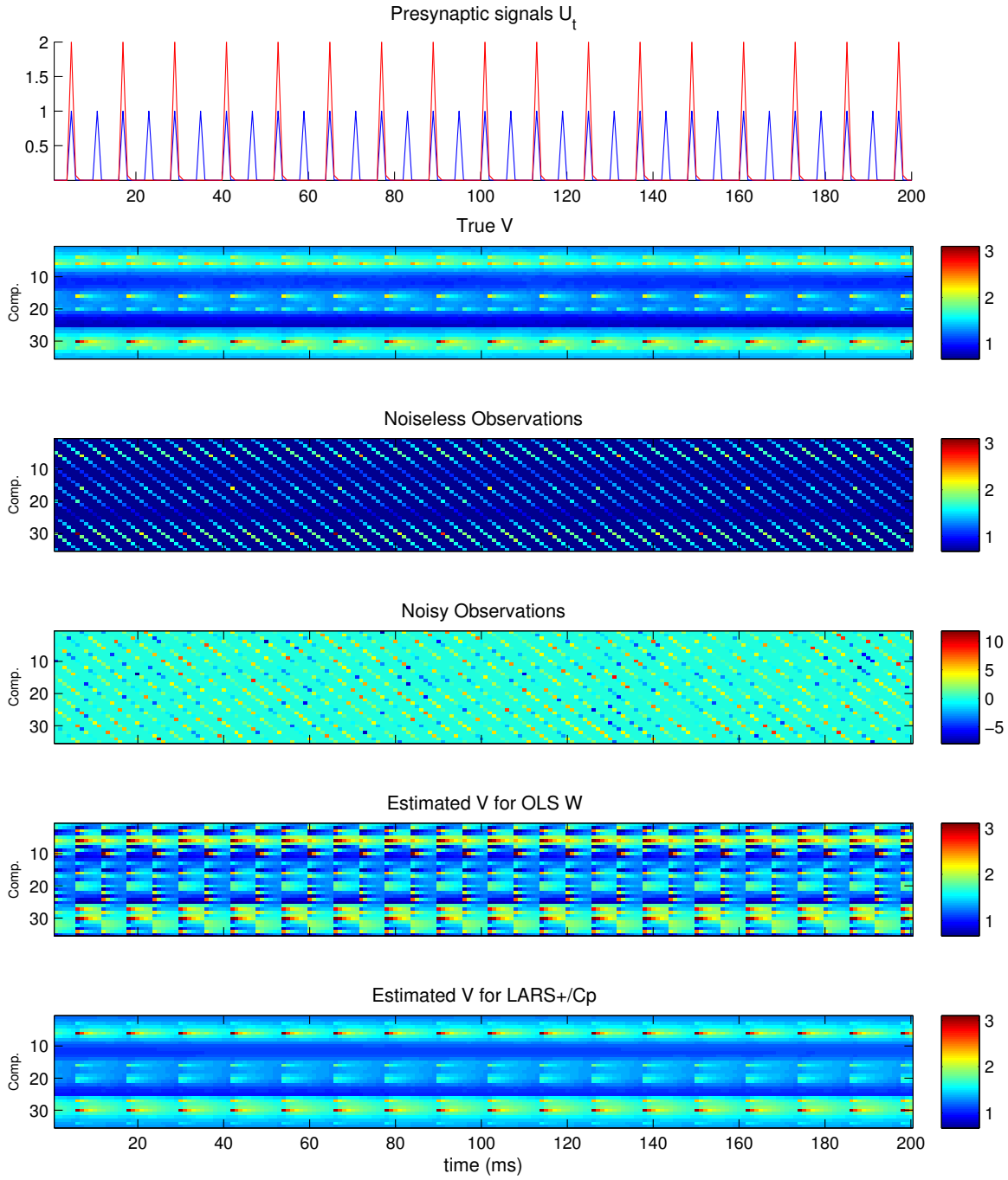

Figure 10: Toy neuron with high noise and two presynaptic inputs. Tracking voltages and observations for the toy neuron of Figure 2 with $C_y = 8$, 7−dimensional observations, *two* pre-synaptic signals ($K = 2$) and SNR $\simeq 0.0016$. Conventions and length of the experiment are as in Figure 4, except top panel shows two presynaptic input signals (one red and one blue). In general we expect the quality of the inferred weights and voltages to be worse as K grows.

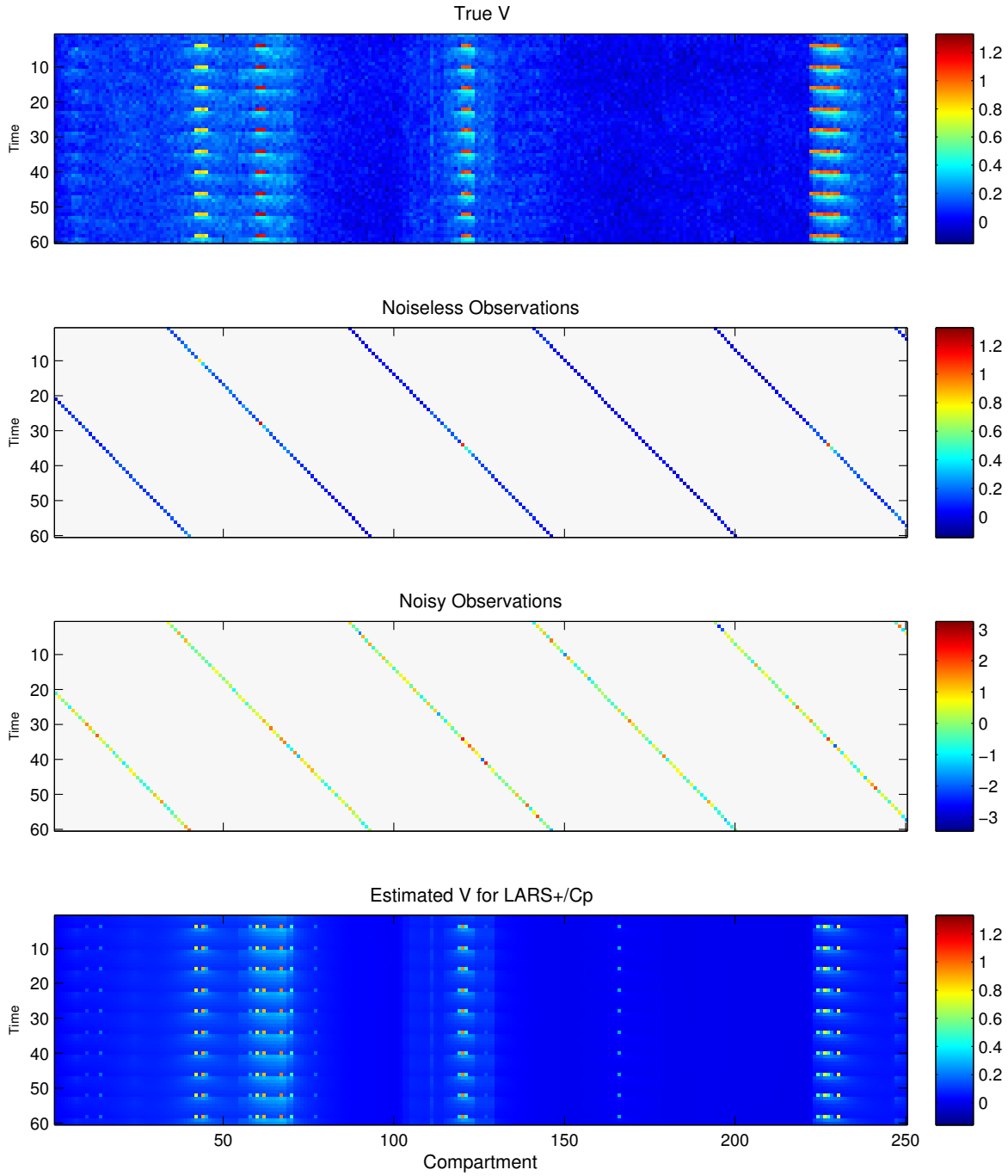

Figure 11: Big neuron with scan observations. Tracking voltages and observations for the THINSTAR neuron with 2133 compartments. Note that for space reasons the axes of the voltage panels are inverted compared to the previous figures. The simulation was run for T=700 ms and the presynaptic neuron spiked every 6 ms. Since the full N x T voltage matrix is too large to examine directly, we only show 250 compartments for the last 60ms. At each time point 40 voltage observations were made.

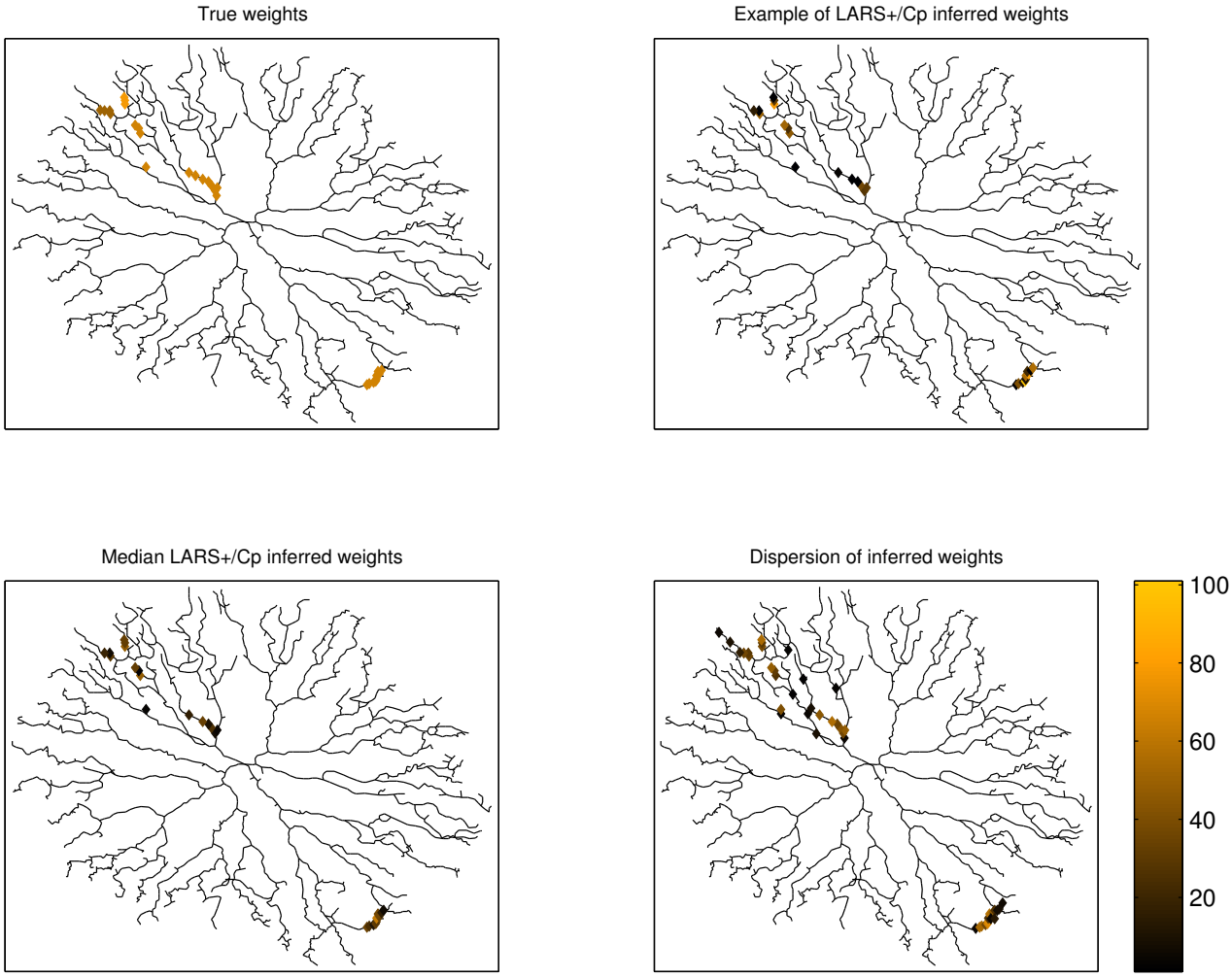

Figure 12: Big neuron with scan observations. True and inferred synaptic weights in the THINSTAR neuron for 20 simulations, with the parameters indicated in Figure 11. Note that the proposed method is able to infer the locations of almost all the synaptic locations, again with a slight downward bias. Upper right: results of a single experiment. Lower left: median results over all 20 experiments. Lower-right: dispersion of results across all 20 experiments (see main text for details). Comparing the dispersion pattern with the true weights shows that there is some variability across the 20 simulations in terms of the *strength* of the inferred weights, but the variability in terms of the *location* of the inferred synapses is small.

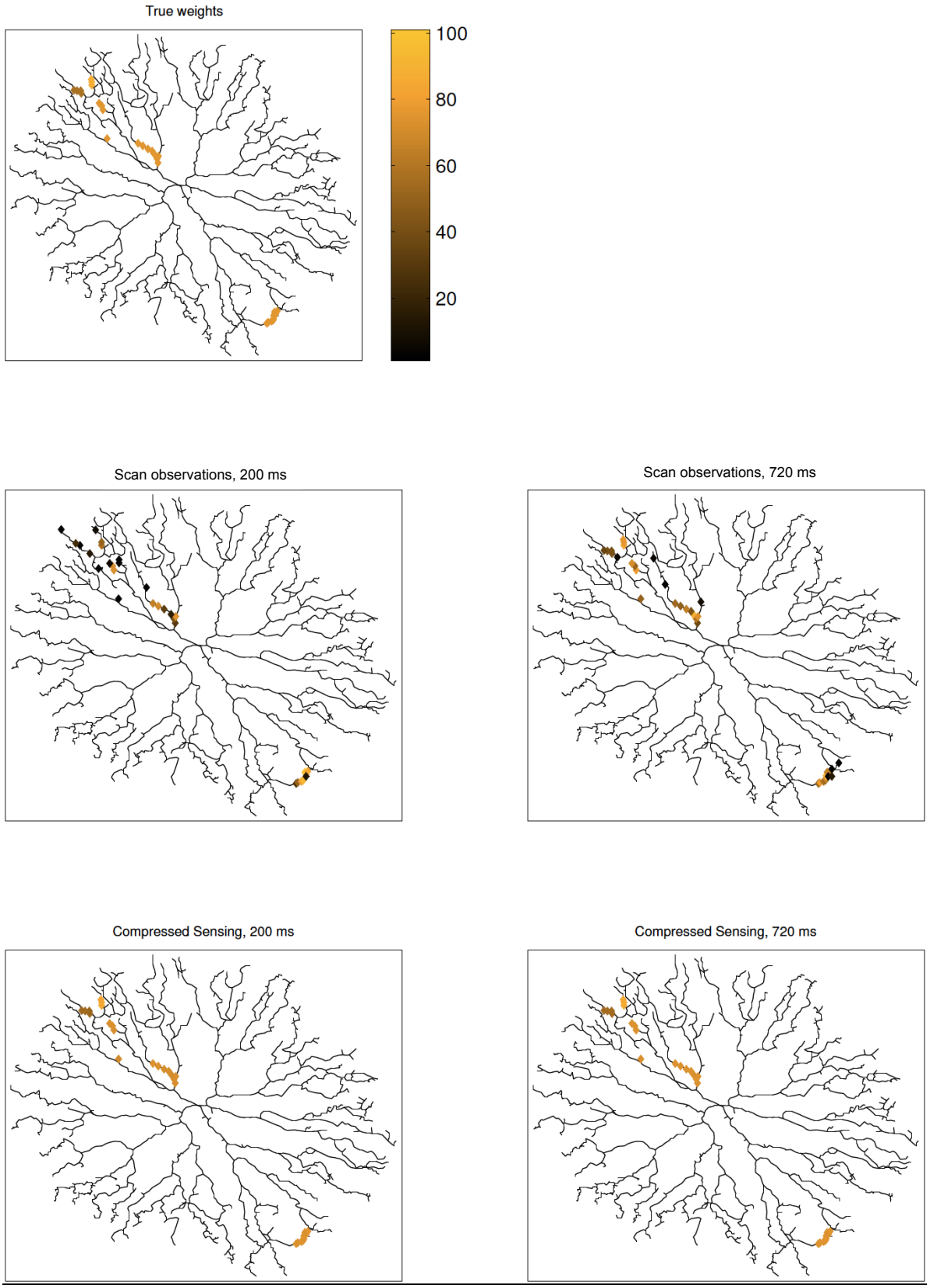

Figure 13: Comparison of observation schemes. True and median of C_p-selected inferred synaptic weights for 10 simulations in the THINSTAR neuron. The scan observation results improve from the short $T = 200$ to the long $T = 720$ experiment. The compressed sensing scheme gives optimal results already in the short experiment.

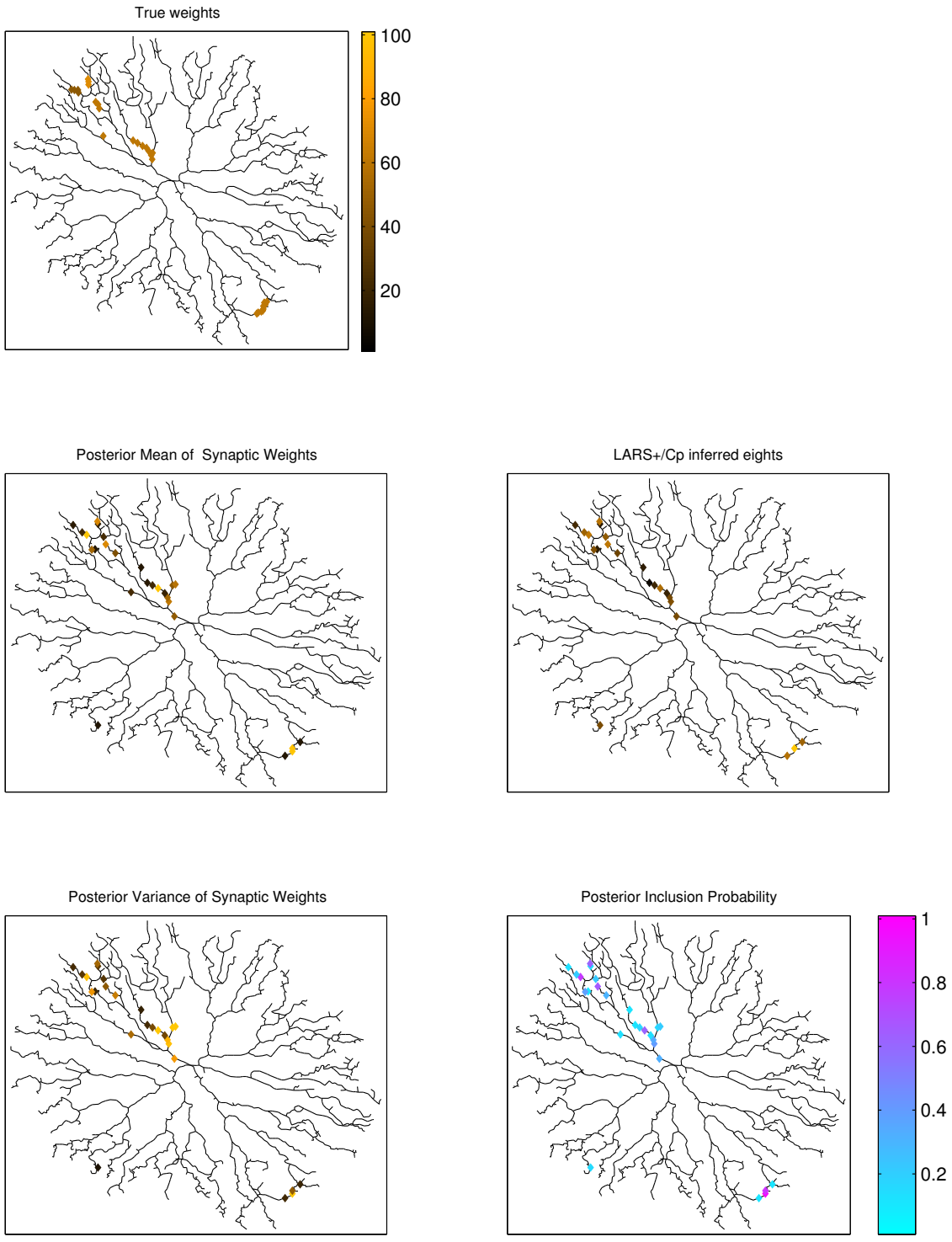

Figure 14: Bayesian Inference with spike-and-slab prior. Inferred weights for a synthetic experiment in the THINSTAR neuron with scan observations and $T = 400$. The Bayesian approach with a spike-and-slab prior quantifies the uncertainty of the inferred weights, both through the posterior variance of the weights (lower left panel) and through the posterior inclusion probability (lower right panel). The latter is defined as $p(s_i|Y)$ for each weight W_i, with s_i the binary variable from the spike-and-slab prior (see (2.18)-(2.19)). The results correspond to 1200 samples, after discarding 300 as burn-in. In the three panels with Bayesian results, we set to zero all weights with a posterior inclusion probability lower than 0.1. The color scale of all the panels is the same, except for the lower right panel.

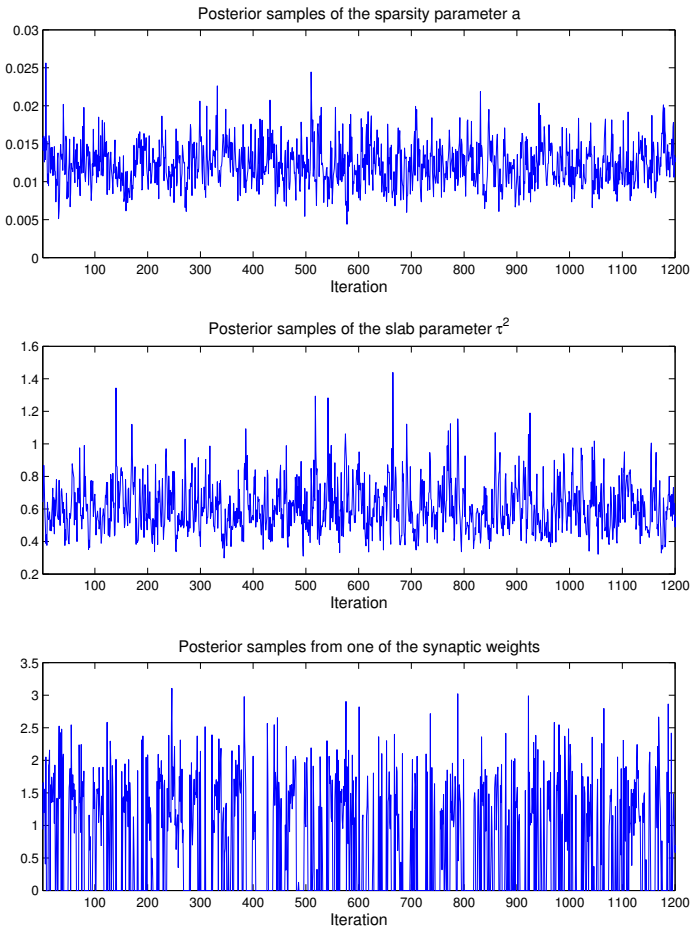

Figure 15: Samples from the Spike-and-Slab posterior distribution. Samples from the posterior distribution (2.23) for the data shown in Figure 14. Show are 1200 samples, after discarding 300 as burn-in. Upper panel: posterior samples from the sparsity parameter a. Middle panel: posterior samples from the slab parameter τ^2. Lower panel: posterior samples from one of the $N = 2133$ synaptic weights. Note that many samples in the lower panel are zero, since the posterior inclusion probability for this weight was $p(s|Y) \simeq 0.48$.

4 Conclusion and Extensions

Our simulations on both toy and real neuronal geometries demonstrate the potential utility of our techniques for inferring the location and strength of synaptic connections when using both scan and compressed sensing observations. Numerical simulations indicate that LARS+ performs better than LARS or OLS and is able to learn the synaptic weights even under low SNR conditions. We close by noting that the basic model we have considered here can accommodate several possible extensions.

Robust observation model. We have considered so far only Gaussian noise, for both the dynamics and the observations. However, it is known that estimates based on a quadratic objective function can be very non-robust (i.e., sensitive to outliers). The standard generalization is to use a log-likelihood that does not grow quadratically, but instead flattens out towards linear growth as

the errors become large, as in the Huber loss function (Huber, 1964)

$$f(x) = \begin{cases} x^2/2 & \text{for } |x| < t\,, \\ t|x| - t^2/2 & \text{for } |x| > t\,. \end{cases} \tag{4.1}$$

Note that this loss function is convex; therefore, log-likelihoods chosen to be proportional to the negative of this loss function will be concave. More generally, if the observation log-likelihood is concave, then the inference method we have introduced remains tractable. To compute the optimal voltage path,

$$\hat{V}(W) \quad = \quad \arg\max_V \log p(V|Y,W)\,, \tag{4.2}$$

we can use Newton's method, where each step requires one call to the Low-Rank Block-Thomas algorithm discussed in Pnevmatikakis and Paninski (2012), which generalizes the method outlined in Appendix C. To estimate the weights, we can use a Laplace approximation to $\log p(Y|W)$, so for each λ we want to solve

$$\hat{W} \quad = \quad \arg\max_W \ \log p(Y|W) + \log p(W|\lambda) \tag{4.3}$$

$$\simeq \quad \arg\max_W \ \log p(\hat{V}(W)|W) + \log p(Y|\hat{V}(W)) - \frac{1}{2}\log|-H_{VV}| + \log p(W|\lambda)\,, \tag{4.4}$$

where the Hessian H_{VV}, defined in (A.6), is evaluated at $\hat{V}$. In the Gaussian case, the first two terms are quadratic in W, and H_{VV} is constant; see Huggins and Paninski (2012) for further details on the evaluation of the $\log|-H_{VV}|$ term. For most reasonable concave observation log-likelihoods, the solution $\hat{W}(\lambda)$ is continuous in λ, and therefore we use the same path-following idea exploited by LARS to efficiently compute the solution path $\hat{W}(\lambda)$. The coordinate-wise descent algorithm of Friedman et al. (2007) provides one efficient solution for finding an optimal W at a given λ value, given a previous solution at a nearby value of λ.

Slow synapses. A slow synapse corresponds in our model to the filtered arrival of the presynaptic signal U_t at several delayed times. We can incorporate such a scenario by modifying the dynamic equation (2.1) as

$$V_{t+dt} \quad = \quad AV_t + \sum_{p=0}^{D} W_p U_{t-p} + \epsilon_t,\ \epsilon_t \sim \mathcal{N}(0, \sigma^2 dt I) \tag{4.5}$$

where each W_p is a $N \times K$ synaptic weights matrix for the stimuli arriving with a delay of p time units. (Equivalently, it is possible to expand U_t in a different basis set of filter functions.) For this case, it is natural to modify the prior $p(W|\lambda)$ to require that all D weights at a given compartment be zero or non-zero jointly. This can be enforced with the *grouped lasso* prior (Lin and Zhang, 2006; Yuan and Lin, 2006),

$$\log p(W|\lambda) = -\lambda \sum_{i,j} \|W^{i,j}\|_2 \tag{4.6}$$

with $\|W^{i,j}\|_2 = \sqrt{\sum_{p=0}^{D} (W_p^{i,j})^2}$, which is known to encourage solutions for which the "group" of elements $W_p^{i,j}$ are held at zero (for a given i, j). Again, the coordinate-wise descent algorithm of Friedman et al. (2007) is applicable here.

More generally, it is worth noting that the l_1 (or group-l_1) prior we have used here is rather simple, and could be generalized considerably. In many cases we may have additional prior information that can be exploited statistically: for example, we may know from previous anatomical studies that

a given presynaptic cell type might prefer to synapse on the postsynaptic neuron at perisomatic but not distal dendritic locations. This can easily be incorporated here by varying the weight of the l_1 penalty in a compartment- and cell-type-dependent manner. See Mishchenko and Paninski (2012) for further discussion.

Other observation models. We can also incorporate more general observation models; for example, some voltage indicators have their own intrinsic dynamics (this is particularly relevant in the case of genetically-encoded indicators (Knopfel et al., 2006)), which can be incorporated into either the dynamics model (as an additional dynamical variable) or the observation model $p(Y|V)$; see Paninski (2010) for details. Another important direction for future work is to incorporate calcium measurements, which provide higher-SNR information about a nonlinearly-thresholded version of the voltage signal (Gobel and Helmchen, 2007; Larkum et al., 2008; Takahashi et al., 2012; Pnevmatikakis, Kelleher, Chen, Saggau, Josić and Paninski, 2012).

Non-linear effects. It will also be important to generalize our methods to increase their robustness to nonlinearities and other departures from the basic model (2.1)-(2.2). In particular, shunting effects may play a role for large synchronous inputs to nearby compartments, so it would be desirable to have conductance terms attached to each compartment (Paninski et al., 2012). Other important effects include synaptic depression and probabilistic release; these random, spike-history dependent terms can be handled in principle using Expectation-Maximization methods (Huys and Paninski, 2009), but further work will be necessary to ascertain the effectiveness of these methods in the context of the type of experimental data considered here.

A The quadratic function and the LARS+ algorithm

In this appendix we provide the details of the algorithm used to obtain the solution $\hat{W}^{ij}(\lambda)$ for all λ, where $i = 1 \dots N$ indicates the neuron compartment and $j = 1 \dots K$ the presynaptic stimulus associated with the weight. To simplify the notation, define

$$Q(W) \equiv \log p(Y|W)\,, \tag{A.1}$$
$$Q(W,V) \equiv \log p(Y,V|W)\,, \tag{A.2}$$

where these expressions are related by

$$p(Y|W) = \int p(Y,V|W)\,dV\,. \tag{A.3}$$

Let us first obtain an explicit expression for $Q(W)$. Recall from Section 2 that

$$\begin{aligned} Q(W,V) &= -\frac{1}{2}\sum_{t=1}^{T}(y_t - B_t V_t)^T C_y^{-1}(y_t - B_t V_t) \\ &\quad -\frac{1}{2}\sum_{t=2}^{T}(V_t - AV_{t-1} - WU_{t-1})^T C_V^{-1}(V_t - AV_{t-1} - WU_{t-1}) \\ &\quad -\frac{1}{2}V_1^T C_0^{-1} V_1 + const.\,. \end{aligned} \tag{A.4}$$

Since $Q(W,V)$ is quadratic and concave in V, we can expand it around its maximum $\hat{V}(W)$ as

$$Q(W,V) = Q(W,\hat{V}) + \frac{1}{2}(V - \hat{V}(W))^T H_{VV}(V - \hat{V}(W))\,, \tag{A.5}$$

where the $NT \times NT$ Hessian

$$H_{VV} = \frac{\partial^2 Q(W,V)}{\partial V \partial V}\,, \tag{A.6}$$

does not depend on Y or W, as is clear from (A.4). Inserting the expansion (A.5) in the integral (A.3) and taking the log, we get

$$Q(W) = Q(W,\hat{V}(W)) + c \tag{A.7}$$
$$= \log p(Y|\hat{V}(W)) + \log p(\hat{V}(W)|W) + c \tag{A.8}$$

where $c = -\frac{1}{2}\log|-H_{VV}| + \frac{TN}{2}\log 2\pi$ is independent of W.

Since $\hat{V}(W)$ is the maximum of $Q(W,V)$, its value is the solution of $\nabla_V Q(W,V) = 0$, given by

$$\hat{V}(W) = -H_{VV}^{-1} Z(W) \tag{A.9}$$

where

$$Z(W) = \nabla_V Q(W,V)|_{V=0} = \begin{pmatrix} B_1^T C_y^{-1} y_1 - A^T C_V^{-1} W U_1 \\ B_2^T C_y^{-1} y_2 - A^T C_V^{-1} W U_2 + C_V^{-1} W U_1 \\ B_3^T C_y^{-1} y_3 - A^T C_V^{-1} W U_3 + C_V^{-1} W U_2 \\ \vdots \end{pmatrix} \in \mathbb{R}^{NT}\,, \tag{A.10}$$

as follows from (A.4). It is useful to expand $Z(W)$ as

$$Z(W) = Z_0 + \sum_{i,j} Z_{ij} W^{ij} \tag{A.11}$$

where the coefficients $Z_0, Z_{ij} \in \mathbb{R}^{NT}$ can be read out from (A.10) and are independent of W. This in turn gives an expansion for $\hat{V}$ in (A.9) as

$$\hat{V}(W) = \hat{V}_0 + \sum_{i,j} \hat{V}_{ij} W^{ij} \quad \in \mathbb{R}^{NT} \tag{A.12}$$

where

$$\hat{V}_0 = -H_{VV}^{-1} Z_0 \quad \in \mathbb{R}^{NT} \tag{A.13}$$

$$\hat{V}_{ij} = -H_{VV}^{-1} Z_{ij} \quad \in \mathbb{R}^{NT} \tag{A.14}$$

are independent of W. Note that $\hat{V}_0$ has components

$$\hat{V}_0 = \begin{pmatrix} (\hat{V}_0)_1 \\ \vdots \\ (\hat{V}_0)_T \end{pmatrix} \tag{A.15}$$

where each $(\hat{V}_0)_t$ is an N-vector, and similarly for each $\hat{V}_{ij}$.

To obtain the explicit form of $Q(W)$ one can insert the expansion (A.12) for $\hat{V}(W)$ in (A.8). But it is easier to notice first, using the chain rule, that

$$\frac{dQ(W, \hat{V}(W))}{dW} = \frac{\partial Q(W, \hat{V}(W))}{\partial W} + \frac{\partial Q(W, \hat{V}(W))}{\partial \hat{V}} \frac{\partial \hat{V}(W)}{\partial W} \tag{A.16}$$

$$= \frac{\partial Q(W, \hat{V}(W))}{\partial W} \tag{A.17}$$

$$= C_V^{-1} \sum_{t=2}^{T} (\hat{V}_t - A\hat{V}_{t-1} - W U_{t-1}) U_{t-1}^T \tag{A.18}$$

where the second term in (A.16) is zero since $\hat{V}(W)$ is the maximum for any W. Thus once $\hat{V}$ is available, the gradient of Q w.r.t. W is easy to compute, since multiplication by the sparse cable dynamics matrix A is fast. We can now insert (A.12) into the much simpler expression (A.18) to get

$$\frac{dQ(W, \hat{V}(W))}{dW^{ij}} = r_{ij} + M_{ij,i'j'} W^{i'j'} \tag{A.19}$$

with $i, i' = 1 \dots N$ and $j, j' = 1 \dots K$ and coefficients

$$r_{ij} = \frac{1}{\sigma^2 dt} \sum_{t=2}^{T} \left((\hat{V}_0)_t - A(\hat{V}_0)_{t-1} \right)_i (U_{t-1})_j \tag{A.20}$$

$$M_{ij,i'j'} = \frac{1}{\sigma^2 dt} \sum_{t=2}^{T} \left[\left((\hat{V}_{i'j'})_t - A(\hat{V}_{i'j'})_{t-1} \right)_i (U_{t-1})_j - (U_{t-1})_j (U_{t-1})_{j'} \delta_{ii'} \right] \tag{A.21}$$

where $\delta_{ii'}$ is Kronecker's delta. The desired expression for $Q(W)$ follows by a simple integration of (A.19) and gives the quadratic expression

$$Q(W) = \sum_{i,j} r_{ij} W^{ij} + \frac{1}{2} W^{ij} M_{ij,i'j'} W^{i'j'} + const. \tag{A.22}$$

where $i = 1 \dots N$, $j = 1 \dots K$. Note that the costly step, computationally, is the linear matrix solve involving H_{VV} in (A.13)-(A.14) to obtain the components of $\hat{V}$, which are then used in (A.20)-(A.21) to obtain p_{ij} and $M_{ij,i'j'}$ in $O(T)$ time. Note that we do not need the explicit form of H_{VV}^{-1}, only its action on the vectors Z_0, Z_{ij}.

Matrix form of coefficients

For just one presynaptic signal ($K = 1$), we can express the coefficients of the log-likelihood (A.22) in a compact form by defining the matrices

$$P = \begin{pmatrix} -A & I_N & & & \\ & -A & I_N & & \\ & & & & \\ & & & -A & I_N \\ & & & & 0 \end{pmatrix} \in \mathbb{R}^{NT \times NT} \tag{A.23}$$

$$U = \begin{pmatrix} U_1 I_N & \cdots & U_{T-1} I_N & 0 \end{pmatrix} \in \mathbb{R}^{N \times NT} \tag{A.24}$$

$$B = \begin{pmatrix} B_1 & & & \\ & B_2 & & \\ & & . & \\ & & & B_T \end{pmatrix} \in \mathbb{R}^{ST \times NT} \tag{A.25}$$

and $C_{yT}^{-1} = C_y^{-1} I_{ST}$, where I_N and I_{ST} are identity matrices of the indicated dimensions. Using these matrices, the expansion (A.12) for the estimated voltages is

$$\hat{V}(W) = V_0 + \hat{V} W \tag{A.26}$$

with

$$V_0 = -H_{VV}^{-1} B^T C_{yT}^{-1} Y \quad \in \mathbb{R}^{NT} \tag{A.27}$$

$$\hat{V} = \begin{pmatrix} \hat{V}_1 \cdots \hat{V}_N \end{pmatrix} \tag{A.28}$$

$$= -H_{VV}^{-1} P^T U^T C_V^{-1} \quad \in \mathbb{R}^{NT \times N} . \tag{A.29}$$

where Y in (A.27) is

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_T \end{pmatrix} \tag{A.30}$$

The coefficients of the quadratic log-likelihood in (A.22) can now be expressed as

$$r = C_V^{-1} U P \hat{V}_0 \tag{A.31}$$

$$= -C_V^{-1} U P H_{VV}^{-1} B^T C_{yT}^{-1} Y \quad \in \mathbb{R}^N \tag{A.32}$$

and

$$M = C_V^{-1} U P \hat{V} - ||U||^2 C_V^{-1} \tag{A.33}$$

$$= -C_V^{-1} U P H_{VV}^{-1} P^T U^T C_V^{-1} - ||U||^2 C_V^{-1} \quad \in \mathbb{R}^{N \times N} \tag{A.34}$$

where we defined $||U||^2 = \sum_{t=1}^{T-1} U_t^2$. Note that this form makes evident that M is symmetric and negative semidefinite, which is not obvious in (A.21). In matrix form, the OLS solution is given by

$$\hat{W} = \arg\max_W W^T r + \frac{1}{2} W^T M W \tag{A.35}$$

$$= -M^{-1} r \tag{A.36}$$

$$= -\left(C_V^{-1} U U^T + C_V^{-1} U P H_{VV}^{-1} P^T U^T C_V^{-1} \right)^{-1} C_V^{-1} U P H_{VV}^{-1} B^T C_{yT}^{-1} Y \tag{A.37}$$

$$= -\frac{UP}{||U||^2} \left(\frac{P^T U^T C_V^{-1} U P}{||U||^2} + H_{VV} \right)^{-1} B^T C_{yT}^{-1} Y , \tag{A.38}$$

where in the last line we used the identity

$$(\mathbf{A}^{-1} + \mathbf{B}^T\mathbf{C}^{-1}\mathbf{B})^{-1}\mathbf{B}^T\mathbf{C}^{-1} = \mathbf{A}\mathbf{B}^T(\mathbf{B}\mathbf{A}\mathbf{B}^T + \mathbf{C})^{-1}\,. \tag{A.39}$$

A.1 LARS-lasso

We will restate here the LARS-lasso algorithm from Efron et al. (2004) for a generic concave quadratic function $Q(W)$. We are interested in solving[2]

$$\hat{W}(\lambda) = \arg\max_W L(W,\lambda) \tag{A.40}$$

where

$$L(W,\lambda) = Q(W) - \lambda\sum_{i=1}^{N}|W^i|\,. \tag{A.41}$$

As we saw in eq.(2.11), the solution for $\hat{W}$ is a piecewise linear function of λ, with components becoming zero or non-zero at the breakpoints.

As a function of W^i, $L(W,\lambda)$ is differentiable everywhere except at $W^i = 0$. Therefore, if W^i is non-zero at the maximum of $L(W,\lambda)$, it follows that

$$\frac{dL(W,\lambda)}{dW^i} = 0 \qquad \text{for} \qquad W^i \neq 0\,, \tag{A.42}$$

or equivalently

$$\nabla_i Q(W) = r_i + M_{i,i'}W^{i'} = \lambda\, sign(W^i) \qquad \text{for} \qquad W^i \neq 0\,, \tag{A.43}$$

which implies

$$|\nabla_i Q(W)| = \lambda \qquad \text{for} \qquad W^i \neq 0\,. \tag{A.44}$$

For $\lambda = \infty$, one can ignore the first term in (A.41), so the solution to (A.40) is clearly $W^i = 0$. One can show that this holds for all $\lambda > \lambda_1$, where

$$\lambda_1 = \max_i |\nabla_i Q|_{W=0} = \max_i |r_i| \tag{A.45}$$

Suppose, without loss of generality, that the maximum in (A.45) occurs for $i = 1$. The condition (A.43) will now be satisfied for non-zero W^1, so we decrease λ and let W^1 change as

$$\lambda = \lambda_1 - \gamma \tag{A.46}$$

$$W^1(\gamma) = \gamma a^1 \qquad \gamma \in [0,\lambda_1] \tag{A.47}$$

while the other W^is are kept to zero. To find a^1, insert (A.47) in (A.43),

$$r_1 + M_{11}\gamma a^1 = (\lambda_1 - \gamma)\, sign(a^1)\,, \tag{A.48}$$

from which we get $a_1 = -r_1/(\lambda_1 M_{11})$. Proceeding in this way, and denoting by $\mathbf{W}_p(\gamma)$ the vector of weights after the p-th breakpoint, in general we will have, after p steps

$$\lambda = \lambda_p - \gamma \tag{A.49}$$

$$\mathbf{W}_p(\gamma) = \text{ linear in } \gamma \text{ with } k \leq p \text{ non-zero components}, \tag{A.50}$$

$$|\nabla_i Q(\mathbf{W}_p(\gamma'))| = \lambda_p - \gamma \qquad i = 1\ldots k \qquad \text{non-zero directions}, \tag{A.51}$$

$$|\nabla_{i'} Q(\mathbf{W}_p(\gamma'))| < \lambda_p - \gamma \qquad i' > k \qquad \text{zero directions}, \tag{A.52}$$

and we let γ grow until either of these conditions occurs:

[2] We omit from here on the indices j, j' in W^{ij} and $M_{ij,i'j'}$ to simplify the notation.

1. At $\gamma = \gamma'$ the gradient along a zero direction, say W^{k+1}, satisfies

$$|\nabla_{k+1} Q(\mathbf{W}_p(\gamma'))| = \lambda_p - \gamma' \,. \tag{A.53}$$

If this happens we let W^{k+1} become active. Define

$$\begin{aligned} \mathbf{W}_p &\equiv \mathbf{W}_p(\gamma') \,, & \text{(A.54)} \\ \lambda_{p+1} &= \lambda_p - \gamma' \,, & \text{(A.55)} \end{aligned}$$

and continue with $k+1$ components as:

$$\mathbf{W}_{p+1}(\gamma) \equiv \mathbf{W}_p + \gamma \mathbf{a} = \begin{pmatrix} W_p^1 \\ \vdots \\ W_p^k \\ 0 \end{pmatrix} + \gamma \begin{pmatrix} a^1 \\ \vdots \\ a^k \\ a^{k+1} \end{pmatrix} \qquad \gamma \in [0, \lambda_{p+1}] \tag{A.56}$$

$$\lambda = \lambda_{p+1} - \gamma \tag{A.57}$$

To find the new velocity $\mathbf{a}$, insert $\mathbf{W}_{p+1}(\gamma)$ into (A.43) to get

$$\begin{pmatrix} M_{11} \dots M_{1(k+1)} \\ \vdots \\ \vdots \\ M_{(k+1)1} \dots M_{(k+1)(k+1)} \end{pmatrix} \begin{pmatrix} a^1 \\ \vdots \\ a^k \\ a^{k+1} \end{pmatrix} = - \begin{pmatrix} sign(W_p^1) \\ \vdots \\ sign(W_p^k) \\ sign(a^{k+1}) \end{pmatrix} \tag{A.58}$$

In this equation we need $sign(a^{k+1})$, which, as we show in Section A.3, coincides with that of the derivative computed in (A.53),

$$sign(a^{k+1}) = sign(\nabla_{k+1} Q(\mathbf{W}_p(\gamma'))) \,. \tag{A.59}$$

2. A component of $\mathbf{W}_p(\gamma)$, say W^k, becomes zero at $\gamma = \gamma'$. (A.60)

If this happens, W^k must drop from the active set because the path of $\mathbf{W}_p(\gamma)$ was obtained assuming a definite sign for W^k in (A.43). So we define

$$\begin{aligned} \mathbf{W}_p &= \mathbf{W}_p(\gamma') \,, & \text{(A.61)} \\ \lambda_{p+1} &= \lambda_p - \gamma' \,, & \text{(A.62)} \end{aligned}$$

drop W^k from the active set and continue with $k-1$ active components as:

$$\mathbf{W}_{p+1}(\gamma) \equiv \mathbf{W}_p + \gamma \mathbf{a} = \begin{pmatrix} W_p^1 \\ \vdots \\ W_p^{k-1} \end{pmatrix} + \gamma \begin{pmatrix} a^1 \\ \vdots \\ a^{k-1} \end{pmatrix} \qquad \gamma \in [0, \lambda_{p+1}] \tag{A.63}$$

$$\lambda = \lambda_{p+1} - \gamma \tag{A.64}$$

To find the new $\mathbf{a}$, inserting $\mathbf{W}_{p+1}(\gamma)$ into (A.43) gives

$$\begin{pmatrix} M_{11} \dots M_{1(k-1)} \\ \vdots \\ M_{(k-1)1} \dots M_{(k-1)(k-1)} \end{pmatrix} \begin{pmatrix} a^1 \\ \vdots \\ a^{k-1} \end{pmatrix} = - \begin{pmatrix} sign(W_p^1) \\ \vdots \\ sign(W_p^{k-1}) \end{pmatrix} \tag{A.65}$$

from which $\mathbf{a}$ can be solved.

As each **a** is found, we decrease λ by increasing γ, and check again for either cases 1 or 2 until we reach $\lambda = 0$, at which point all directions will be active and the weights will correspond to the global maximum of $Q(W)$.

Having presented the algorithm, let us discuss its computational cost. To obtain p_i we need to act with H_{VV}^{-1} on Z_0 (see (A.13) and (A.20)). Similarly, for each new active weight W^{k+1} the $(k+1)$-th column of M is needed in (A.58), which comes from acting with H_{VV}^{-1} on Z_{k+1} (see (A.14) and (A.21)). The action of H_{VV}^{-1} has a runtime of $O(TN^3)$, but in Appendix C we show how to reduce it to $O(TNS^2)$ with a low-rank approximation. For the total computational cost, we have to add the runtime of solving (A.58). Since at each breakpoint the matrix in the left-hand side of (A.58) only changes by the addition of the $(k+1)$th row and column, the solution takes $O(k^2)$ instead of $O(k^3)$ (Efron et al., 2004). Running the LARS algorithm through k steps, the total cost is then $O(kTNS^2 + k^3)$ time.

A.2 Enforcing a sign for the inferred weights

We can enforce a definite sign for the non-zero weights by a simple modification of the LARS-lasso. Assuming for concreteness an excitatory synapse, the solution to (A.40) for all λ and subject to

$$W^i \geq 0$$

can be obtained by allowing a weight to become active only if its value along the new direction is positive. The enforcement of this condition for the linear regression case was discussed in Efron et al. (2004). In our formulation of the LARS-lasso algorithm, the positivity can be enforced by requiring that the first weight becomes active when

$$\lambda_1 = \max_i r_i \qquad r_i > 0 \tag{A.66}$$

and by replacing the condition that triggers the introduction of new active weights, denoted above as condition 1, by

1. At $\gamma = \gamma'$ the gradient along a zero direction, say W^{k+1}, satisfies

$$\nabla_{k+1} Q(\mathbf{W}_p(\gamma')) = +(\lambda_p - \gamma')\,. \qquad \gamma' \in [0, \lambda_p] \tag{A.67}$$

By requiring the derivative along W^{k+1} to be positive at the moment of joining the active set, we guarantee that W^{k+1} will be positive due to the result of Section A.3.

When λ reaches zero, the weights, some of which may be zero, are the solution to the quadratic program

$$\hat{W} = \arg\max_W Q(W)\,, \qquad W^i \geq 0\,. \tag{A.68}$$

We will refer to the LARS-lasso algorithm with the modification (A.67) as LARS+. In practice, the measurements can be so noisy that the algorithm may have to be run assuming both non-negative and non-positive weights, and the nature of the synapse can be established by comparing the likelihood of both results at their respective maxima. More generally, if $K > 1$ we have to estimate the sign of each presynaptic neuron; this can be done by computing the likelihoods for each of the 2^K possible sign configurations. This exhaustive approach is tractable since we are focusing here on the small-K setting; for larger values of K, approximate greedy approaches may be necessary Mishchenko et al. (2011).

A.3 The sign of a new active variable

Property: the sign of a new variable W^{k+1} which joins the active group is the sign of $\nabla_{k+1}Q(W)$ at the moment of joining.

Proof:[3] Remember that the matrix $M_{ii'}$ is negative definite and, in particular, its diagonal elements are negative

$$M_{ii} < 0\,, \qquad i = 1 \dots N \tag{A.69}$$

As we saw in Section A.1, if the first variable to become active is

$$\mathbf{W}_1(\gamma) \;=\; \gamma a^1 \qquad\qquad \gamma \in [0, \lambda_1] \tag{A.70}$$

with

$$\lambda_1 \;=\; \max_i |\nabla_i Q|_{W=0} = |r_1|\,, \tag{A.71}$$

we have

$$a_1 = -\frac{r_1}{\lambda_1 M_{11}} \tag{A.72}$$

and using (A.69) and $\lambda_1 > 0$ we get

$$sgn(a_1) = sgn(r_1) \tag{A.73}$$

as claimed. Suppose now that there are k active coordinates and our solution is

$$\mathbf{W}_p(\gamma) = \begin{pmatrix} W_p^1(\gamma) \\ \cdot \\ \cdot \\ W_p^k(\gamma) \end{pmatrix} \qquad\qquad \gamma \in [0, \lambda_p] \tag{A.74}$$

Define

$$c_j(\gamma) = \nabla_j Q(\mathbf{W}_p(\gamma))\,, \tag{A.75}$$

and note that

$$|c_j(\gamma)| = \lambda_p - \gamma \qquad j = 1 \dots k\,. \tag{A.76}$$

Suppose a new variable W^{k+1} enters the active set at $\gamma = \gamma'$ such that

$$|c_{k+1}(\gamma')| = \lambda_p - \gamma' \tag{A.77}$$

It is easy to see that when taking γ all the way to λ_p, the sign of $c_{k+1}(\gamma)$ does not change

$$sgn(c_{k+1}(\gamma')) = sgn(c_{k+1}(\lambda_p)) \tag{A.78}$$

since the $c_j(\gamma)$ $(j = 1, \dots k)$ go faster towards zero than $c_{k+1}(\gamma)$. To make the variable W^{k+1} active, define

$$\lambda_{p+1} \;=\; \lambda_p - \gamma'\,, \tag{A.79}$$

$$\mathbf{W}_p \;\equiv\; \mathbf{W}_p(\gamma')\,, \tag{A.80}$$

and continue with $k+1$ components as:

[3] This is a recasting of Lemma 4 in Efron et al. (2004).

$$\mathbf{W}_{p+1}(\gamma) \equiv \mathbf{W}_p + \gamma\mathbf{a} = \begin{pmatrix} W_p^1 \\ \vdots \\ W_p^k \\ 0 \end{pmatrix} + \gamma \begin{pmatrix} a^1 \\ \vdots \\ a^k \\ a^{k+1} \end{pmatrix} \qquad \gamma \in [0, \lambda_{p+1}] \tag{A.81}$$

$$\lambda = \lambda_{p+1} - \gamma \tag{A.82}$$

To find $\mathbf{a}$, impose on (A.81) the conditions (A.43) that give

$$\mathbf{p} + \mathbf{M}_{(k+1,k+1)}\mathbf{W}_{\mathbf{p+1}}(\gamma) = (\lambda_{p+1} - \gamma)\mathbf{s} \tag{A.83}$$

where $\mathbf{p} = (p_1, \ldots, p_{k+1})^T$, $\mathbf{M}_{(k+1,k+1)}$ is the $(k+1) \times (k+1)$ submatrix of M_{ij}, and

$$\mathbf{s} = \begin{pmatrix} sgn(W_p^1) \\ . \\ . \\ sgn(W_p^k) \\ sgn(a^{k+1}) \end{pmatrix} \tag{A.84}$$

Since (A.83) holds for any γ, we get the two equations

$$\mathbf{p} + \mathbf{M}_{(k+1,k)}\mathbf{W}_p = \lambda_{p+1}\mathbf{s} \tag{A.85}$$

and

$$\mathbf{M}_{(k+1,k+1)}\mathbf{a} = -\mathbf{s}\,. \tag{A.86}$$

where $\mathbf{M}_{(k+1,k)}$ is obtained from $\mathbf{M}_{(k+1,k+1)}$ by eliminating the last column. Inserting (A.85) into (A.86) we get

$$\mathbf{a} = -\frac{1}{\lambda_{p+1}}\mathbf{M}^{-1}_{(k+1,k+1)}(\mathbf{p} + \mathbf{M}_{(k+1,k)}\mathbf{W}_p) \tag{A.87}$$

$$= -\frac{1}{\lambda_{p+1}}\mathbf{M}^{-1}_{(k+1,k+1)}\left(\mathbf{p} + \mathbf{M}_{(k+1,k)}\mathbf{W}_p(\lambda_p) - \mathbf{M}_{(k+1,k)}\mathbf{W}_p(\lambda_p) + \mathbf{M}_{(k+1,k)}\mathbf{W}_p\right) \tag{A.88}$$

$$= -\frac{1}{\lambda_{p+1}}\mathbf{M}^{-1}_{(k+1,k+1)}\begin{pmatrix} \mathbf{0} \\ c_{k+1}(\lambda_p) \end{pmatrix} - \frac{1}{\lambda_{p+1}}\left(\mathbf{W}_p - \mathbf{W}_p(\lambda_p)\right) \tag{A.89}$$

where $\mathbf{0}$ has k elements. Since the $(k+1)$-th element of the second term in (A.89) is zero, we get

$$a^{k+1} = -\frac{\left(\mathbf{M}^{-1}_{(k+1,k+1)}\right)_{(k+1)(k+1)}}{\lambda_{p+1}}\, c_{k+1}(\lambda_p)\,. \tag{A.90}$$

Since $\mathbf{M}^{-1}_{(k+1,k+1)}$ is negative definite, we have $(\mathbf{M}^{-1}_{(k+1,k+1)})_{(k+1)(k+1)} < 0$, so using (A.78), the result

$$sgn(a^{k+1}) = sgn(c_{k+1}(\gamma')) \tag{A.91}$$

follows. □

B The C_p criterion for low SNR

In the limit of very low signal-to-noise ratio, we can ignore the dynamic noise term in eq. (2.1) and consider

$$V_{t+dt} = AV_t + WU_t \tag{B.1}$$

$$y_t = B_tV_t + \eta_t, \qquad \eta_t \sim \mathcal{N}(0, C_yI). \tag{B.2}$$

Let us assume that the number of presynaptic neurons is $K = 1$ to simplify the formulas. The results can be easily extended to the general case. We can combine the above equations as

$$Y = XW + \eta\,, \tag{B.3}$$

where we defined

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_T \end{pmatrix} \qquad \eta = \begin{pmatrix} \eta_1 \\ \vdots \\ \eta_T \end{pmatrix} \tag{B.4}$$

and the matrix X is given by the product

$$X = BC \qquad \in \mathbb{R}^{ST\times N}\,, \tag{B.5}$$

where B was defined in (A.25) and

$$C = \begin{pmatrix} 0 \\ U_1 \\ AU_1 + U_2 \\ A^2U_1 + AU_2 + U_3 \\ . \\ . \\ A^{T-2}U_1 + \cdots + U_{T-1} \end{pmatrix} \qquad \in \mathbb{R}^{NT\times N}\,. \tag{B.6}$$

Equation (B.3) corresponds to a standard linear regression problem and the l_1-penalized posterior log-likelihood to maximize is now

$$\log p(W|Y,\lambda) = -\frac{1}{2}||Y - XW||^2 - \lambda\sum_{i=1}^{N}|W^i|\,. \tag{B.7}$$

The solution $\hat{W}(\lambda)$ that maximizes (B.7) is obtained, as in the general case, using the LARS/LARS+ algorithm, and the fitted observations are given by

$$\hat{Y}(\lambda) \quad = \quad BC\hat{W}(\lambda)\,. \tag{B.8}$$

One can show that each row in $C\hat{W}(\lambda)$ corresponds to the $C_V \to 0$ limit of the expected voltage $\hat{V}_t(\lambda)$ defined in (2.12). Given an experiment (Y, U), consider the training error

$$\mathrm{err}(\lambda) = ||Y - \hat{Y}(\lambda)||^2 \tag{B.9}$$

and the in-sample error

$$\mathrm{Err}_{in}(\lambda) = \mathbb{E}_{\tilde{Y}}\left[||\tilde{Y} - \hat{Y}(\lambda)||^2\right]\,. \tag{B.10}$$

In $\mathrm{Err}_{in}(\lambda)$, we compute the expectation over new observations $\tilde{Y}$ for the same stimuli U_t and compare them to the predictions $\hat{Y}(\lambda)$ obtained with the initial experiment (Y, U). Thus, $\mathrm{Err}_{in}(\lambda)$ gives a measure of the generalization error of our results. $\mathrm{Err}_{in}(\lambda)$ itself cannot be computed directly, but we can compute its expectation with respect to the original observations Y. For this, let us consider first the difference between Err_{in} and err, called the *optimism* (Friedman et al., 2008). Denoting the components of Y with an index i, it is easy to verify that the expected optimism with respect to Y is

$$\omega(\lambda) \quad \equiv \quad \langle \mathrm{Err}_{in}(\lambda) - \mathrm{err}(\lambda)\rangle \tag{B.11}$$

$$= \quad 2\sum_{i=1}^{ST}\langle Y_i\hat{Y}_i(\lambda)\rangle - \langle Y_i\rangle\langle\hat{Y}_i(\lambda)\rangle \tag{B.12}$$

$$= \quad 2\sum_{i=1}^{ST}\mathrm{Cov}(Y_i, \hat{Y}_i(\lambda))\,. \tag{B.13}$$

For the general case $K \geq 1$, we will have $X \in \mathbb{R}^{ST \times NK}$. Let us assume that $ST > NK$ and that X is full rank, that is, $\text{rank}(X) = NK$. Then in (Zou et al., 2007) it was shown that if we define $d(\lambda) = ||\hat{W}(\lambda)||_0$ as the number of non-zero components in $\hat{W}(\lambda)$, we have[4]

$$\omega(\lambda) = 2\langle d(\lambda) \rangle \, C_y. \tag{B.14}$$

Thus $2d(\lambda)\, C_y$ is an unbiased estimate of $\omega(\lambda)$, and is also consistent (Zou et al., 2007). With this result, and using $\text{err}(\lambda)$ as an estimate of $\langle \text{err}(\lambda) \rangle$, we obtain an estimate of the average generalization error $\langle \text{Err}_{in}(\lambda) \rangle$ as

$$C_p(\lambda) = ||Y - \hat{Y}(\lambda)||^2 + 2d(\lambda)\, C_y \, . \tag{B.15}$$

This quantity can be used to select the best λ as that value that minimizes $C_p(\lambda)$. Since the first term is a non-decreasing function of λ (Zou et al., 2007), it is enough to evaluate $C_p(\lambda)$ for each d at the smallest value of λ at which there are d active weights in $W(\lambda)$. With a slight abuse of notation, the resulting set of discrete values of (B.15) will be denoted as $C_p(d)$.

C The low-rank block-Thomas algorithm

In this appendix we will present a fast approximation technique to perform multiplications by the inverse Hessian H_{VV}^{-1}. The $NT \times NT$ Hessian H_{VV} in (A.6) takes the block-tridiagonal form

$$\begin{aligned} H_{VV} \;=\; & \begin{pmatrix} -C_0^{-1} - A^T A & A^T & \mathbf{0} & \dots & \\ A & -I - A^T A & A^T & \mathbf{0} & \dots \\ \mathbf{0} & A & -I - A^T A & A^T & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \\ & \dots & & A & -I \end{pmatrix} \\ & - \begin{pmatrix} B_1^T C_y^{-1} B_1 & & & \\ & B_2^T C_y^{-1} B_2 & & \\ & & & \\ & & & B_T^T C_y^{-1} B_T \end{pmatrix} \end{aligned} \tag{C.1}$$

where we have set $C_V = I$ to simplify the notation. We will restore it below to a generic value.

It will be convenient, following Paninski (2010), to adopt for C_0, the covariance of the initial voltage V_1, the value

$$C_0 = \sum_{i=0}^{\infty} (AA^T)^i = (I - AA^T)^{-1} \tag{C.2}$$

(note that the dynamics matrix A is stable here, ensuring the convergence of this infinite sum). This is the stationary prior covariance of the voltages V_t in the absence of observations y_t, and with this value for C_0, the top left entry in the first matrix in (C.1) simplifies to $-C_0^{-1} - A^T A = -I$.

We want to calculate

$$H_{VV}^{-1}\mathbf{b} = H_{VV}^{-1} \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_T \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_T \end{pmatrix} = \mathbf{x}, \tag{C.3}$$

where $\mathbf{b}$ can be an arbitrary NT-dimensional vector and each b_i and x_i is a column vector with length N. We can calculate this using the block Thomas algorithm for tridiagonal systems of equations (Press et al., 1992), which in general requires $O(N^3 T)$ time and $O(N^2 T)$ space, as shown in Algorithm 1.

[4]We have verified, through Monte Carlo simulations similar to those in (Zou et al., 2007), that this result also holds in the positive constrained case.

Algorithm 1 Standard Block Thomas Algorithm for calculating $H_{VV}^{-1}\mathbf{b}$

$\alpha_1 \leftarrow -(I + B_1^T C_y^{-1} B_1)$
$\gamma_1 \leftarrow \alpha_1^{-1} A^T$
$y_1 \leftarrow \alpha_1^{-1} b_1$
for $i = 2$ to $T-1$ **do**
  $\alpha_i \leftarrow -(I + A^T A + B_i^T C_y^{-1} B_i + A\gamma_{i-1})$
  $\gamma_i \leftarrow \alpha_i^{-1} A^T$
  $y_i \leftarrow \alpha_i^{-1}(b_i - A y_{i-1})$
end for
$\alpha_T \leftarrow -(I + B_T^T C_y^{-1} B_T + A\gamma_{T-1})$
$x_T \leftarrow \alpha_T^{-1}(b_T - A y_{T-1})$
for $i = T-1$ to 1 **do**
  $x_i \leftarrow y_i - \gamma_i x_{i+1}$
end for

We can adapt this algorithm to yield an approximate solution to (C.3) in $O(TNS^2)$ time by using low-rank perturbation techniques similar to those used in Paninski (2010); Huggins and Paninski (2012); Pnevmatikakis, Paninski, Rad and Huggins (2012). The first task is to calculate α_1^{-1}. Using the Woodbury matrix lemma, we get

$$\begin{aligned} \alpha_1^{-1} &= -(I + B_1^T C_y^{-1} B_1)^{-1} && \text{(C.4)} \\ &= -I + B_1^T (C_y + B_1 B_1^T)^{-1} B_1 && \text{(C.5)} \\ &= -I + L_1 D_1 L_1^T \quad \in \mathbb{R}^{N\times N} && \text{(C.6)} \end{aligned}$$

where

$$L_1 = B_1^T \quad \in \mathbb{R}^{N\times S} \tag{C.7}$$

and

$$D_1 = (C_y + B_1 B_1^T)^{-1} \quad \in \mathbb{R}^{S\times S}. \tag{C.8}$$

Note that the simple expression (C.6) for α_1^{-1} follows from the form we chose in (C.2) for C_0. Plugging α_1^{-1} into the Algorithm 1's expression for γ_1 gives

$$\begin{aligned} \gamma_1 &= \alpha_1^{-1} A^T && \text{(C.9)} \\ &= -A^T + L_1 D_1 L_1^T A^T \quad \in \mathbb{R}^{N\times N}. && \text{(C.10)} \end{aligned}$$

To continue the recursion for the other α_i^{-1}s, the idea is to approximate these matrices as low-rank perturbations to $-I$,

$$\alpha_i^{-1} \approx -I + L_i D_i L_i^T \quad \in \mathbb{R}^{N\times N}, \tag{C.11}$$

where D_i is a small $d_i \times d_i$ matrix with $d_i \ll N$ and $L_i \in \mathbb{R}^{N\times d_i}$. This in turn leads to a form similar to (C.10) for γ_i,

$$\gamma_i \approx -A^T + L_i D_i L_i^T A^T. \tag{C.12}$$

Therefore we can write

$$\begin{aligned} \alpha_i^{-1} &= -(I + A^T A + B_i^T C_y^{-1} B_i + A\gamma_{i-1})^{-1} && \text{(C.13)} \\ &\approx -(I + A^T A + B_i^T C_y^{-1} B_i - AA^T + AL_{i-1} D_{i-1} L_{i-1}^T A^T)^{-1} && \text{(C.14)} \\ &\approx -(I + B_i^T C_y^{-1} B_i + AL_{i-1} D_{i-1} L_{i-1}^T A^T)^{-1}. && \text{(C.15)} \end{aligned}$$

This expression justifies our approximation of α_i^{-1}s as a low rank perturbation to $-I$: the term $B_i^T C_y^{-1} B_i$ is low rank because the number of measurements is $S \ll N$, and the second term is low rank because the condition $eigs(A) < 1$ tends to suppress at step i the contribution of the previous step encoded in $L_{i-1} D_{i-1} L_{i-1}^T$. See Pnevmatikakis, Paninski, Rad and Huggins (2012) for details.

To apply Woodbury we choose a basis for the two non-identity matrices,

$$O_i = [AL_{i-1} \quad B_i^T] \in \mathbb{R}^{N \times (S+d_{i-1})} \tag{C.16}$$

and write

$$B_i^T C_y^{-1} B_i + AL_{i-1} D_{i-1} L_{i-1}^T A^T = O_i M_i O_i^T, \tag{C.17}$$

where

$$M_i = \begin{pmatrix} D_{i-1} & \\ & C_y^{-1} \end{pmatrix} \in \mathbb{R}^{(S+d_{i-1}) \times (S+d_{i-1})}$$

Applying Woodbury gives

$$\begin{aligned} \alpha_i^{-1} &= -(I + O_i M_i O_i^T)^{-1} && \text{(C.18)} \\ &= -I + O_i (M_i^{-1} + O_i^T O_i)^{-1} O_i^T. && \text{(C.19)} \end{aligned}$$

We obtain L_i and D_i by truncating the SVD of the expression on the right-hand side: in Matlab, for example, do

$$[L', D'] = svd(O_i (M_i^{-1} + O_i^T O_i)^{-1/2}, \text{'econ'}), \tag{C.20}$$

then choose L_i as the first d_i columns of L' and D_i as the square of the first d_i diagonal elements D', where d_i is chosen to be large enough (for accuracy) and small enough (for computational tractability).

We must handle α_T^{-1} slightly differently because of the boundary condition. Making use of the fact that $C_0^{-1} = I - AA^T$ and the Woodbury identity, we get

$$\begin{aligned} \alpha_T^{-1} &= -(I + B_T^T C_y^{-1} B_T + A\gamma_{T-1})^{-1} && \text{(C.21)} \\ &= -(I + B_T^T C_y^{-1} B_T - AA^T + AL_{T-1} D_{T-1} L_{T-1}^T A^T)^{-1} && \text{(C.22)} \\ &= -(C_0^{-1} + O_T M_T O_T^T)^{-1} && \text{(C.23)} \\ &= -C_0 + C_0 O_T (M_T^{-1} + O_T^T C_0 O_T)^{-1} O_T^T C_0 && \text{(C.24)} \\ &= -C_0 + L_T D_T L_T^T, && \text{(C.25)} \end{aligned}$$

where

$$L_T = C_0 O_T \tag{C.26}$$

and

$$D_T = (M_T^{-1} + O_T^T C_0 O_T)^{-1}. \tag{C.27}$$

Multiplications by α_T^{-1} are efficient since we can multiply by C_0 in $O(N)$ time, expoiting the sparse structure of A (see (Paninski, 2010) for details). It is unnecessary to control the rank because we will only be performing one multiplication with α_T^{-1} and calculating the SVD is a relatively expensive operation.

The updates for calculating y_i and x_i are straightforward:

$$\begin{aligned} y_1 &= \alpha_1^{-1} b_1 && \text{(C.28)} \\ &= -b_1 + L_1 D_1 L_1^T b_1 && \text{(C.29)} \\ y_i &= \alpha_i^{-1}(b_i - C_V^{-1} A y_{i-1}) && \text{(C.30)} \\ &= (-I + L_i D_i L_i^T)(b_i - A y_{i-1}) && \text{(C.31)} \\ x_T &= \alpha_T^{-1}(b_T - C_V^{-1} A y_{T-1}) && \text{(C.32)} \\ &= (-C_0 + L_T D_T L_T^T)(b_T - A y_{T-1}) && \text{(C.33)} \\ x_i &= y_i - \gamma_i x_{i+1} && \text{(C.34)} \\ &= y_i + A^T x_{i+1} - L_i D_i L_i^T A^T x_{i+1}. && \text{(C.35)} \end{aligned}$$

Algorithm 2 summarizes the full procedure. One can verify that the total computational cost scales like $O(TNS^2)$ (see Pnevmatikakis, Paninski, Rad and Huggins (2012) for details).

Finally, note that for repeated calls to $H_{VV}^{-1}\mathbf{b}$, we can compute the matrices L_i, D_i once and store them. For the case when C_V is not the identity we can apply a linear whitening change of variables $V_t' = C_V^{-1/2} V_t$. We solve as above except we make the substitution $B_t \to B_t C_V^{1/2}$ and our final solution now has the form

$$\mathbf{x} = \left(I_T \otimes C_V^{1/2}\right) H_{VV}^{-1} \left(I_T \otimes C_V^{1/2}\right)^T \mathbf{b}.$$

Algorithm 2 Low Rank Block Thomas Algorithm for calculating $H_{VV}^{-1}\mathbf{b}$

$L_1 \leftarrow B_1^T$
$D_1 \leftarrow (C_y + B_1 B_1^T)^{-1}$
$y_1 \leftarrow -b_1 + L_1 D_1 L_1^T b_1$
for $i = 2$ to T **do**
  $O_i \leftarrow [A L_{i-1} \quad B_i^T]$
  $M_i \leftarrow \text{diag}(D_{i-1}, C_y^{-1})$
  if $i \neq T$ **then**
    $[L_i', D_i'] \leftarrow svd(O_i(M_i^{-1} + O_i^T O_i)^{-1/2}, \text{`econ'})$
    $y_i \leftarrow (-I + L_i' D_i' L_i'^T)(b_i - A y_{i-1})$
    control rank of L_i' and D_i' to obtain L_i and D_i
  else
    $L_i \leftarrow C_0 O_i$
    $D_i \leftarrow (M_i^{-1} + O_i^T C_0 O_i)^{-1}$
  end if
end for
$x_T \leftarrow (-C_0 + L_T D_T L_T^T)(b_T - A y_{T-1})$
for $i = T - 1$ to 1 **do**
  $x_i \leftarrow y_i + A^T x_{i+1} - L_i D_i L_i^T A^T x_{i+1}$
end for

Acknowledgements

This work was supported by an NSF CAREER grant, a McKnight Scholar award, and by NSF grant IIS-0904353. This material is based upon work supported by, or in part by, the U. S. Army Research Laboratory and the U. S. Army Research Office under contract number W911NF-12-1-0594. JHH was partially supported by the Columbia College Rabi Scholars Program. AP was partially supported by the Swartz Foundation. The computer simulations were done in the Hotfoot HPC Cluster of Columbia University. We thank E. Pnevmatikakis for helpful discussions and comments.

References

Barbour, B., Brunel, N., Hakim, V. and Nadal, J.-P. (2007), 'What can we learn from synaptic weight distributions?', *TRENDS in Neurosciences* **30**(12), 622–629.

Bloomfield, S. and Miller, R. (1986), 'A functional organization of ON and OFF pathways in the rabbit retina', *J. Neurosci.* **6**(1), 1–13.

Candes, E., Romberg, J. and Tao, T. (2006), 'Stable signal recovery from incomplete and inaccurate measurements', *Communications on pure and applied mathematics* **59**(8), 1207–1223.

Candès, E. and Wakin, M. (2008), 'An introduction to compressive sampling', *Signal Processing Magazine, IEEE* **25**(2), 21–30.

Canepari, M., Djurisic, M. and Zecevic, D. (2007), 'Dendritic signals from rat hippocampal CA1 pyramidal neurons during coincident pre- and post-synaptic activity: a combined voltage- and calcium-imaging study', *J Physiol* **580**(2), 463–484.

Canepari, M., Vogt, K. and Zecevic, D. (2008), 'Combining voltage and calcium imaging from neuronal dendrites', *Cellular and Molecular Neurobiology* **28**, 1079–1093.

Djurisic, M., Antic, S., Chen, W. R. and Zecevic, D. (2004), 'Voltage imaging from dendrites of mitral cells: EPSP attenuation and spike trigger zones', *J. Neurosci.* **24**(30), 6703–6714.

Djurisic, M., Popovic, M., Carnevale, N. and Zecevic, D. (2008), 'Functional structure of the mitral cell dendritic tuft in the rat olfactory bulb', *J. Neurosci.* **28**(15), 4057–4068.

Dombeck, D. A., Blanchard-Desce, M. and Webb, W. W. (2004), 'Optical recording of action potentials with second-harmonic generation microscopy', *J. Neurosci.* **24**(4), 999–1003.

Durbin, J., Koopman, S. and Atkinson, A. (2001), *Time series analysis by state space methods*, Vol. 15, Oxford University Press Oxford.

Efron, B. (2004), 'The estimation of prediction error', *Journal of the American Statistical Association* **99**(467), 619–632.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004), 'Least angle regression', *Annals of Statistics* **32**, 407–499.

Fisher, J. A. N., Barchi, J. R., Welle, C. G., Kim, G.-H., Kosterin, P., Obaid, A. L., Yodh, A. G., Contreras, D. and Salzberg, B. M. (2008), 'Two-photon excitation of potentiometric probes enables optical recording of action potentials from mammalian nerve terminals in situ', *J Neurophysiol* **99**(3), 1545–1553.

Friedman, J., Hastie, T., Höfling, H. and Tibshirani, R. (2007), 'Pathwise coordinate optimization', *The Annals of Applied Statistics* **1**(2), 302–332.

Friedman, J., Hastie, T. and Tibshirani, R. (2008), *The Elements of Statistical Learning*, Spinger.

Gelman, A., Carlin, J., Stern, H. and Rubin, D. (2004), *Bayesian data analysis*, CRC press.

Geman, S., Bienenstock, E. and Doursat, R. (1992), 'Neural networks and the bias/variance dilemma', *Neural computation* **4**(1), 1–58.

Gobel, W. and Helmchen, F. (2007), 'New angles on neuronal dendrites in vivo', *J Neurophysiol* **98**(6), 3770–3779.

Hines, M. (1984), 'Efficient computation of branched nerve equations', *International Journal of Bio-Medical Computing* **15**(1), 69 – 76.

Huber, P. (1964), 'Robust estimation of a location parameter', *The Annals of Mathematical Statistics* **35**(1), 73–101.

Huggins, J. and Paninski, L. (2012), 'Optimal experimental design for sampling voltage on dendritic trees', *In Press, Journal of Computational Neuroscience* .

Huys, Q., Ahrens, M. and Paninski, L. (2006), 'Efficient estimation of detailed single-neuron models', *Journal of Neurophysiology* **96**, 872–890.

Huys, Q. and Paninski, L. (2009), 'Model-based smoothing of, and parameter estimation from, noisy biophysical recordings', *PLOS Computational Biology* **5**, e1000379.

Iyer, V., Hoogland, T. M. and Saggau, P. (2006), 'Fast functional imaging of single neurons using random-access multiphoton (RAMP) microscopy', *J Neurophysiol* **95**(1), 535–545.

Knopfel, T., Diez-Garcia, J. and Akemann, W. (2006), 'Optical probing of neuronal circuit dynamics: genetically encoded versus classical fluorescent sensors', *Trends in Neurosciences* **29**, 160–166.

Kralj, J., Douglass, A., Hochbaum, D., Maclaurin, D. and Cohen, A. (2011), 'Optical recording of action potentials in mammalian neurons using a microbial rhodopsin', *Nature Methods* .

Larkum, M. E., Watanabe, S., Lasser-Ross, N., Rhodes, P. and Ross, W. N. (2008), 'Dendritic properties of turtle pyramidal neurons', *J Neurophysiol* **99**(2), 683–694.

Lin, Y. and Zhang, H. (2006), 'Component selection and smoothing in multivariate nonparametric regression', *The Annals of Statistics* **34**(5), 2272–2297.

Mallows, C. (1973), 'Some comments on Cp', *Technometrics* pp. 661–675.

Milojkovic, B. A., Zhou, W.-L. and Antic, S. D. (2007), 'Voltage and calcium transients in basal dendrites of the rat prefrontal cortex', *J Physiol* **585**(2), 447–468.

Mishchenko, Y. and Paninski, L. (2012), 'A Bayesian compressed-sensing approach for reconstructing neural connectivity from subsampled anatomical data', *Under review* .

Mishchenko, Y., Vogelstein, J. and Paninski, L. (2011), 'A Bayesian approach for inferring neuronal connectivity from calcium fluorescent imaging data', *Annals of Applied Statistics* **5**, 1229–1261.

Mitchell, T. J. and Beauchamp, J. J. (1988), 'Bayesian variable selection in linear regression', *Journal of the American Statistical Association* **83**(404), 1023–1032.

Nikolenko, V., Watson, B., Araya, R., Woodruff, A., Peterka, D. and Yuste, R. (2008), 'SLM microscopy: Scanless two-photon imaging and photostimulation using spatial light modulators', *Frontiers in Neural Circuits* **2**, 5.

Nuriya, M., Jiang, J., Nemet, B., Eisenthal, K. and Yuste, R. (2006), 'Imaging membrane potential in dendritic spines', *PNAS* **103**, 786–790.

Packer, A. M., Peterka, D. S., Hirtz, J. J., Prakash, R., Deisseroth, K. and Yuste, R. (2012), 'Two-photon optogenetics of dendritic spines and neural circuits', *Nature methods* **9**(12), 1202–1205.

Pakman, A. and Paninski, L. (2013), 'Exact hamiltonian monte carlo for truncated multivariate gaussians', *Journal of Computational and Graphical Statistics, preprint arXiv:1208.4118* .

Paninski, L. (2010), 'Fast Kalman filtering on quasilinear dendritic trees', *Journal of Computational Neuroscience* **28**, 211–28.

Paninski, L. and Ferreira, D. (2008), 'State-space methods for inferring synaptic inputs and weights', *COSYNE* .

Paninski, L., Vidne, M., DePasquale, B. and Ferreira, D. (2012), 'Inferring synaptic inputs given a noisy voltage trace via sequential monte carlo methods', *In Press, Journal of Computational Neuroscience* .

Peterka, D., Takahashi, H. and Yuste, R. (2011), 'Imaging voltage in neurons', *Neuron* **69**(1), 9–21.

Pnevmatikakis, E. A., Kelleher, K., Chen, R., Saggau, P., Josić, K. and Paninski, L. (2012), Fast spatiotemporal smoothing of calcium measurements in dendritic trees. submitted.

Pnevmatikakis, E. A. and Paninski, L. (2012), 'Fast interior-point inference in high-dimensional sparse, penalized state-space models', *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics (AISTATS) 2012, La Palma, Canary Islands. Volume XX of JMLR: W&CP XX.* .

Pnevmatikakis, E. A., Paninski, L., Rad, K. R. and Huggins, J. (2012), 'Fast Kalman filtering and forward-backward smoothing via a low-rank perturbative approach', *Journal of Computational and Graphical Statistics* . in press.

Press, W., Teukolsky, S., Vetterling, W. and Flannery, B. (1992), *Numerical recipes in C*, Cambridge University Press.

Reddy, G. D. and Saggau, P. (2005), 'Fast three-dimensional laser scanning scheme using acousto-optic deflectors', *J Biomed Opt* **10**(6), 064038.

Sacconi, L., Dombeck, D. A. and Webb, W. W. (2006), 'Overcoming photodamage in second-harmonic generation microscopy: Real-time optical recording of neuronal action potentials', *Proceedings of the National Academy of Sciences* **103**(9), 3124–3129.

Sjostrom, P. J., Rancz, E. A., Roth, A. and Hausser, M. (2008), 'Dendritic Excitability and Synaptic Plasticity', *Physiol. Rev.* **88**(2), 769–840.

Smith, C. (2013), 'Low-rank graphical models and Bayesian analysis of neural data', *PhD Thesis, Columbia University* .

Song, S., Sjöström, P. J., Reigl, M., Nelson, S. and Chklovskii, D. B. (2005), 'Highly nonrandom features of synaptic connectivity in local cortical circuits', *PLoS biology* **3**(3), e68.

Studer, V., Bobin, J., Chahid, M., Mousavi, H., Candes, E. and Dahan, M. (2012), 'Compressive fluorescence microscopy for biological and hyperspectral imaging', *Proceedings of the National Academy of Sciences* **109**(26), E1679–E1687.

Takahashi, N., Kitamura, K., Matsuo, N., Mayford, M., Kano, M., Matsuki, N. and Ikegaya, Y. (2012), 'Locally synchronized synaptic inputs', *Science* **335**(6066), 353–356.

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

Vucinic, D. and Sejnowski, T. J. (2007), 'A compact multiphoton 3d imaging system for recording fast neuronal activity', *PLoS ONE* **2**(8), e699.

Yuan, M. and Lin, Y. (2006), 'Model selection and estimation in regression with grouped variables', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68**(1), 49–67.

Zou, H. and Hastie, T. (2005), 'Regularization and variable selection via the elastic net', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**(2), 301–320.

Zou, H., Hastie, T. and Tibshirani, R. (2007), 'On the "degrees of freedom" of the lasso', *The Annals of Statistics* **35**(5), 2173–2192.