SIAM J. APPL. MATH.
Vol. 68, No. 2, pp. 354–391

EXPLOITING HISTORY-DEPENDENT EFFECTS TO INFER NETWORK CONNECTIVITY[*]

DUANE Q. NYKAMP[†]

Abstract. We present an approach to distinguish between causal connections and common input connections among nodes in a network. By modeling how the activity of a node depends on its own recent history, we demonstrate how this history dependence predicts different patterns of activity depending on the nature of the network connectivity. In particular, a causal connection between a pair of observed nodes can be distinguished from common input connections that originate from nodes whose activity remains unobserved. This work builds on previous results where this same distinction was made based on modeling how the activity of a node depends on measured external variables such as stimuli. The results have a potentially broad range of application as the analysis can be based on a fairly generic class of models.

Key words. neural networks, correlations, causality, maximum likelihood, point process, autocorrelation

AMS subject classification. 92C20

DOI. 10.1137/070683350

1. Introduction.

The determination of causal connections among nodes within a network is a difficult challenge. This challenge is magnified in the presence of hidden nodes, the effects of which can mimic the presence of causal connections among the set of measured nodes. For example, the connection from a hidden node onto two measured nodes could introduce correlations in the activity of the measured nodes that resemble the effect of a causal connection between the measured nodes (see Figure 1).

We have recently introduced [14, 13, 12] an approach for identifying causal connections in the presence of hidden nodes that is based on modeling the relationship between the activity of nodes and measurable external variables, such as those representing a stimulus. In the original formulation of this approach, the activity of any node could be only weakly dependent on the history of its activity. However, in general, the activity of a node could depend strongly on the recent activity of that node. For example, this approach was originally designed for neuronal networks, and the spiking activity of a neuron is strongly modulated by that neuron's spike history. After firing a spike, a neuron cannot immediately fire a second spike due to its refractory period. Some neurons tend to fire spikes in bursts so that, once the refractory period is over, the probability of firing a spike is transiently much higher than baseline. These history-dependent effects were neglected in our original formulation.

We have now discovered that, if one models how the activity of a node depends on its recent history, one's ability to distinguish causal connections within a network is enhanced. The reason that modeling history dependence can help determine causal connections is caricatured in Figure 2. For the purpose of illustration, imagine that the nodes are neurons and that the measured activity is the times of the neurons' output spikes. Moreover, imagine that neuron 1 tends to fire spikes in pairs (note the

[*] Received by the editors February 22, 2007; accepted for publication (in revised form) July 25, 2007; published electronically November 9, 2007. This research was supported by National Science Foundation grant DMS-0415409.

http://www.siam.org/journals/siap/68-2/68335.html

[†] School of Mathematics, University of Minnesota, Minneapolis, MN 55455 (nykamp@math.umn.edu).

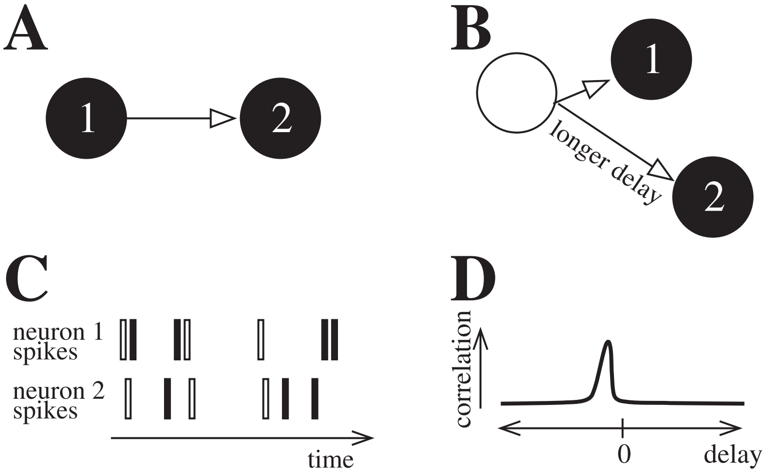

Fig. 1. *The effect of hidden nodes (unfilled circle) can be to mimic causal connections among measured nodes (filled circles).* (A) *A causal connection from measured node* 1 *onto measured node* 2*.* (B) *A common connection from a hidden node onto two measured nodes, where the connection onto measured node* 2 *has a longer delay.* (C) *Both the common input configuration (A) and the causal connection configuration (B) produce similar correlations in the activity of the measured nodes. For concreteness, let the nodes be neurons whose activity is a sequence of spike times illustrated by the temporal sequence of rectangles. Both the networks (A and B) will increase the probability that neuron* 2 *will fire a spike immediately after neuron* 1*. (These spike combinations are highlighted by the unfilled rectangles.)* (D) *Schematic of the correlation induced by either network (A or B). Neuron* 2 *is highly likely to fire a spike a certain delay after neuron* 1 *fires. There is a peak in the correlation measured at that delay. (We arbitrarily use a negative delay when neuron* 2 *follows neuron* 1*.) Since both the common input (A) and the causal connection (B) configurations induce similar correlations in the activity of the measured nodes, our goal is to distinguish which configuration underlies the measured activity of the two nodes.*

pairs of closely spaced spikes in the output of neuron 1 in the right panels of Figure 2). We argue that neuron 2 should respond differently to the spike pairs depending on whether the network contains a causal connection (Figure 2(A)) or common input connections (Figure 2(B)).

To further simplify the situation, imagine that the spike trains of neither neuron 2 nor the hidden neuron have a significant dependence on their history. Then, as portrayed in Figure 2(A), if neuron 1 has a causal connection onto neuron 2, neuron 2 will respond equally well to both spikes in the spike pairs emitted by neuron 1. Neuron 2 will receive both spikes in the pair as inputs, so neuron 2 will be likely to fire a spike immediately after both of these inputs. On the other hand, in the common input configuration of Figure 2(B), neuron 2 does not receive neuron 1's spike pairs as inputs. When the hidden neuron fires a single spike, it may elicit a spike pair from neuron 1. However, neuron 2 just receives the single input from the hidden neuron, so neuron 2 will not be driven to fire twice. When looking at just the spike trains of neuron 1 and 2, it may appear, for example, that neuron 2 is responding to just the first spike of each pair from neuron 1 and ignoring the second spike. The key intuition to gain from this example is that, for the common input configuration, it looks as though neuron 2 does not respond to spikes that can be predicted by the history dependence of neuron 1.

Of course, any real situation will be far more complicated than this exaggerated example. For instance, all of the nodes could have a strong history dependence to their activity, which will confound the simple reasoning given above. Moreover,

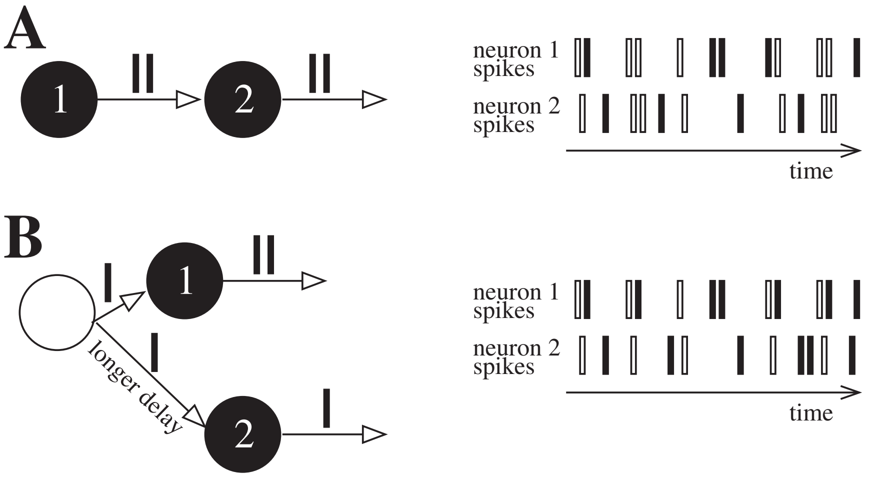

Fig. 2. *Illustration of the different effects of history dependence based on the underlying circuitry. Nodes are spiking neurons as in Figure* 1*. For this illustration, we assume neuron* 1*'s activity is strongly dependent on its spiking history; it is highly likely to fire spikes in pairs. We also assume that the spikes of neuron* 2 *and the unmeasured neuron are largely independent of their respective history.* (A) *In a causal connection configuration, neuron* 2 *may respond to all of neuron* 1*'s spikes. As schematized on the left, when neuron* 1 *fires a pair of spikes (black rectangles), neuron* 2 *is likely to spike after each one and so may spike twice. The right panel illustrates a possible temporal sequence of spikes from both neurons. Spike combinations where neuron* 2 *fires immediately after neuron* 1 *are highlighted by unfilled rectangles. Neuron* 2 *is highly likely to spike both after the first spike and after the second spike in each spike pair from neuron* 1*.* (B)*. In a common input configuration, neuron* 2 *does not receive the spike pairs from neuron* 1*. Since a single spike from the hidden neuron can evoke the spike pair from neuron* 1*, neuron* 2 *receives only one input that is correlated with the spike pair from neuron* 1*. If the connection from the hidden neuron onto neuron* 2 *has a slightly longer delay than the connection onto neuron* 1*, neuron* 2 *will be likely to fire immediately after the first spike in each pair from neuron* 1*. It will not be likely to spike after the second spike of the pair, as illustrated in the right panel.*

the influence of the connections between a pair of nodes will typically be weaker than illustrated here, as input received via any one connection will be just one small influence on a node bathed with inputs from other nodes in the network. Hence, exploiting such history-dependent effects to infer connectivity requires some form of analysis that can synthesize the various ways in which internode connectivity and intranode history-dependent effects interact to influence nodes' activities. Nonetheless, the mathematical analysis we present will confirm that intuition gleaned from this exaggerated example does apply to the more complex situation (see section 3.4.1).

This paper presents a mathematical analysis through which one can employ a model of history-dependent effects to develop estimates of the network connectivity among measured nodes. In section 2, we describe the class of models that we consider. In section 3, we present the analysis to determine the connectivity. We demonstrate the results applied to simulated networks in section 4 and discuss the results in section 5.

2. The history-dependent model.

We present our model and analysis in fairly abstract terms. As detailed in [14], we employ a modular approach where the details of the single-node model are ignored in the network analysis. To employ the results to analyze a particular dataset, one must select an appropriate model, the form of which can be "plugged into" the network analysis.

2.1. The general model formulation. The model is formulated in discrete time. Let R_s^i be a random variable that represents the activity of node s at time point i. Since our examples will involve models of neurons, we will assume that R_s^i is a discrete random variable. However, the analysis proceeds analogously for a continuous random variable. Ignoring the activity of other nodes for a moment, the probability distribution of R_s^i will depend both on the history of node s and on some measurable external variables. Let $\mathbf{R}_s^{<i}$ be the vector of the history of the activity of node s (i.e., the vector with values of R_s^k for $k < i$). Denote the external variable vector by $\mathbf{X}$. The vector $\mathbf{X}$ could represent any quantity or set of quantities whose values are known and that modulate the activity of the nodes. For example, in neuroscience applications, $\mathbf{X}$ could correspond to a sequence of stimuli or a sequence of animal positions. (See [14] for a discussion on external variables. Note that $\mathbf{X}$ could depend on time, although the notation does not make that explicit.)

The activity of a given node on the network also depends on activity of other nodes. We denote the network connectivity by $\bar{W}_{\tilde{s},s}^{\tilde{\imath},i}$, which indicates the magnitude of the effect of the activity of node $\tilde{s}$ at time $\tilde{\imath}$ (i.e., $R_{\tilde{s}}^{\tilde{\imath}}$) on the activity of node s at time i (i.e., R_s^i). We assume that all connections are causal so that $\bar{W}_{\tilde{s},s}^{\tilde{\imath},i} = 0$ for $\tilde{\imath} \geq i$. We model the effect of $R_{\tilde{s}}^{\tilde{\imath}}$ on the probability distribution of R_s^i as a function of the product $\bar{W}_{\tilde{s},s}^{\tilde{\imath},i} R_{\tilde{s}}^{\tilde{\imath}}$. Moreover, we simply linearly sum the coupling effects from all nodes and previous time steps, modeling the total coupling effect of all nodes on the probability distribution of R_s^i as a function of the sum

$$\sum_{\tilde{s} \neq s} \sum_{\tilde{\imath} < i} \bar{W}_{\tilde{s},s}^{\tilde{\imath},i} R_{\tilde{s}}^{\tilde{\imath}}.$$

To summarize, we model the probability distribution of R_s^i as a parametric function of the history $\mathbf{R}_s^{<i}$ of node s, the external variables $\mathbf{X}$, and the past activity of all nodes as

$$\Pr(R_s^i = r_s^i \,|\, \mathbf{R}^{<i} = \mathbf{r}^{<i}, \mathbf{X} = \mathbf{x}) = P_s\left(r_s^i, \mathbf{r}_s^{<i}, \mathbf{x}, \sum_{\tilde{s} \neq s} \sum_{\tilde{\imath} < i} \bar{W}_{\tilde{s},s}^{\tilde{\imath},i} r_{\tilde{s}}^{\tilde{\imath}}; \bar{\theta}_s^i \right), \tag{2.1}$$

where P_s is some discrete probability distribution in its first argument and $\bar{\theta}_s^i$ is a vector of parameters. The quantity $\mathbf{R}^{<i}$ (without a subscript) is the history of all nodes, i.e., has components $R_{\tilde{s}}^k$ for all $\tilde{s}$ and all $k < i$. If $\mathbf{R}$ represents all of the activity of all nodes (i.e., has components $R_{\tilde{s}}^k$ for all $\tilde{s}$ and k), then, by Bayes' law, the probability distribution of $\mathbf{R}$, given the value of the external variable vector $\mathbf{X}$, is

$$\begin{aligned} \Pr(\mathbf{R} = \mathbf{r} \,|\, \mathbf{X} = \mathbf{x}) &= \prod_s \prod_i \Pr(R_s^i = r_s^i \,|\, \mathbf{R}^{<i} = \mathbf{r}^{<i}, \mathbf{X} = \mathbf{x}) \\ &= \prod_s \prod_i P_s\left(r_s^i, \mathbf{r}_s^{<i}, \mathbf{x}, \sum_{\tilde{s} \neq s} \sum_{\tilde{\imath} < i} \bar{W}_{\tilde{s},s}^{\tilde{\imath},i} r_{\tilde{s}}^{\tilde{\imath}}; \bar{\theta}_s^i \right). \end{aligned} \tag{2.2}$$

To obtain (2.2), we exploited the fact that nodes influence each other only through causal connections. Hence, conditioned on the history $\mathbf{R}^{<i}$ of the network and the external variables, the activities R_s^i of nodes in a single time step i are independent. (In other words, we assume the time bins are small enough so that interactions involve a delay of at least one time bin.)

2.2. Assumptions required for analysis. With the exception of the linear coupling among nodes, (2.2) is a fairly generic description of a network in discrete time. (Recall that the equations could be trivially modified to allow the activity R_s^i to be a continuous random variable.) However, to proceed with our analysis we make a few strong assumptions about the network. These assumptions are similar to those detailed in [14]. (The biggest difference is that here we make no assumptions about the dependence of a node on its own history.) For this reason, we present only a brief discussion of these assumptions here and refer the reader to the more detailed discussion in the former article.

First, we assume that an algorithm exists to fit the activity of a single node to the same parametric model with the coupling factors $\bar{W}_{\tilde{s},s}^{\tilde{\imath},i}$ set to zero. Note that this particular assumption is only about choice of models; it is not an assumption about the network activity. We assume that, from measurements of the activity of just a single node s (i.e., of the vector $\mathbf{R}_s$ composed of R_s^i for all i), one has an algorithm to determine effective parameters θ_s^i by fitting the averaged model[1]

$$\Pr(\mathbf{R}_s = \mathbf{r}_s \,|\, \mathbf{X} = \mathbf{x}) = \prod_i P_s\bigl(r_s^i, \mathbf{r}_s^{<i}, \mathbf{x}, 0; \theta_s^i\bigr). \tag{2.3}$$

Although the activity of all other nodes $R_{\tilde{s}}^{\tilde{\imath}}$ is ignored in this fitting procedure, we view the R_s^i as really generated from the full network via model (2.2). Therefore, the effective parameters θ_s^i do include the averaged effects of the coupling from other nodes. Our analysis will rely heavily on these effective parameters; hence, the results depend on having chosen a good model P_s and fitting algorithm so that the averaged model (2.3) captures key elements of the activity of each node. This assumption puts stringent limits on the model P_s. For example, one cannot use detailed biophysical models, as all of the parameters of such models cannot be determined by $\mathbf{R}_s$ and $\mathbf{X}$ alone. Neither could one allow the θ_s^i to be independent for each time i. (See [14] for more details.)

Second, we assume that the coupling $\bar{W}_{\tilde{s},s}^{\tilde{\imath},i}$ is weak so that we can expand the full model (2.2) in a Taylor series in $\bar{W}_{\tilde{s},s}^{\tilde{\imath},i}$ and retain terms only through second order. Since we assume that P_s is C^2 in its fourth argument, our analysis will have an error that is $O(\bar{W}^3)$. As detailed in [14], the assumption has the following important consequences: The average coupling strength must scale like $1/N$, where N is the network size; the identity of the nodes that appear in (2.2) must be regarded as "lumped" models that already incorporate effects of nodes projecting to them; and the network topology is highly simplified, as the second order Taylor series will represent combinations of at most two edges of the network graph. If the actual connectivity is too strong to strictly justify this assumption, the resulting connectivity estimates may need to be reinterpreted as an effective connectivity (see the discussion in section 5).

Third, once we have calculated θ_s^i by fitting the averaged model (2.3), we assume that the model is constructed so that we can calculate $P_s(r_s^i, \mathbf{r}_s^{<i}, \mathbf{x}, w; \theta_s^i)$ for any value of w. This is a strong assumption on the allowed form of the model function P_s, as the averaged model (2.3) is not based on $P_s(r_s^i, \mathbf{r}_s^{<i}, \mathbf{x}, w; \theta_s^i)$ for any nonzero w. This assumption also implies that we can calculate $\frac{\partial}{\partial w} P_s(r_s^i, \mathbf{r}_s^{<i}, \mathbf{x}, w; \theta_s^i)$ and

[1] The probability of the left-hand side of model (2.3) is the marginal distribution of the probability of the left-hand side of model (2.2), averaged over the activity of all other nodes.

$\frac{\partial^2}{\partial w^2} P_s(r_s^i, \mathbf{r}_s^{<i}, \mathbf{x}, w; \theta_s^i)$. In fact, this assumption implies that these derivatives must be equivalent to derivatives with respect to some function of $\mathbf{r}_s^{<i}$, $\mathbf{x}$, and θ_s^i.

Last, unless one could repeatedly sample the activity of the nodes from the same time points, one couldn't hope to be able to determine arbitrary connectivity $\bar{W}_{\tilde{s},s}^{\tilde{\imath},i}$ that varies freely with the time point. (This is the same reason θ_s^i cannot be allowed to vary freely with the time point, as mentioned above.) When we actually implement the approach, we will eventually (see section 3.3.3) allow $\bar{W}_{\tilde{s},s}^{\tilde{\imath},i}$ to depend on $\tilde{\imath}$ and i only through the delay $i - \tilde{\imath}$. (One could also allow the coupling to adapt slowly with time.) During most of our analysis, we will keep the notation where $\bar{W}_{\tilde{s},s}^{\tilde{\imath},i}$ varies freely with the time point, as it adds no complexity to the equations.

3. The analysis.

We begin by giving a short overview of the analysis. We operate under the assumption that model (2.2) gives the true probability distribution of the activity $\mathbf{R}$ of the entire network. However, we assume that one can observe just a small number of nodes with indices q in some subset $\mathcal{Q}$. We denote by $\mathbf{R}_{\mathcal{Q}}$ the activity of all of these measured nodes. (The components of $\mathbf{R}_{\mathcal{Q}}$ are a subset of those of $\mathbf{R}$.)

The first step of the analysis will be to derive an expression for the probability distribution of the activity of just the measured nodes. We will derive an expression for this probability, which we denote by $\Pr(\mathbf{R}_{\mathcal{Q}}|\mathbf{X})$, by taking the expression for $\Pr(\mathbf{R}|\mathbf{X})$ given in (2.2) and averaging it over the activity of all hidden nodes. This step will rely heavily on the weak coupling assumption described above.

The resulting expression for $\Pr(\mathbf{R}_{\mathcal{Q}}|\mathbf{X})$ will depend on all of the unknown parameters $\bar{\theta}_s^i$ and $\bar{W}_{\tilde{s},s}^{\tilde{\imath},i}$. Given that many nodes remain hidden, we don't have any hope of obtaining estimates of the original parameters $\bar{\theta}_s^i$. However, we do, by assumption, have an algorithm for determining the effective parameters θ_q^i of any measured node q by fitting the averaged model (2.3) to the activity of that measured node. To take advantage of this information, our second step will be to derive an expression for the original parameters $\bar{\theta}$ in terms of the effective parameters θ.

Our third step is to combine the results of steps one and two to arrive at an expression for the probability distribution $\Pr(\mathbf{R}_{\mathcal{Q}}|\mathbf{X})$ of the measured node activity in terms of the effective parameters. With one further approximation, we can group all of the effects of the hidden nodes into a small number of parameters. In the end, our expression for $\Pr(\mathbf{R}_{\mathcal{Q}}|\mathbf{X})$ will contain just two sets of unknown parameters: the effective causal connection parameters (which we'll denote by an unbarred W) and the effective common input parameters (which we'll denote by U). Given a measurement of the activity of the measured nodes, we use our expression for $\Pr(\mathbf{R}_{\mathcal{Q}}|\mathbf{X})$ to compute maximum likelihood estimates of the W and U. The effective causal connection W will be our estimate of the connectivity among the measured nodes.

3.1. Step one: Average for measured node probability distribution.

Our first step is to average the full model (2.2) over all possible values of the activity of hidden nodes to obtain an expression for the probability distribution of measured node activity. Before we compute the average, we use the weak coupling assumption described in section 2.2 to simplify (2.2).

We invoke the weak coupling assumption to expand the full model (2.2) as a Taylor series in $\bar{W}$. To simplify the presentation, we define the following shorthand notation for the probability distribution of the activity of node s (over all time points

i) that would result if all coupling was set to zero:

$$\bar{P}_s = \prod_i P_s\big(r_s^i, \mathbf{r}_s^{<i}, \mathbf{x}, 0; \bar{\theta}_s^i, \big). \tag{3.1}$$

Similarly, we define shorthand notation for the derivatives of $\bar{P}_s$ with respect to the $r_{\tilde{s}}^{\tilde{\imath}}$ for $\tilde{s} \neq s$:

$$\begin{aligned} \frac{\partial \bar{P}_s}{\partial r_{\tilde{s}}^{\tilde{\imath}}} &= \frac{\partial}{\partial r_{\tilde{s}}^{\tilde{\imath}}} \left(\prod_i P_s \left(r_s^i, \mathbf{r}_s^{<i}, \mathbf{x}, \sum_{\acute{s} \neq s} \sum_{\acute{\imath} < i} \bar{W}_{\acute{s},s}^{\acute{\imath},i} r_{\acute{s}}^{\acute{\imath}}; \bar{\theta}_s^i \right) \right) \Bigg|_{\{r_{\acute{s}}^{\acute{\imath}} = 0 | \acute{s} \neq s\}}, \\ \frac{\partial^2 \bar{P}_s}{\partial r_{\tilde{s}_1}^{\tilde{\imath}_1} \partial r_{\tilde{s}_2}^{\tilde{\imath}_2}} &= \frac{\partial^2}{\partial r_{\tilde{s}_1}^{\tilde{\imath}_1} \partial r_{\tilde{s}_2}^{\tilde{\imath}_2}} \left(\prod_i P_s \left(r_s^i, \mathbf{r}_s^{<i}, \mathbf{x}, \sum_{\acute{s} \neq s} \sum_{\acute{\imath} < i} \bar{W}_{\acute{s},s}^{\acute{\imath},i} r_{\acute{s}}^{\acute{\imath}}; \bar{\theta}_s^i \right) \right) \Bigg|_{\{r_{\acute{s}}^{\acute{\imath}} = 0 | \acute{s} \neq s\}}. \end{aligned} \tag{3.2}$$

Since the $\bar{W}$ appear only in the combination $\bar{W}_{\tilde{s},s}^{\tilde{\imath},i} r_{\tilde{s}}^{\tilde{\imath}}$, we can write our Taylor series in $\bar{W}$ as though it were a Taylor series in the $r_{\tilde{s}}^{\tilde{\imath}}$. This notation will turn out to be more convenient for the analysis because the key factors of $r_{\tilde{s}}^{\tilde{\imath}}$ will be written out explicitly. Note that, for each node s, we make no assumptions about the effect of its own history $\mathbf{r}_s^{<i}$ and do *not* expand out this history dependence in a Taylor series.

Using the above shorthand notation, the Taylor series of (2.2) is

$$\begin{aligned} \Pr(\mathbf{R} = \mathbf{r} | \mathbf{X} = \mathbf{x}) = \prod_s \bar{P}_s &+ \sum_{\substack{s_1, \tilde{s}_1 \\ \tilde{s}_1 \neq s_1}} \sum_{\tilde{\imath}_1} \frac{\partial \bar{P}_{s_1}}{\partial r_{\tilde{s}_1}^{\tilde{\imath}_1}} r_{\tilde{s}_1}^{\tilde{\imath}_1} \prod_{\substack{s_2 \\ s_2 \neq s_1}} \bar{P}_{s_2} \\ &+ \frac{1}{2} \sum_{\substack{s_1, \tilde{s}_1, \tilde{s}_2 \\ \tilde{s}_1 \neq s_1, \tilde{s}_2 \neq s_1}} \sum_{\tilde{\imath}_1, \tilde{\imath}_2} \frac{\partial^2 \bar{P}_{s_1}}{\partial r_{\tilde{s}_1}^{\tilde{\imath}_1} \partial r_{\tilde{s}_2}^{\tilde{\imath}_2}} r_{\tilde{s}_1}^{\tilde{\imath}_1} r_{\tilde{s}_2}^{\tilde{\imath}_2} \prod_{\substack{s_2 \\ s_2 \neq s_1}} \bar{P}_{s_2} \\ &+ \frac{1}{2} \sum_{\substack{s_1, s_2, \tilde{s}_1, \tilde{s}_2 \\ s_2 \neq s_1, \tilde{s}_1 \neq s_1 \\ \tilde{s}_2 \neq s_2}} \sum_{\tilde{\imath}_1, \tilde{\imath}_2} \frac{\partial \bar{P}_{s_1}}{\partial r_{\tilde{s}_1}^{\tilde{\imath}_1}} \frac{\partial \bar{P}_{s_2}}{\partial r_{\tilde{s}_2}^{\tilde{\imath}_2}} r_{\tilde{s}_1}^{\tilde{\imath}_1} r_{\tilde{s}_2}^{\tilde{\imath}_2} \prod_{\substack{s_3 \\ s_3 \neq s_1 \\ s_3 \neq s_2}} \bar{P}_{s_3} + O(\bar{W}^3). \end{aligned} \tag{3.3}$$

Note that the derivative $\partial \bar{P}_s / \partial r_{\tilde{s}}^{\tilde{\imath}}$ corresponds to the effect of a connection from node $\tilde{s}$ onto node s. If we wrote out the derivative explicitly, it would contain a sum of terms involving $\bar{W}_{\tilde{s},s}^{\tilde{\imath},i}$ for all $i > \tilde{\imath}$. It represents the change in the distribution of all R_s^i for $i > \tilde{\imath}$ given a change in $R_{\tilde{s}}^{\tilde{\imath}}$ (calculated at $R_{\tilde{s}}^{\tilde{\imath}} = 0$).

We can now write down an expression for the activity of all measured nodes by averaging over all possible values of the activity of the hidden nodes. As mentioned above, let $\mathcal{Q}$ denote the set of node indices corresponding to all measured nodes. Similarly, let $\mathcal{P}$ denote the set of node indices corresponding to all hidden nodes. Then $\mathcal{Q} \cup \mathcal{P}$ corresponds to the entire network. To simplify the notation, we will make the following notational conventions. We will use the index s and its variants to index all nodes in the network; i.e., we implicitly assume that $s \in \mathcal{Q} \cup \mathcal{P}$. We will use the indices p and q (and their variants) to index hidden and measured nodes, respectively; i.e., we implicitly assume that $p \in \mathcal{P}$ and $q \in \mathcal{Q}$. Last, we let $\mathbf{R}_{\mathcal{Q}}$ and $\mathbf{R}_{\mathcal{P}}$ represent all measured node activity R_q^i and all hidden node activity R_p^i, respectively.

To derive an expression for the probability distribution of all measured activity, we average (3.3) over all possible values of $\mathbf{R}_\mathcal{P}$. The probability distribution of $\mathbf{R}_\mathcal{Q}$ is therefore

$$\begin{aligned}\Pr(\mathbf{R}_\mathcal{Q} = \mathbf{r}_\mathcal{Q} | \mathbf{X} = \mathbf{x}) &= \sum_{\mathbf{r}_\mathcal{P}} \Pr(\mathbf{R} = \mathbf{r} | \mathbf{X} = \mathbf{x}) \\ &= \sum_{\mathbf{r}_\mathcal{P}} \prod_s \bar{P}_s + \sum_{\mathbf{r}_\mathcal{P}} \sum_{\substack{s_1, \tilde{s}_1 \\ \tilde{s}_1 \neq s_1}} \sum_{\tilde{\imath}_1} \frac{\partial \bar{P}_{s_1}}{\partial r^{\tilde{\imath}_1}_{\tilde{s}_1}} r^{\tilde{\imath}_1}_{\tilde{s}_1} \prod_{\substack{s_2 \\ s_2 \neq s_1}} \bar{P}_{s_2} \\ &\quad + \frac{1}{2} \sum_{\mathbf{r}_\mathcal{P}} \sum_{\substack{s_1, \tilde{s}_1, \tilde{s}_2 \\ \tilde{s}_1 \neq s_1, \tilde{s}_2 \neq s_1}} \sum_{\tilde{\imath}_1, \tilde{\imath}_2} \frac{\partial^2 \bar{P}_{s_1}}{\partial r^{\tilde{\imath}_1}_{\tilde{s}_1} \partial r^{\tilde{\imath}_2}_{\tilde{s}_2}} r^{\tilde{\imath}_1}_{\tilde{s}_1} r^{\tilde{\imath}_2}_{\tilde{s}_2} \prod_{\substack{s_2 \\ s_2 \neq s_1}} \bar{P}_{s_2} \\ &\quad + \frac{1}{2} \sum_{\mathbf{r}_\mathcal{P}} \sum_{\substack{s_1, s_2, \tilde{s}_1, \tilde{s}_2 \\ s_2 \neq s_1, \tilde{s}_1 \neq s_1 \\ \tilde{s}_2 \neq s_2}} \sum_{\tilde{\imath}_1, \tilde{\imath}_2} \frac{\partial \bar{P}_{s_1}}{\partial r^{\tilde{\imath}_1}_{\tilde{s}_1}} \frac{\partial \bar{P}_{s_2}}{\partial r^{\tilde{\imath}_2}_{\tilde{s}_2}} r^{\tilde{\imath}_1}_{\tilde{s}_1} r^{\tilde{\imath}_2}_{\tilde{s}_2} \prod_{\substack{s_3 \\ s_3 \neq s_1 \\ s_3 \neq s_2}} \bar{P}_{s_3} + O(\bar{W}^3), \end{aligned} \tag{3.4}$$

where the sum over $\mathbf{r}_\mathcal{P}$ indicates a sum over all possible values of the hidden node activity.

It turns out that we can explicitly compute the sum over $\mathbf{r}_\mathcal{P}$. Note that the value r^i_s of a given random variable can appear in (3.4) either explicitly or in the probability distribution $\bar{P}_s$ (or its derivatives). It is not hidden in any other factors. Therefore, to compute a sum over all possible values of the activity of a node indexed by some s, we can factor out everything except one factor of $\bar{P}_s$ (or a derivative of $\bar{P}_s$) and a polynomial in the r^i_s. Hence, we need to derive expressions for the average of such quantities.

The average of a polynomial in the r^i_s multiplied by the undifferentiated $\bar{P}_s$ will simply be the expected value of that polynomial, under the probability distribution $\bar{P}_s$ with the W argument set to zero. Taking the average of expressions involving the derivatives of $\bar{P}_s$ is more complicated. In Appendix A.1, we outline how to compute such averages. The important point is that one can compute these averages explicitly in terms of the model parameters and the probability distributions $P_s(\cdot)$. We end up with the lengthy expression for $\Pr(\mathbf{R}_\mathcal{Q}|\mathbf{X})$ given by (A.5).

3.2. Step two: Original parameters in terms of effective parameters. One of the assumptions given in section 2.2 is the existence of an algorithm to calculate the effective parameters θ^i_s by fitting the averaged model (2.3) to the activity of node s (while ignoring the activity of all other nodes). Hence, we can regard the effective parameters θ^i_q as known for all measured nodes $q \in \mathcal{Q}$. In the previous step, we obtained an expression for the probability distribution of the measured node activity $\Pr(\mathbf{R}_\mathcal{Q}|\mathbf{X})$ in terms of the unknown original model parameters $\bar{\theta}^i_s$. In this second step of the analysis, we will derive a relationship between the effective parameters θ^i_s and the original paramters $\bar{\theta}^i_s$. This relationship will allow us to rewrite our equation for $\Pr(\mathbf{R}_\mathcal{Q}|\mathbf{X})$ in terms of the effective parameters.

We define equivalent shorthand notation for expressions involving the effective parameters as we did for expressions involving the original model parameters. We define P_s to be the probability distribution that we fit from the averaged model (2.3):

$$P_s = \prod_i P_s\big(r^i_s, \mathbf{r}^{<i}_s, \mathbf{x}, 0; \theta^i_s,\big) = \Pr(\mathbf{R}_s = \mathbf{r}_s \,|\, \mathbf{X} = \mathbf{x}). \tag{3.5}$$

We then define the derivatives P_s just as we did for $\bar{P}_s$:

$$\frac{\partial P_s}{\partial r_{\tilde{s}}^{\tilde{\imath}}} = \frac{\partial}{\partial r_{\tilde{s}}^{\tilde{\imath}}} \left(\prod_i P_s \left(r_s^i, \mathbf{r}_s^{<i}, \mathbf{x}, \sum_{\acute{s}\neq s} \sum_{\acute{\imath}<i} \bar{W}_{\acute{s},s}^{\acute{\imath},i} r_{\acute{s}}^{\acute{\imath}}; \theta_s^i \right) \right) \Bigg|_{\{r_{\acute{s}}^{\acute{\imath}}=0|\acute{s}\neq s\}} \tag{3.6}$$

and analogously for the second derivatives. We also define the expected values:

$$E_0(g(\mathbf{R})) = \sum_{\mathbf{r}} g(\mathbf{r}) \prod_s P_s, \tag{3.7a}$$

$$E_0\left(\frac{\partial R_s^i}{\partial R_{\tilde{s}}^{\tilde{\imath}}}\right) = \frac{\partial}{\partial r_{\tilde{s}}^{\tilde{\imath}}} E(R_s^i | \mathbf{R}^{<i} = \mathbf{r}^{<i}) \Bigg|_{\{\mathbf{r}_{\acute{s}}=\mathbf{0}|\acute{s}\neq s\}} = \sum_{\mathbf{r}_s} r_s^i \frac{\partial P_s}{\partial r_{\tilde{s}}^{\tilde{\imath}}}. \tag{3.7b}$$

These expected values are analogous to the barred versions given in Appendix A.1 ((A.2) and (A.4)) except that they are based on the averaged model (2.3). We assume that the chosen model and fitting algorithm for θ_s^i results in the averaged model (2.3) being a good approximation. Then (3.7a) does indeed represent the expected value of any function for the activity. For example, $E_0(R_s^i)$ is the expected value of the activity of node s at time i. We will also use the statistic $E_0(R_s^{i_1} R_s^{i_2}) - E_0(R_s^{i_1})E_0(R_s^{i_2})$, which represents the covariance of the activity of node s at the times i_1 and i_2.

The derivative of (3.7b) represents how the average activity of node s at time i changes with the activity of node $\tilde{s}$ at time $\tilde{\imath}$. (Since we assume causal connections, this is nonzero only if $\tilde{\imath} < i$.) See Appendix A.1 for further discussion on the properties of such derivatives.

Although we know the effective parameters only for measured nodes, we can still define the (unknown) effective parameters for hidden nodes using the averaged model (2.3). Using effective parameters for all nodes will simplify the form of our equation for $\Pr(\mathbf{R}_{\mathcal{Q}}|\mathbf{X})$.

It turns out that we have already done much of the work toward deriving an equation for effective parameters in step one, above. In that first step, we derived an expression for $\Pr(\mathbf{R}_{\mathcal{Q}}|\mathbf{X})$, which is the marginal distribution (of the full distribution $\Pr(\mathbf{R}|\mathbf{X})$ given by model (2.2)) for the activity of a set of measured nodes. The averaged model (2.3) is based on $\Pr(\mathbf{R}_s|\mathbf{X})$, which we can regard as the marginal distribution for the activity of a single node. If we replace the set $\mathcal{Q}$ of measured nodes in (A.5) with just the single node s, then (A.5) becomes the marginal distribution for the activity of a single node. In this way, we obtain an expression for $\Pr(\mathbf{R}_s|\mathbf{X})$ in terms of the original model parameters. Given the definition (2.3) of the effective parameters, we have obtained an expression for the effective parameters θ in terms of the original model parameters $\bar{\theta}$.

However, we need to go the other direction: to transform expressions involving the original model parameters $\bar{\theta}$ in terms of the effective parameters θ. Using the procedure outlined in Appendix A.2, we can solve for the original uncoupled probability $\bar{P}_s$ (which is a function of the $\bar{\theta}_s^i$) in terms of the effective probability P_s (which is a function of the θ_s^i). We obtain the following relationship:

$$\begin{aligned}\bar{P}_s = P_s &- \sum_{\substack{\tilde{s}_1 \\ \tilde{s}_1\neq s}} \sum_{\tilde{\imath}_1} \frac{\partial P_s}{\partial r_{\tilde{s}_1}^{\tilde{\imath}_1}} E_0(R_{\tilde{s}_1}^{\tilde{\imath}_1}) \\ &- \frac{1}{2} \sum_{\substack{\tilde{s}_1 \\ \tilde{s}_1\neq s}} \sum_{\tilde{\imath}_1,\tilde{\imath}_2} \frac{\partial^2 P_s}{\partial r_{\tilde{s}_1}^{\tilde{\imath}_1} \partial r_{\tilde{s}_1}^{\tilde{\imath}_2}} [E_0(R_{\tilde{s}_1}^{\tilde{\imath}_1} R_{\tilde{s}_1}^{\tilde{\imath}_2}) - E_0(R_{\tilde{s}_1}^{\tilde{\imath}_1}) E_0(R_{\tilde{s}_1}^{\tilde{\imath}_2})]\end{aligned}$$

$$- \sum_{\substack{\tilde{s}_1\\ \tilde{s}_1\neq s}} \sum_{\substack{\tilde{\imath}_1,\tilde{\imath}_2\\ \tilde{\imath}_2<\tilde{\imath}_1}} \frac{\partial P_s}{\partial r_{\tilde{s}_1}^{\tilde{\imath}_1}} E_0\left(\frac{\partial R_{\tilde{s}_1}^{\tilde{\imath}_1}}{\partial R_s^{\tilde{\imath}_2}}\right)[r_s^{\tilde{\imath}_2} - E_0(R_s^{\tilde{\imath}_2})]$$

$$+ \frac{1}{2} \sum_{\substack{\tilde{s}_1,\tilde{s}_2\\ \tilde{s}_1\neq s,\tilde{s}_2\neq s}} \sum_{\tilde{\imath}_1,\tilde{\imath}_2} \frac{\partial^2 P_s}{\partial r_{\tilde{s}_1}^{\tilde{\imath}_1}\partial r_{\tilde{s}_2}^{\tilde{\imath}_2}} E_0(R_{\tilde{s}_1}^{\tilde{\imath}_1})E_0(R_{\tilde{s}_2}^{\tilde{\imath}_2}) + O(\bar{W}^3). \tag{3.8}$$

Each term on the right-hand side of (3.8) has a significant meaning and illustrates the process of approximating a full network (2.2) by an averaged model (2.3). The sum on the first line is simply the change in the probability distribution of node s caused by the average effect of connections from other nodes $\tilde{s}_1$. Intuitively, this change is the average activity of node $\tilde{s}_1$ times the effect of node $\tilde{s}_1$ on the probability distribution of node s (i.e., the derivative of P_s). Since the effective distribution P_s includes the average influence of connections from other nodes, this term must be subtracted from P_s to regain the original uncoupled distribution $\bar{P}_s$.

The term from the second line accounts for second-order effects from a connection from node $\tilde{s}_1$. First consider the case where $\tilde{\imath}_1 = \tilde{\imath}_2$. Now imagine that the effect of the connection from node $\tilde{s}_1$ onto node s lasts multiple time steps.[2] Then the connection from node $\tilde{s}_1$ will introduce correlations in the activity of node s. (Recall how common input from a node onto two different nodes can introduce correlations between those two nodes. The effect of the second line of (3.8) is identical except that in this case we have "common input" onto the same node but at different times, which creates correlations within that one node's activity.) This correlation will be proportional to the variance of $R_{\tilde{s}_1}^{\tilde{\imath}_1}$.

The case with $\tilde{\imath}_1 \neq \tilde{\imath}_2$ is similar. If $R_{\tilde{s}_1}^{\tilde{\imath}_1}$ is correlated with $R_{\tilde{s}_1}^{\tilde{\imath}_2}$ (due to the history dependence of the activity of node $\tilde{s}_1$), then the combined effect of the activity of node $\tilde{s}_1$ at times $\tilde{\imath}_1$ and $\tilde{\imath}_2$ will induce correlations in the activity of node s. This correlation will be proportional to the covariance of $R_{\tilde{s}_1}^{\tilde{\imath}_1}$ and $R_{\tilde{s}_1}^{\tilde{\imath}_2}$.

The reason this source of correlation must be subtracted from P_s in the second line of (3.8) is as follows. When fitting the averaged model (2.3) for node s, one is averaging over the activity of all other nodes, including node $\tilde{s}_1$. The induced correlations due to the connection from node $\tilde{s}_1$ will still be present in the activity of node s. Hence, the averaged model P_s (and its parameters θ_s) will take into account this additional correlation, and the additional correlation will appear in the averaged model as part of the history dependence of node s. However, the original uncoupled model represented by $\bar{P}_s$ (and its parameters $\bar{\theta}_s$) will not include effects due to coupling from other nodes. This history dependence of $\bar{P}_s$ would not include these additional correlations due to the connection from node $\tilde{s}_1$. Hence, the effect of these correlations must be subtracted from the effective distribution P_s to regain the original distribution $\bar{P}_s$, as is done in the second line of (3.8).

The term from the third line of (3.8) is similar in that it accounts for additional correlations in the activity of node s due to connections involving other nodes. In this case, the correlations are induced by indirect connections from node s onto itself via one of the other nodes $\tilde{s}_1$. This effect has three components as shown by the three factors. The right factor is the deviation of the activity of node s at time $\tilde{\imath}_2$

[2] Since P_s, as defined in (3.5), models the activity of node s for all time steps, the derivative $\partial^2 P_s/(\partial r_{\tilde{s}_1}^{\tilde{\imath}_1})^2$ includes the effects of $R_{\tilde{s}_1}^{\tilde{\imath}_1}$ on the activity of node s at all times. In particular, this derivative captures how the activity of node $\tilde{s}_1$ at a single time point $\tilde{\imath}_1$ can influence the activity of node s at two different times, thus causing correlations in the activity of node s at those two times.

from its expected activity as predicted by the averaged model (2.3). The middle factor is the effect of the activity of node s at time $\tilde{\imath}_2$ on the activity of node $\tilde{s}_1$ at time $\tilde{\imath}_1$. The left factor is the effect of the activity of node $\tilde{s}_1$ at time $\tilde{\imath}_1$ on the probability distribution of node s. The resulting correlation in the activity of node s from this chain of connection would be included in the history dependence of the effective distribution P_s. But, since these correlations depend on connections, their effect would not be included in the original uncoupled distribution $\bar{P}_s$. Hence, their effect must be subtracted from P_s to regain the original distribution $\bar{P}_s$.

The sum from the last line of (3.8) is simply a second-order effect of single connections onto node s. Equation (3.8) is accurate up to second order in $\bar{W}$. The sum of the first line is only a first-order approximation of the change in P_s due to the average effect of connections from other nodes. The addition of the last line gives the correct second-order approximation.

3.3. Step three: Measured node distribution in terms of effective parameters.

Our third step is to derive an expression for the probability distribution $\Pr(\mathbf{R}_\mathcal{Q}|\mathbf{X})$ of the measured node activity in terms of the effective parameters θ^i_s. Once we have written down an initial form of this distribution, we can simplify it by grouping the effects of hidden nodes into two sets of parameters: an effective causal connection W and an effective common input U. Then, by making one further assumption, we can sufficiently reduce the degrees of freedom within W and U so that computing their solution becomes tractable.

3.3.1. The initial form of the measured node probability distribution.

In the first step of our analysis, we obtained a lengthy expression for $\Pr(\mathbf{R}_\mathcal{Q}|\mathbf{X})$, the probability distribution of the measured node activity. (It is given by (A.5) in Appendix A.1.) However, this expression is in terms of the original model parameters $\bar{\theta}^i_s$ which remain unknown. As outlined in Appendix A.3, we rewrite this expression in terms of the effective parameters. Appendix A.4 describes how we transform the result into the form of a true probability (which we need since we wish to use it to develop maximum likelihood estimates of network parameters). We show in Appendix A.4 that this step requires one small deviation from a true second-order approximation, so we will use the ≈ symbol in our result. We also use the shorthand notation[3]

$$\begin{aligned} P^i_s &= P_s\big(r^i_s, \mathbf{r}^{<i}_s, \mathbf{x}, 0; \theta^i_s\big), \\ \frac{\partial P^i_s}{\partial w} &= \frac{\partial}{\partial w} P_s\big(r^i_s, \mathbf{r}^{<i}_s, \mathbf{x}, w; \theta^i_s\big)\Big|_{w=0}. \end{aligned} \tag{3.9}$$

In the end, we obtain the following expression for the probability distribution of the measured nodes' activity:

$$\Pr(\mathbf{R}_\mathcal{Q} = \mathbf{r}_\mathcal{Q}|\mathbf{X} = \mathbf{x}) \approx \prod_q \prod_i P_q\big(r^i_q, \mathbf{r}^{<i}_q, \mathbf{x}, \widetilde{W}^i_q; \theta^i_q\big) + O(\bar{W}^3), \tag{3.10a}$$

[3] Note the subtle difference between the new notation P^i_s and $\partial P^i_s/\partial w$ (defined by (3.9)) on one hand and the similar notation P_s and $\partial P_s/\partial r^{\tilde{\imath}}_{\tilde{s}}$ (defined by (3.5) and (3.6)) on the other hand. One key difference is that the new notation contains a superscript i, which means it refers to the distribution of the activity of node s just at time point i.

where

$$
\begin{aligned}
\widetilde{W}_q^i = &\sum_{\substack{\tilde{q}\\ \tilde{q}\neq q}} \sum_{\substack{\tilde{\imath}_1\\ \tilde{\imath}_1<i}} \bar{W}_{\tilde{q},q}^{\tilde{\imath}_1,i} [r_{\tilde{q}}^{\tilde{\imath}_1} - E_0(R_{\tilde{q}}^{\tilde{\imath}_1})] \\
&+ \sum_{\substack{p,\tilde{q}\\ \tilde{q}\neq q}} \sum_{\substack{\tilde{\imath}_1,\tilde{\imath}_2\\ \tilde{\imath}_2<\tilde{\imath}_1<i}} \bar{W}_{p,q}^{\tilde{\imath}_1,i} E_0\left(\frac{\partial R_p^{\tilde{\imath}_1}}{\partial R_{\tilde{q}}^{\tilde{\imath}_2}}\right)[r_{\tilde{q}}^{\tilde{\imath}_2} - E_0(R_{\tilde{q}}^{\tilde{\imath}_2})] \\
&+ \sum_{\substack{p,\tilde{q}\\ \tilde{q}\neq q}} \sum_{\substack{\tilde{\imath}_1,\tilde{\imath}_2,\tilde{\imath}_3\\ \tilde{\imath}_2<\tilde{\imath}_3<i\\ \tilde{\imath}_1<i}} \bar{W}_{p,q}^{\tilde{\imath}_1,i} \bar{W}_{p,\tilde{q}}^{\tilde{\imath}_2,\tilde{\imath}_3} \frac{\partial P_{\tilde{q}}^{\tilde{\imath}_3}}{\partial w} \frac{1}{P_{\tilde{q}}^{\tilde{\imath}_3}} [E_0(R_p^{\tilde{\imath}_1} R_p^{\tilde{\imath}_2}) - E_0(R_p^{\tilde{\imath}_1}) E_0(R_p^{\tilde{\imath}_2})] \\
&+ \sum_{\substack{p,\tilde{q}\\ \tilde{q}<q}} \sum_{\substack{\tilde{\imath}_1,\tilde{\imath}_2\\ \tilde{\imath}_1,\tilde{\imath}_2<i}} \bar{W}_{p,q}^{\tilde{\imath}_1,i} \bar{W}_{p,\tilde{q}}^{\tilde{\imath}_2,i} \frac{\partial P_{\tilde{q}}^{i}}{\partial w} \frac{1}{P_{\tilde{q}}^{i}} [E_0(R_p^{\tilde{\imath}_1} R_p^{\tilde{\imath}_2}) - E_0(R_p^{\tilde{\imath}_1}) E_0(R_p^{\tilde{\imath}_2})] \\
&- \sum_{\substack{\tilde{q}\\ \tilde{q}\neq q}} \sum_{\substack{\tilde{\imath}_1,\tilde{\imath}_2,\tilde{\imath}_3\\ \tilde{\imath}_2<\tilde{\imath}_3<i\\ \tilde{\imath}_1<i}} \bar{W}_{\tilde{q},q}^{\tilde{\imath}_1,i} \bar{W}_{\tilde{q},q}^{\tilde{\imath}_2,\tilde{\imath}_3} \frac{\partial P_{q}^{\tilde{\imath}_3}}{\partial w} \frac{1}{P_{q}^{\tilde{\imath}_3}} [E_0(R_{\tilde{q}}^{\tilde{\imath}_1} R_{\tilde{q}}^{\tilde{\imath}_2}) - E_0(R_{\tilde{q}}^{\tilde{\imath}_1}) E_0(R_{\tilde{q}}^{\tilde{\imath}_2})] \\
&- \sum_{\substack{\tilde{q}\\ \tilde{q}\neq q}} \sum_{\substack{\tilde{\imath}_1,\tilde{\imath}_2\\ \tilde{\imath}_2<\tilde{\imath}_1<i}} \bar{W}_{\tilde{q},q}^{\tilde{\imath}_1,i} E_0\left(\frac{\partial R_{\tilde{q}}^{\tilde{\imath}_1}}{\partial R_{q}^{\tilde{\imath}_2}}\right)[r_{q}^{\tilde{\imath}_2} - E_0(R_{q}^{\tilde{\imath}_2})].
\end{aligned}
\tag{3.10b}
$$

Though this expression is somewhat lengthy, each line of (3.10b) represents the effect of a connection or combination of connections on the probability distribution of the measured nodes' activity. Just as we did for the single-node results (3.8), we briefly describe the effects of the connections as embedded in (3.10).

The sum from the first line of (3.10b) represents a direct causal connection from measured node $\tilde{q}$ onto measured node q. The second line represents an indirect causal connection from measured node $\tilde{q}$ onto measured node q via a hidden node p. Both lines describe a change in the distribution of the activity of node q due to a deviation in the activity of node $\tilde{q}$ from that predicted by the averaged model.

The third and fourth lines are the common input onto measured nodes q and $\tilde{q}$ from a hidden node p. The common input effect is proportional to the (unknown) (co)variance of the activity of node p (compare to the second line of (3.8)). We separated out the common input that reaches nodes q and $\tilde{q}$ simultaneously (fourth line of (3.10b)). We arbitrarily put this common input effect into the $\widetilde{W}_q$ of the node with the higher index (as we restrict the sum to $\tilde{q} < q$). The goal of this analysis will be to distinguish the common input from the third line from the causal connections of the first two lines. (We will assume any correlations at zero delay are due to the common input described on the fourth line.)

The fifth and sixth lines of (3.10b) involve only measured nodes. These lines are similar to the second and third lines of (3.8), and their presence in (3.10) has a similar origin. When the effective parameters of node q were determined, the activity of node $\tilde{q}$ was ignored. Nonetheless, connections from node $\tilde{q}$ onto node q still influenced the activity of node q. As we described in the context of (3.8), the activity of node $\tilde{q}$ could induce correlations in the activity of node q if it had connections onto node q that lasted multiple time steps. Similarly, node $\tilde{q}$ could induce correlations in node q via an indirect connection from node q onto itself through node $\tilde{q}$.

If the network contains such a pattern of connections, the effective distribution P_q of node q would already contain such correlations as part of the history dependence of the model. Hence, the probability distribution of $R_{\mathcal{Q}}$ in (3.10) would contain these correlations in the activity of node q even if all $\bar{W}$ were zero. However, these correlations in the activity of node q were caused by connections (i.e., nozero $\bar{W}$) between node q and $\tilde{q}$, as described above. When these individual connections are added to (3.10) via the direct causal connections of the first line of (3.10b), the resulting correlations in the activity of node q will have been added to (3.10) twice. To correct for this, we need to explicitly subtract them off via the fifth and sixth lines of (3.10b).

3.3.2. Grouping the effects of hidden nodes. Once the effective parameters θ^i_q have been determined for all measured nodes q, the only unknowns in (3.10) are the connectivity factor $\bar{W}$ and all expressions involving hidden nodes p. We group these unknowns into two expressions:

$$\begin{aligned}
W^{i_2,i_1}_{q_2,q_1} &= \bar{W}^{i_2,i_1}_{q_2,q_1} + \sum_{p} \sum_{\substack{\tilde{i} \\ i_1>\tilde{i}>i_2}} \bar{W}^{\tilde{i},i_1}_{p,q_1} E_0\left(\frac{\partial R^{\tilde{i}}_p}{\partial R^{i_2}_{q_2}}\right), \\
U^{i_2,i_1}_{q_2,q_1} &= \sum_{p} \sum_{\substack{\tilde{i}_1,\tilde{i}_2 \\ \tilde{i}_1<i_1,\tilde{i}_2<i_2}} \bar{W}^{\tilde{i}_1,i_1}_{p,q_1} \bar{W}^{\tilde{i}_2,i_2}_{p,q_2} [E_0(R^{\tilde{i}_1}_p R^{\tilde{i}_2}_p) - E_0(R^{\tilde{i}_1}_p) E_0(R^{\tilde{i}_2}_p)],
\end{aligned} \tag{3.11}$$

defined for $q_2 \neq q_1$. The causal connection factor $W^{i_2,i_1}_{q_2,q_1}$ is the effective causal connection from node q_2 onto node q_1. It includes an indirect causal connection via a hidden node p. The direct and indirect causal connections are lumped together as we cannot distinguish between them. The common input factor $U^{i_2,i_1}_{q_2,q_1}$ is the effective common input from hidden nodes that arrives at node q_2 at time i_2 and at node q_1 and time i_1. Both $W^{i_2,i_1}_{q_2,q_1}$ and $U^{i_2,i_1}_{q_2,q_1}$ include sums over arbitrary hidden nodes. Although we cannot resolve the individual contributions of the hidden nodes, we will be able to solve for these effective parameters.

We also rewrite the expression for the expected value of the derivative to pull out the hidden factor of $\bar{W}$ contained in it. From the definition (3.7) as well as the definitions of the derivatives (3.6) and (3.9), we write[4]

$$\begin{aligned}
E_0\left(\frac{\partial R^i_s}{\partial R^{\tilde{i}}_{\tilde{s}}}\right) &= \sum_{\mathbf{r}_s} r^i_s \frac{\partial P_s}{\partial r^{\tilde{i}}_{\tilde{s}}} = \sum_{\mathbf{r}_s} r^i_s \sum_{\substack{i_2 \\ \tilde{i}<i_2\leq i}} \bar{W}^{\tilde{i},i_2}_{\tilde{s},s} \frac{\partial P^{i_2}_s}{\partial w} \frac{1}{P^{i_2}_s} \prod_{i_3} P^{i_3}_s \\
&= \sum_{\substack{i_2 \\ \tilde{i}<i_2\leq i}} \bar{W}^{\tilde{i},i_2}_{\tilde{s},s} E_0\left(R^i_s \frac{\partial P^{i_2}_s}{\partial w} \frac{1}{P^{i_2}_s}\right).
\end{aligned} \tag{3.12}$$

With these definitions of W and U, $\widetilde{W}^i_q$ becomes

$$\begin{aligned}
\widetilde{W}^i_q = &\sum_{\substack{\tilde{q} \\ \tilde{q}\neq q}} \sum_{\substack{\tilde{i} \\ \tilde{i}<i}} W^{\tilde{i},i}_{\tilde{q},q} [r^{\tilde{i}}_{\tilde{q}} - E_0(R^{\tilde{i}}_{\tilde{q}})] \\
&+ \sum_{\substack{\tilde{q} \\ \tilde{q}\neq q}} \sum_{\substack{\tilde{i} \\ \tilde{i}<i}} U^{\tilde{i},i}_{\tilde{q},q} \frac{\partial P^{\tilde{i}}_{\tilde{q}}}{\partial w} \frac{1}{P^{\tilde{i}}_{\tilde{q}}} + \sum_{\substack{\tilde{q} \\ \tilde{q}<q}} U^{i,i}_{\tilde{q},q} \frac{\partial P^i_{\tilde{q}}}{\partial w} \frac{1}{P^i_{\tilde{q}}}
\end{aligned}$$

[4] One subtlety in (3.12) is the fact that we restrict $i_2 \leq i$. If $i_2 > i$, then the term disappears due to a similar argument as underlying the identities in (A.3).

$$
\begin{aligned}
&-\sum_{\substack{\tilde{q}\\ \tilde{q}\neq q}} \sum_{\substack{\tilde{\imath}_1,\tilde{\imath}_2,\tilde{\imath}_3\\ \tilde{\imath}_2<\tilde{\imath}_3<i\\ \tilde{\imath}_1<i}} W^{\tilde{\imath}_1,i}_{\tilde{q},q} W^{\tilde{\imath}_2,\tilde{\imath}_3}_{\tilde{q},q} \frac{\partial P^{\tilde{\imath}_3}_q}{\partial w} \, \frac{1}{P^{\tilde{\imath}_3}_q} [E_0(R^{\tilde{\imath}_1}_{\tilde{q}} R^{\tilde{\imath}_2}_{\tilde{q}}) - E_0(R^{\tilde{\imath}_1}_{\tilde{q}}) E_0(R^{\tilde{\imath}_2}_{\tilde{q}})]\\
&-\sum_{\substack{\tilde{q}\\ \tilde{q}\neq q}} \sum_{\substack{\tilde{\imath}_1,\tilde{\imath}_2,\tilde{\imath}_3\\ \tilde{\imath}_2<\tilde{\imath}_3\leq\tilde{\imath}_1<i}} W^{\tilde{\imath}_2,\tilde{\imath}_3}_{q,\tilde{q}} W^{\tilde{\imath}_1,i}_{\tilde{q},q} E_0\left(R^{\tilde{\imath}_1}_{\tilde{q}} \frac{\partial P^{\tilde{\imath}_3}_{\tilde{q}}}{\partial w} \frac{1}{P^{\tilde{\imath}_3}_{\tilde{q}}} \right) [r^{\tilde{\imath}_2}_q - E_0(R^{\tilde{\imath}_2}_q)]. \qquad (3.13)
\end{aligned}
$$

Note that, according to (3.12), the quantity $E_0\big(\partial R^{\tilde{\imath}}_{\tilde{s}}/\partial R^i_s\big)$ is $O(\bar{W})$. Hence, the definition (3.11) of W shows that W is a first-order approximation to $\bar{W}$ (i.e., $W^{\tilde{\imath},i}_{\tilde{q},q} = \bar{W}^{\tilde{\imath},i}_{\tilde{q},q} + O(\bar{W}^2)$). This means that, in terms that are quadratic in $\bar{W}$, we can replace $\bar{W}$ with W and still maintain our second-order approximation (as the error is cubic in $\bar{W}$). This allowed us to write (3.13) in terms of just the effective W.

Our expression for the probability distribution $\Pr(\mathbf{R}_{\mathcal{Q}}|\mathbf{X})$ of the measured activity is now (3.10a) combined with (3.13). Given the effective parameters θ^i_q, we can calculate the P^i_q and the $\partial P^i_q/\partial w$ via (3.9). We can also calculate, in principle, all of the expressions involving $E_0(\cdot)$ using the definitions in (3.7). (We estimate these via Monte Carlo simulations, as described in Appendix C.) Therefore, the only remaining unknown factors are the causal connection factors W and the common input factors U.

3.3.3. A further assumption for a tractable solution. Our goal is to estimate W and U by finding their values that maximize our approximation of the probability distribution of measured activity. In other words, we seek maximum likelihood estimators of W and U. However, there are still too many unknowns to make the solution tractable, as we still have more unknowns than we would have data points (we have only one measurement of activity per measured node per time point).[5] To reduce the number of unknowns, we assume that W and U depend only on the difference between their temporal indicies, i.e.,

$$W^{i-j,i}_{q_1,q_2} = W^j_{q_1,q_2} \qquad \text{and} \qquad U^{i-j,i}_{q_1,q_2} = U^j_{q_1,q_2}. \qquad (3.14)$$

(One could presumably weaken this assumption by allowing W and U to change slowly over time at the cost of additional computational complexity and increased data requirements.)

This assumption for W has no hidden surprises, as it is equivalent to assuming that the underlying connectivity $\bar{W}$ depends only on the difference in temporal indicies.[6] However, this assumption for U is more significant than may appear at first glance. It turns out that this assumption is really about the hidden nodes and affects how one can interpret the meaning of W and U.

To demonstrate this, we rewrite the definition of U from (3.11) using the index j to indicate the difference between temporal indices:

$$
\begin{aligned}
&U^{i_1-j_1,i_1}_{q_2,q_1} =\\
&\sum_p \sum_{\substack{j_2,j_3\\ j_2>0,j_3>0}} \bar{W}^{i_1-j_2,i_1}_{p,q_1} \bar{W}^{i_1-j_1-j_3,i_1-j_1}_{p,q_2} [E_0(R^{i_1-j_2}_p R^{i_1-j_1-j_3}_p) - E_0(R^{i_1-j_2}_p) E_0(R^{i_1-j_1-j_3}_p)].
\end{aligned}
$$

[5] Perhaps one could solve for W and U in full generality if one could repeatedly sample from a small number of time bins and one assumed that the $\bar{W}$ could vary over the time bins but were identical for each repetition.

[6] The effect of the connectivity, however, could vary with time, as each $P_s(\cdot)$ could change with time.

Our assumption on $U^{i_1-j_1,i_1}_{q_2,q_1}$ is that it is independent of i_1. If the $\bar{W}$ depend only on the difference in temporal indices, the only place on the right-hand side where i_1 doesn't immediately drop out is in the (co)variance of activity of the hidden node p. Hence, by insisting that $U^{i_1-j_1,i_1}_{q_2,q_1}$ be independent of i_1, we are really approximating the covariance of each hidden node p as though it were independent of time bin i_1. Equivalently, we could view this approximation as replacing the covariance of node p with its average over all time bins i_1.

As detailed in [14], such an approximation leads to a certain degree of ambiguity in the identification of causal connections, which we refer to as *subpopulation ambiguity.* This ambiguity contains subtleties that are out of the scope of this article and are discussed extensively in [14]. We illustrate the basic consequences of the ambiguity with simulation results (see section 4.4). Note also that, as described in [14], this ambiguity is already present in many experimental contexts (such as those commonly used in neuroscience); hence, in those contexts, this approximation does not add additional ambiguity.

Putting this all together, our procedure to construct the causal connections among measured nodes is as follows. For each measured node indexed by $q \in \mathcal{Q}$, determine the effective parameters θ^i_q by fitting the averaged model (2.3) to the external variables $\mathbf{X}$ and the activity R^i_q of node q. (We assume such an algorithm for determining the θ^i_q is known.) Then determine the effective causal connections $W^j_{q_1,q_2}$ and the effective common input $U^j_{q_1,q_2}$ from the external variables and the activity $\mathbf{R}_{\mathcal{Q}}$ of all measured nodes by finding the values of $W^j_{q_1,q_2}$ and $U^j_{q_1,q_2}$ that maximize the log-likelihood modeled by the equation

$$\log \Pr(\mathbf{R}_{\mathcal{Q}} = \mathbf{r}_{\mathcal{Q}} | \mathbf{X} = \mathbf{x}) = \sum_q \sum_i \log P_q\big(r^i_q, \mathbf{r}^{<i}_q, \mathbf{x}, \widetilde{W}^i_q; \theta^i_q\big), \tag{3.15a}$$

where

$$\begin{aligned}\widetilde{W}^i_q = &\sum_{\substack{\tilde{q}\\ \tilde{q}\neq q}} \sum_{\substack{j\\ j>0}} W^j_{\tilde{q},q}[r^{i-j}_{\tilde{q}} - E_0(R^{i-j}_{\tilde{q}})]\\
&+ \sum_{\substack{\tilde{q}\\ \tilde{q}\neq q}} \sum_{\substack{j\\ j>0}} U^j_{\tilde{q},q} \frac{\partial P^{i-j}_{\tilde{q}}}{\partial w} \frac{1}{P^{i-j}_{\tilde{q}}} + \sum_{\substack{\tilde{q}\\ \tilde{q}<q}} U^0_{\tilde{q},q} \frac{\partial P^i_{\tilde{q}}}{\partial w} \frac{1}{P^i_{\tilde{q}}}\\
&- \sum_{\substack{\tilde{q}\\ \tilde{q}\neq q}} \sum_{\substack{j_1,j_2,j_3\\ j_1,j_2,j_3>0}} W^{j_1}_{\tilde{q},q} W^{j_2}_{\tilde{q},q} \frac{\partial P^{i-j_3}_q}{\partial w} \frac{1}{P^{i-j_3}_q} [E_0(R^{i-j_1}_{\tilde{q}} R^{i-j_3-j_2}_{\tilde{q}}) - E_0(R^{i-j_1}_{\tilde{q}}) E_0(R^{i-j_3-j_2}_{\tilde{q}})]\\
&- \sum_{\substack{\tilde{q}\\ \tilde{q}\neq q}} \sum_{\substack{j_1,j_2,j_3\\ j_1,j_2>0\\ j_3\geq 0}} W^{j_2}_{q,\tilde{q}} W^{j_1}_{\tilde{q},q} E_0\bigg(R^{i-j_1}_{\tilde{q}} \frac{\partial P^{i-j_1-j_3}_{\tilde{q}}}{\partial w} \frac{1}{P^{i-j_1-j_3}_{\tilde{q}}}\bigg)[r^{i-j_1-j_3-j_2}_q - E_0(xR^{i-j_1-j_3-j_2}_q)].\end{aligned}$$

(3.15b)

Unfortunately, especially with the terms that are quadratic in W, one cannot be certain that the log-likelihood is free of nonglobal local maxima. So, in general, one needs to be aware that one could get trapped in such a local maximum in the process of looking for the global maximum. The likelihood surface may be better behaved if one ignores the quadratic terms (the final two terms in (3.15b)). We next present an example probability distribution P_s where the likelihood surface has no nonglobal local maxima in the absence of the quadratic terms. We use that fact to find a local maximum of the full log-likelihood that, at least in our tests, gives good results.

3.4. Special case: A Poisson distribution. We present a special case of the results when the activity of each node at each time step is drawn from a Poisson distribution. We use such a distribution because, for small time bins, the averaged model approximates a generic history-dependent point process [4, 5], which one can use to model the spike times of a neuron. Moreover, the results with the Poisson distribution illustrate how history dependence can distinguish common input from causal connections, as discussed below. We use the Poisson model when we demonstrate the results via simulations.

Since we assume that R^i_s, the activity of node s at time bin i, is a Poisson random variable, we simply need to specify its mean. We can write the probability distribution of R^i_s as

$$P_s(r^i_s, \mathbf{r}^{<i}_s, \mathbf{x}, w; \theta^i_s) = \Gamma(r^i_s, \lambda_s(\mathbf{r}^{<i}_s, \mathbf{x}, w; \theta^i_s)), \tag{3.16a}$$

where

$$\Gamma(n, \lambda) = \frac{1}{n!}\lambda^n e^{-\lambda}. \tag{3.16b}$$

The function $\lambda_s(\mathbf{r}^{<i}_s, \mathbf{x}, w; \theta^i_s)$ defines how the expected value of R^i_s depends on the history $\mathbf{r}^{<i}_s$ of node s, the external variables $\mathbf{x}$, and the total input w from other neurons. We rewrite the log-likelihood (3.15) as

$$\log \Pr(\mathbf{R}_{\mathcal{Q}} = \mathbf{r}_{\mathcal{Q}} | \mathbf{X} = \mathbf{x}) = \sum_{q,i} r^i_q \log \lambda_q(\mathbf{r}^{<i}_q, \mathbf{x}, \widetilde{W}^i_q; \theta^i_q) - \sum_{q,i} \lambda_q(\mathbf{r}^{<i}_q, \mathbf{x}, \widetilde{W}^i_q; \theta^i_q) + C, \tag{3.17a}$$

where

$$\begin{aligned}\widetilde{W}^i_q = {} & \sum_{\substack{\tilde{q}\\ \tilde{q}\neq q}} \sum_{\substack{j\\ j>0}} W^j_{\tilde{q},q}[r^{i-j}_{\tilde{q}} - E_0(\lambda_{\tilde{q}}(\mathbf{R}^{<i-j}_{\tilde{q}}, \mathbf{x}, 0; \theta^{i-j}_{\tilde{q}}))] \\ & + \sum_{\substack{\tilde{q}\\ \tilde{q}\neq q}} \sum_{\substack{j\\ j>0}} U^j_{\tilde{q},q}[r^{i-j}_{\tilde{q}} - \lambda_{\tilde{q}}(\mathbf{r}^{<i-j}_{\tilde{q}}, \mathbf{x}, 0; \theta^{i-j}_{\tilde{q}})] \frac{\partial_w \lambda_{\tilde{q}}(\mathbf{r}^{<i-j}_{\tilde{q}}, \mathbf{x}, 0; \theta^{i-j}_{\tilde{q}})}{\lambda_{\tilde{q}}(\mathbf{r}^{<i-j}_{\tilde{q}}, \mathbf{x}, 0; \theta^{i-j}_{\tilde{q}})} \\ & + \sum_{\substack{\tilde{q}\\ \tilde{q}<q}} U^0_{\tilde{q},q}[r^i_{\tilde{q}} - \lambda_{\tilde{q}}(\mathbf{r}^{<i}_{\tilde{q}}, \mathbf{x}, 0; \theta^i_{\tilde{q}})] \frac{\partial_w \lambda_{\tilde{q}}(\mathbf{r}^{<i}_{\tilde{q}}, \mathbf{x}, 0; \theta^i_{\tilde{q}})}{\lambda_{\tilde{q}}(\mathbf{r}^{<i}_{\tilde{q}}, \mathbf{x}, 0; \theta^i_{\tilde{q}})} \\ & + \text{quadratic terms.}\end{aligned} \tag{3.17b}$$

The constant $C = -\sum_{q,i} \log((r^i_q)!)$ can be ignored since we simply want to maximize (3.17) over W and U with everything else fixed. We use the notation $\partial_w \lambda_{\tilde{q}}(\mathbf{r}^{<i-j}_{\tilde{q}}, \mathbf{x}, w; \theta^{i-j}_{\tilde{q}})$ for the partial derivative of $\lambda_{\tilde{q}}(\cdot)$ with respect to w.

The quadratic terms are the last two lines of (3.15b); we gain no insight by rewriting them in terms of the Poisson distribution. As detailed in section 3.3.1, they are needed to have a correct second-order expression. However, they don't directly contribute to the distinction between causal connections and common input.

3.4.1. Different effects of causal connections and common input. From (3.17), we see two important differences in the way that the causal connections W and the common input U affect the probability distribution $\Pr(R_{\mathcal{Q}}|\mathbf{X})$ of the measured node activity. Our ability to successfully distinguish causal connections from common input connections is based on these two differences.

The first difference is that the common input terms have an additional $\partial_w \lambda_{\tilde{q}}/\lambda_{\tilde{q}}$ factor. In previous work [14], this factor was the only difference that appeared because the analysis did not exploit history-dependent effects. As detailed in [14], this difference alone can distinguish causal connections from common input in many cases. Even if one did not model history-dependent effects, the relationship among external variables (such as a stimulus) and the activity of measured nodes would distinguish common input from causal connection, and this difference is captured by the $\partial_w \lambda_{\tilde{q}}/\lambda_{\tilde{q}}$ factor.

The second difference in the way W and U appear in (3.17) is due to the history-dependent effects. This second difference is the focus of this paper. It turns out that this difference is exactly what we observed in the exaggerated example presented in the introduction and illustrated in Figure 2. For both the causal connection W term and the common input U term of (3.17b), a certain quantity is subtracted from the activity $r_{\tilde{q}}^{i-j}$. The difference between these quantities can distinguish a causal connection from common input. In what follows, we will show that the key difference is that the activity predicted by a node's history dependence is subtracted only from the common input term.

Equation (3.17) shows that a causal connection from node $\tilde{q}$ onto node q induces a change in the probability distribution of node q proportional to the deviation of the activity of node $\tilde{q}$ from that predicted by the averaged model (2.3). That is, in the causal connection term (first line of (3.17b)), a contribution is added to $\widetilde{W}_q^i$ when the measured activity of node $\tilde{q}$ (i.e., $r_{\tilde{q}}^{i-j}$) differs from its expected value $E_0(R_{\tilde{q}}^{i-j}) = E_0(\lambda_{\tilde{q}}(\mathbf{R}_{\tilde{q}}^{<i-j}, \mathbf{x}, 0; \theta_{\tilde{q}}^{i-j}))$ given by the averaged model.

An important point is that, once the effective parameters ($\theta_{\tilde{q}}^{\tilde{\imath}}$ for all $\tilde{\imath}$) have been determined, this expected value $E_0(R_{\tilde{q}}^{i-j})$ does not depend on the actual history $\mathbf{r}_{\tilde{q}}^{<i-j}$ of node $\tilde{q}$. This expected value is an average over all possible histories of node $\tilde{q}$, given the effective parameters and the external variables.[7] For example, imagine that $R_{\tilde{q}}^i$ corresponds to the number of spikes of neuron $\tilde{q}$ in time bin i (with a sufficiently small time bin so that $R_{\tilde{q}}^i > 1$ with vanishingly small probability). Imagine, moreover, that (similar to neuron 1 in Figure 2) neuron $\tilde{q}$ tended to spike in pairs so that if it spiked in time bin $i-1$ but not in time bin $i-2$, it was very likely to spike in time bin i: $\Pr(R_{\tilde{q}}^i = 1|R_{\tilde{q}}^{i-1} = 1 \ \&\ R_{\tilde{q}}^{i-2} = 0) \approx 1$. If one used an appropriate model, then the averaged model (2.3) would capture this tendency to fire in pairs once the parameters $\theta_{\tilde{q}}$ were fit to the spikes $\mathbf{R}_{\tilde{q}}$ of neuron $\tilde{q}$. Even so, the expected value $E_0(R_{\tilde{q}}^i)$ would not depend on the presence or absence of spikes in the previous two time bins; it is independent of the specific history of node $\tilde{q}$. Even if $r_{\tilde{q}}^{i-1} = 1$ and $r_{\tilde{q}}^{i-2} = 0$, the expected value $E_0(R_{\tilde{q}}^i)$ would not be close to one. If indeed $r_{\tilde{q}}^i = 1$, then both the spike at time bin $i-1$ and the spike at time bin i would contribute equally to the causal connection term in the first line of (3.17b).[8]

[7] We calculate this value via Monte Carlo. We repeatedly generate a realization of the activity of node $\tilde{q}$ for all time points according to the averaged model (2.3). The average activity at each time point i over many such realizations is our estimate of $E_0(R_{\tilde{q}}^i)$. See Appendix C.

[8] One would get a similar result if neuron $\tilde{q}$ had a *refractory period* where, for example, it could not spike in time i if it spiked in time bin $i-1$: $\Pr(R_{\tilde{q}}^i = 1|R_{\tilde{q}}^{i-1} = 1) = 0$. Even if neuron $\tilde{q}$ did spike at time bin $i-1$, the presence of the refractory period would not affect $E_0(R_{\tilde{q}}^i)$.

We contrast this observation with the common input term from the second line of (3.17b). In the common input term, the activity $r_{\tilde{q}}^{i-j}$ of node $\tilde{q}$ is subtracted by the mean $\lambda_{\tilde{q}}$ of the Poisson distribution, given the specific history $\mathbf{r}_{\tilde{q}}^{<i-j}$ measured from node $\tilde{q}$. Unlike the causal connection term, this quantity is the expected value of $R_{\tilde{q}}^{i-j}$, *conditioned on the measured history* $\mathbf{r}_{\tilde{q}}^{<i-j}$: $\lambda_{\tilde{q}}(\mathbf{r}_{\tilde{q}}^{<i-j}, \mathbf{x}, 0; \theta_{\tilde{q}}^{i-j}) = E_0(R_{\tilde{q}}^{i-j} \,|\, \mathbf{R}_{\tilde{q}}^{<i-j} = \mathbf{r}_{\tilde{q}}^{<i-j})$. This is still an expected value based on the averaged model, but it is not an average over all possible histories of node $\tilde{q}$. As above, imagine that node $\tilde{q}$ was a neuron that tended to fire pairs of spikes and that one used a model that accurately captured this firing pattern. Then, if $r_{\tilde{q}}^{i-1} = 1$ and $r_{\tilde{q}}^{i-2} = 0$, the expected value $\lambda_{\tilde{q}}(\mathbf{r}_{\tilde{q}}^{<i}, \mathbf{x}, 0; \theta_{\tilde{q}}^{i})$ would be close to one because the model predicts that neuron $\tilde{q}$ should immediately fire a second spike. If indeed $r_{\tilde{q}}^{i} = 1$, this spike would have little contribution to the second line of (3.17b) as $r_{\tilde{q}}^{i} - \lambda_{\tilde{q}}(\mathbf{r}_{\tilde{q}}^{<i}, \mathbf{x}, 0; \theta_{\tilde{q}}^{i})$ would be small.

Equation (3.17) therefore demonstrates that the intuition we gain from the exaggerated example of Figure 2 is applicable to the more realistic situation we used to derive (3.17). For example, although it isn't intuitively obvious what should happen when all nodes have strong history dependence, (3.17) shows that one may estimate the connectivity even in that case, provided one has a model through which one can accurately capture the history dependence of the measured nodes.

3.4.2. Tractable computation of maximum likelihood estimators. In order to efficiently compute maximum likelihood estimators of W and U, we'd like to make sure that any local maxima of the log-likelihood (3.17) are indeed global maxima. As discussed below, if one ignores the quadratic terms in (3.17b), one can develop a condition on the form of λ_s to ensure all local maxima are global maxima. One can then use the solution to the reduced problem (without quadratic terms) to guide the search for a solution to the full problem.

If we ignore the quadratic terms from (3.17b), then $\widetilde{W}_q^i$ is linear in W and U, and the log-likelihood (3.17) has the same form of dependence on W and U as discussed in [14]. Since concavity is preserved under addition and $r_q^i \geq 0$, the log-likelihood will be concave in W and U if $\lambda_q(\mathbf{r}_q^{<i}, \mathbf{x}, w; \theta_q^i)$ is convex in w and $\log \lambda_q(\mathbf{r}_q^{<i}, \mathbf{x}, w; \theta_q^i)$ is concave in w. Reference [17] describes this condition in a more general setting and outlines the ensuing requirements on λ_q, such as the fact that λ_q must be monotonically increasing in w and must grow at least linearly in w. If the log-likelihood is concave in W and U, there can be no nonglobal local maxima.

We base our search for a maximizer of the full log-likelihood (3.17) on the maximizer for reduced log-likelihood (ignoring the quadratic terms of (3.17b)). We form a homotopy from the reduced problem to the full problem by multiplying the quadratic terms by some number $\gamma \in [0, 1]$. After maximizing the reduced log-likelihood ($\gamma = 0$), we form a series of log-likelihoods with increasing γ. For each problem, we use the maximizer of the previous problem as the initial condition. We end up with a maximizer of the full log-likelihood ($\gamma = 1$). Although we cannot guarantee that we have found a global maximizer, we have achieved good results using this algorithm in our simulation tests. (To calculate the maximizer for a given γ, we iterate to a critical point of (3.17) using a modified version of Powell's hybrid method as implemented in the GNU Scientific Library [6].)

4. Results.

4.1. Overview of simulations. To test the performance of our analysis, we simulated small networks of simplified neurons responding to a stimulus $\mathbf{X}$. Our goal is to demonstrate that we can distinguish the common input and causal connection networks schematized in Figure 1, under the condition that the common input neuron is unmeasured.

4.1.1. The stimulus. The external variables $\mathbf{X}$ represented the same one-dimensional (i.e., constant along vertical lines) visual stimulus as detailed in [14]. This stimulus was a movie of a sequence of sinusoidal gratings $\mathbf{I}^k$ with wave number k. The jth line of $\mathbf{I}^k$ was $I^k_j = \operatorname{cas}(2\pi kj/N_0)$, where $\operatorname{cas} x = \cos x + \sin x$, $N_0 = 100$, and $0 \le j \le N_0 - 1$. Every 10 simulated milliseconds, a new image was selected, with replacement, from the set composed of the $\mathbf{I}^k$ and $-\mathbf{I}^k$, for $k = -10, -9, \dots, 9, 10$. The movie was one simulated minute long.

4.1.2. The simulated neurons. In the simulated networks, we let each neuron be a generalized linear model (also called a linear-nonlinear model). We discretized time into $\Delta t_{\text{sim}} = 0.5$ ms time bins. In each time bin i, we let the probability that a neuron spiked be a linear function of its spiking history, the stimulus $\mathbf{X}$, and previous spikes of other neurons, composed with a half-squaring nonlinearity,

$$
\begin{aligned}
&\Pr(R^i_p = 1|\mathbf{R}^{<i} = \mathbf{r}^{<i}, \mathbf{X} = \mathbf{x}) \\
&\qquad = A\Delta t_{\text{sim}} \left[\sum_{j>0} \bar{h}^j_{\text{hist},p} r^{i-j}_p + \bar{\mathbf{h}}^i_{\text{ext},p} \cdot \mathbf{x} + \sum_{q\neq p}\sum_{j>0} \bar{W}^j_{q,p} r^{i-j}_q + \bar{y}_p\right]^2_+ ,
\end{aligned} \tag{4.1}
$$

where $[y]^2_+ = y^2$ if $y > 0$ and is zero otherwise. The activity variable $R^i_p = 1$ if neuron p spiked in time bin i and $R^i_p = 0$ otherwise. We set $A = 0.01$ ms^{-1}. The value of the threshold parameters $\bar{y}_p$, coupling parameters $\bar{W}^j_{q,p}$, and other parameters that appear below are given in the context of specific simulations. If (4.1) resulted in a probability greater than one, it was truncated to one.

The linear kernel $\bar{\mathbf{h}}_{\text{hist},p}$ specified the spike-history dependence of neuron p. We included a refractory period of length τ^{ref}_p by setting $\bar{h}^j_{\text{hist},p} = -100$ for $j\Delta t_{\text{sim}} \le \tau^{\text{ref}}_p$. (Since -100 was much larger in magnitude than other parameters in (4.1), $\Pr(R^i_p = 1)$ was zero for an interval of τ^{ref}_p after each spike.) After the refractory period, we let the history-dependent term transiently increase the probability of a spike by setting

$$\bar{h}^j_{\text{hist},p} = a_{\text{hist},p} e^{-j\Delta t_{\text{sim}}/\tau_{\text{hist},p}} \qquad \text{for } j\Delta t_{\text{sim}} > \tau^{\text{ref}}_p.$$

As our purpose is to demonstrate the effect of history dependence, we included strong history dependence in each model neuron, setting $a_{\text{hist},p}$ relatively large and positive. Hence, the history-dependence term created a tendency for spikes to occur in bursts, leading to significant peaks in autocorrelation, such as shown in Figure 3.

We used the same spatiotemporal kernels $\bar{\mathbf{h}}_{\text{ext},p}$ as in [14], retaining the convention that $\bar{\mathbf{h}}^i_{\text{ext},p}$ was the kernel $\bar{\mathbf{h}}_{\text{ext},p}$ shifted for time point i. For line $j = 0, 1, \dots, N_0$ and temporal index t, we used the form

$$\bar{h}_{\text{ext},p}(j,t) = (t - b_p)\exp\left(-\frac{t - b_p}{\tau_{\text{ext},p}} - \frac{(j-c)^2}{2\sigma^2_p}\right)\cos(2\pi f_p(j - c) + \phi_p)$$

for $t > b_p$ and $\bar{h}_{\text{ext},p}(j,t) = 0$ otherwise [10]. To center the kernels on the image, we set $c = (N_0 - 1)/2$. The vector $\bar{\mathbf{h}}_{\text{ext},p}$ corresponded to $\bar{h}_{\text{ext},p}(j, k\Delta t_{\text{sim}})$ for integer k with $k\Delta t_{\text{sim}} < 200$ ms. We normalized $\bar{\mathbf{h}}_{\text{ext},p}$ so that the standard deviation of $\bar{\mathbf{h}}^i_{\text{ext},p} \cdot \mathbf{X}$ was equal to the parameter $a_{\text{ext},p}$; hence, $a_{\text{ext},p}$ specified how strongly neuron p responded to the stimulus.

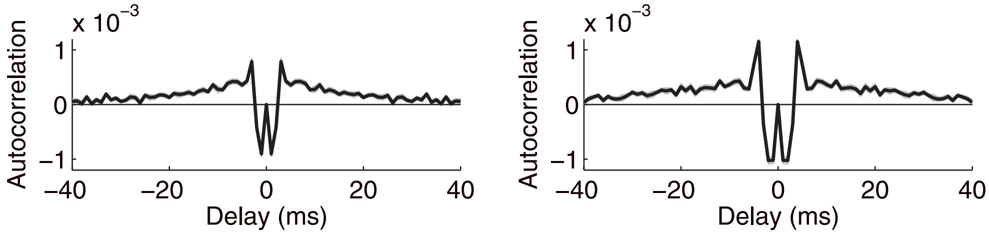

FIG. 3. *Examples of the large autocorrelations due strong history dependence included in simulated models. The autocorrelation of neuron p at delay j is $\langle R_p^i R_p^{i-j}\rangle - \langle\langle R_p^i|\mathbf{X}\rangle\langle R_p^{i-j}|\mathbf{X}\rangle\rangle$, where $\langle\cdot|\mathbf{X}\rangle$ indicates averaging over all repeats of the stimulus and $\langle\cdot\rangle$ indicates averaging over all time points. Shown are the autocorrelations from neuron* 1 *(left) and neuron* 2 *(left) in the simulation of Figure* 4(A). *Autocorrelation at zero delay has been truncated to zero.*

We used interneuronal coupling of the form

$$\bar{W}_{pq}^{j} = B_{pq}\frac{j\Delta t_{\mathrm{sim}} - d_{pq}}{\tau_w^2}\exp\left(-\frac{j\Delta t_{\mathrm{sim}} - d_{pq}}{\tau_W}\right)$$

for $j\Delta t_{\mathrm{sim}} > d_{pq}$ and $\bar{W}_{pq}^{j} = 0$ otherwise. Hence, d_{pq} represented the delay and B_{pq} the strength of the connection. For all connections, we set the time scale to $\tau_W = 0.5$ ms.

4.1.3. The model used in the analysis. We also used a generalized linear model (or linear-nonlinear model) for the analysis. (In [14], we test an earlier version of the analysis for stronger deviations from the simulated model.) In the analysis, we uses a temporal discretization of $\Delta t = 1$ ms.

We modeled the activity of each neuron in time bin i as a Poisson distribution (section 3.4) with the expected value given by

$$\lambda_s(\mathbf{r}_s^{<i}, \mathbf{x}, w; \theta_s^i) = A_s \log\left(1 + \exp\left[\sum_{j>0} h_{\mathrm{hist},s}^{j} r_s^{i-j} + \mathbf{h}_{\mathrm{ext},s}^{i}\cdot\mathbf{x} + w + y_s\right]\right). \tag{4.2}$$

The parameters θ_s^i correspond to A_s and y_s, as well as the parameters within $\mathbf{h}_{\mathrm{hist},s}$ and $\mathbf{h}_{\mathrm{ext},s}$. We used this form of the nonlinearity so that λ_s would be convex and $\log\lambda_s$ would be concave, a requirement for tractable numerical computations discussed in section 3.4.2 and [17]. We discuss how to determine the parameters θ_s^i in Appendix B.

4.2. Distinguishing common input from direct connection. We simulated two networks analogous to those schematized in Figure 1. In the first network, neuron 2 had a direct connection onto neuron 1. In the second network, a third, unmeasured neuron had a direct connection onto both neurons 1 and 2, with a longer delay onto neuron 1. In both cases, the spikes of neuron 1 were correlated with a delayed version of the spikes of neuron 2.

We simulated the response of each network to ten repetitions of the minute-long movie described above. Then we set the thresholds $\bar{y}_p$ so that each neuron spiked approximately 1,000 times during each presentation of the movie, obtaining approximately 10,000 spikes per neuron. The spikes from the third, common input neuron were discarded, as we treated that neuron as an unmeasured neuron.

Since we analyze just the spikes of two neurons, we will plot both the causal connection factor W and the common input factor U as a function of the delay j

defined as spike time of neuron 1 minus spike time of neuron 2. Hence, our plots will use the convention

$$W^j = \begin{cases} W_{12}^{-j} & \text{for } j < 0, \\ 0 & \text{for } j = 0, \\ W_{21}^{j} & \text{for } j > 0, \end{cases}$$

$$U^j = \begin{cases} U_{12}^{-j} & \text{for } j \leq 0, \\ U_{21}^{j} & \text{for } j > 0. \end{cases}$$

As shown in Figure 4, we were able to successfully distinguish the common input network from the direct connection network, despite the fact that the correlations between neurons 1 and 2 looked the same in both cases. The causal connection measure W was positive in the direct connection network; the common input measure U was positive in the common input network.

Section 3.4.1 outlines two differences between causal connections and common input that our analysis exploits to make this distinction. Only one of those differences was due exclusively to the history-dependence modeling that is the focus of this paper. To test the relative importance of the history-dependent factor, we reanalyzed the simulation of Figure 4 while ignoring any history-dependent effects. We set $\mathbf{h}_{\text{hist},p}$ in (4.2) to zero, essentially reverting our analysis back to an earlier version [14]. In this case, we model the expected activity of a node as independent of its measured history (conditioned on the external variables $\mathbf{X}$), so we remove the difference between the causal connection and common input terms of (3.17b) that is attributed to this history dependence.

The results after ignoring history-dependent effects (not shown) differed only slightly from the results when employing the full model. As in Figure 4, W was positive in the direct connection network, and U was positive in the common input network. Note that the simulations were generated with strong history dependence (yielding autocorrelations as in Figure 3). The fact that we achieved good results even while assuming no history dependence indicates that the analysis is at least somewhat robust to deviations from model assumptions.

4.3. Improvement from modeling history dependence.

Although the above results do indicate that the analysis that includes history dependence can succeed in distinguishing causal connections from common input, we wish to demonstrate that we have gained analytic power from our history-dependent modeling. Adding history-dependent effects to our modeling introduced significant complexity compared to an earlier version of the analysis [14]. To justify such complexity, we must demonstrate an improved ability to distinguish connectivity.

One limitation of earlier versions [13, 14] of this analysis is that they require that the neural activity be strongly related to measurable external variables (such as stimuli) in a manner that one can capture with a model. In many experimental contexts, such as when recording from brain areas that are not closely linked to a stimulus, such a strong relationship between external variables and neuronal activity may not be available. In such cases, the earlier versions of the analysis may not apply. On the other hand, if such neurons have a strong history dependence that can be captured by a model, the additional handle provided by history-dependent modeling may allow one to apply the analysis to these systems.

To demonstrate how the history-dependent modeling can improve the results, we repeated the simulation of Figure 4 but weakened the relationship between the

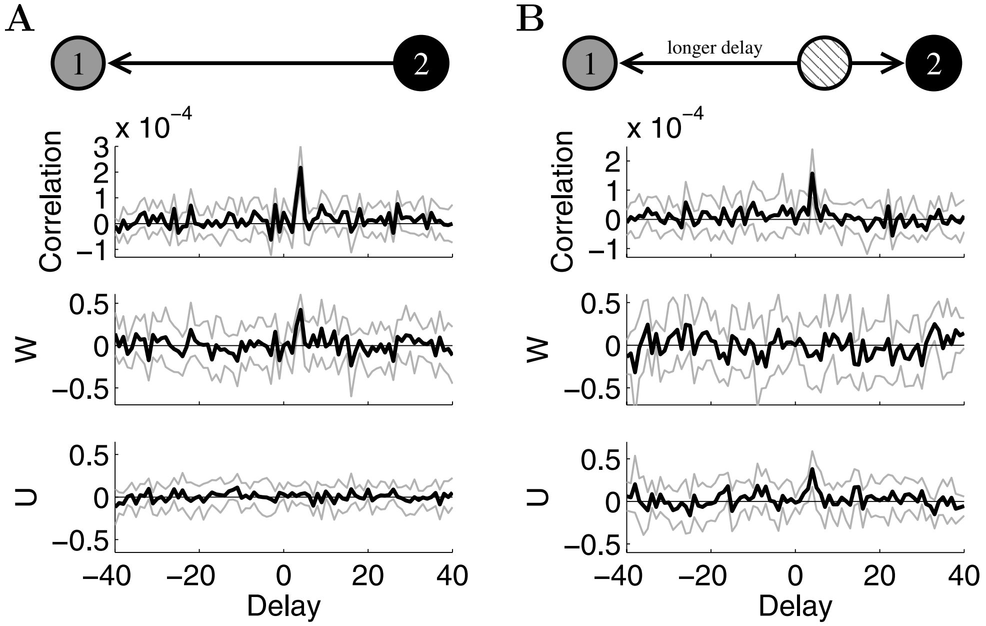

FIG. 4. *Successfully distinguishing a causal connection from common input.* (A) *Results from analyzing a network where neuron* 2 *has a direct connection onto neuron* 1*, as schematized at top. The correlation (shuffle-corrected correlogram or covariogram* [19, 1, 16]*) at delay j is* $\langle R_1^i R_2^{i-j}\rangle - \langle\langle R_1^i|\mathbf{X}\rangle\langle R_2^{i-j}|\mathbf{X}\rangle\rangle$*, where the averaging* $\langle\cdot\rangle$ *is defined as in Figure* 3*. The direct connection leads to a peak in the correlation at a positive delay. The causal connection measure W (but not the common input measure U) has a positive peak at the same delay, indicating the presence of a causal connection from neuron* 2 *onto neuron* 1*. (At the peak, W was seven standard errors from zero.) Thin gray lines indicate a bootstrap estimate of three standard errors, calculated by resampling from the set of stimulus repetitions* 50 *times. Simulation parameters:* $a_{\text{hist},1} = 1.2$, $a_{\text{hist},2} = 1.5$, $\tau_{\text{hist},1} = 10$ ms, $\tau_{\text{hist},2} = 12$ ms, $\bar{y}_1 = 0.5$, $\bar{y}_2 = 0.7$, $b_1 = b_2 = 0$, $a_{\text{ext},1} = a_{\text{ext},2} = 1$, $\tau_{\text{ext},1} = 40$ ms, $\tau_{\text{ext},2} = 50$ ms, $\sigma_1 = 10$, $\sigma_2 = 15$, $f_1 = 0.08$, $f_2 = 0.04$, $\phi_1 = 0$, $\phi_2 = 2\pi/3$, $B_{21} = 1.2$, $d_{21} = 3$, $B_{11} = B_{12} = B_{22} = 0$. (B) *Results from analyzing a network where an unmeasured neuron (hatched circle in schematic at top) has a connection onto neuron* 1 *and onto neuron* 2*. Since the connection onto neuron* 1 *has a longer delay, there is a peak in the correlation at a positive delay that is indistinguishable from a peak in correlation due to a direct connection from neuron* 2 *onto neuron* 1*. Only U, and not W, has a positive peak at the same delay, indicating that the correlation was due to common input rather than any causal connection from neuron* 2 *onto neuron* 1*. (At the peak, U was five standard errors from zero.) Most parameters as in panel A. Exceptions and additional parameters (the unmeasured neuron is indexed by* 3*):* $a_{\text{hist},3} = 1.0$, $\tau_{\text{hist},3} = 6$ ms, $\bar{y}_2 = 0.6$, $\bar{y}_3 = 0.8$, $b_3 = 0$, $a_{\text{ext},3} = 1$, $\tau_{\text{ext},3} = 45$ ms, $\sigma_3 = 20$, $f_3 = 0.06$, $\phi_3 = 4\pi/3$, $d_{31} = 4$, $d_{32} = 0$, $B_{31} = B_{32} = 4.5$, $B_{ij} = 0$ *for all other i and j.*

neuronal activity and the stimulus. We reduced the magnitude of the external variable terms $\mathbf{h}_{\text{ext},p} \cdot \mathbf{X}$ by a factor of 5 (reducing their standard deviation $a_{\text{ext},p}$ from 1 to 0.2). As this greatly increased the difficulty of the network analysis, we also doubled the simulation length to 20 simulated minutes (20 repeats of the movie), obtaining around 20,000 spikes from each neuron.

The results of the analysis based on the full model (4.2) are shown in Figure 5. Despite the weak dependence on the stimulus, the analysis was still able to determine which network contained the causal connection and which network contained common input.

In this case, since the neurons' activities were only weakly related to the stimulus, the history-dependent effects played a bigger role in determining the connectivity. To

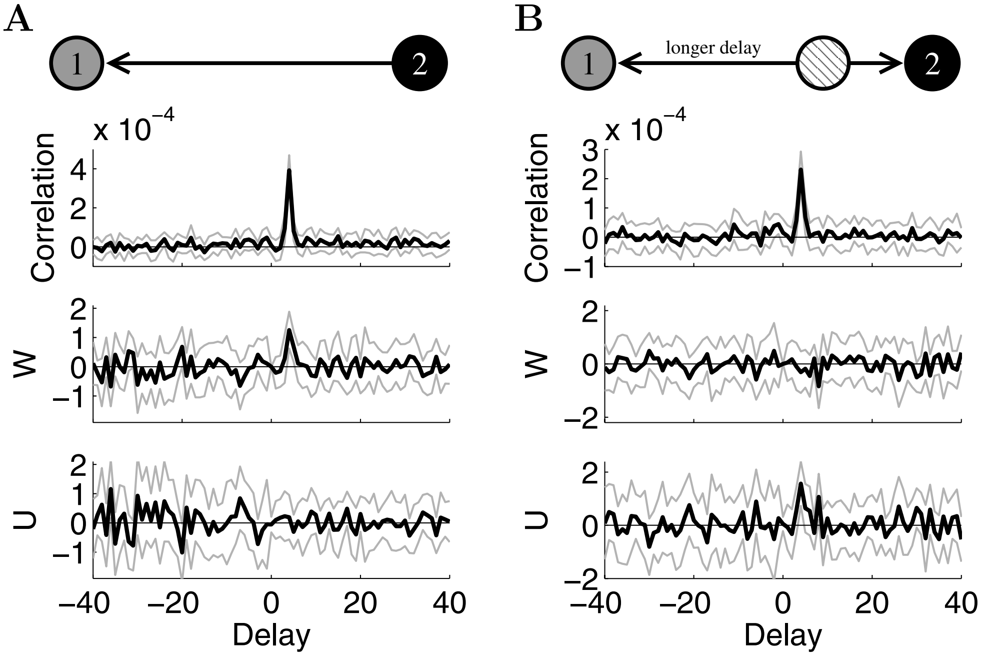

FIG. 5. *Determining circuitry even when neuronal activity is only weakly related to the stimulus. The same networks as in Figure* 4 *were simulated, except that the magnitude of the stimulus input was decreased by a factor of* 5 *and the simulation length was doubled. The causal connection network was still distinguished from the common input network, primarily due to exploitation history-dependent effects (cf. Figure* 6*, where these effects were ignored). Panels as in Figure* 4*.* (A) *The causal connection measure W has a peak at the same delay as the correlation, indicating the correlation was due to a causal connection from neuron* 2 *onto neuron* 1*. (At the peak, W was six standard errors from zero.) Parameters as in Figure* 4(A)*, except that $a_{\text{ext},1} = a_{\text{ext},2} = 0.2$, $\bar{y}_1 = 1.1$, and $\bar{y}_2 = 1.1$.* (B) *The common input measure U has a peak at the same delay as the correlation, indicating the correlation was due to common input. (At the peak, U was five standard errors from zero.) Parameters as in Figure* 4(B)*, except that $a_{\text{ext},1} = a_{\text{ext},2} = a_{\text{ext},3} = 0.2$, $\bar{y}_1 = 1.0$, $\bar{y}_2 = 1.0$, $\bar{y}_3 = 1.2$, and $B_{31} = B_{32} = 4$.*

demonstrate the role of the history-dependent model, we reanalyzed the simulation results of Figure 5 while ignoring history-dependent effects (as above, we set $\mathbf{h}_{\text{hist},p}$ in (4.2) to zero). This time, the analysis was unable to make a clear distinction between the direct connection network and the common input network, as shown in Figure 6. In the direct connection network of Figure 6(A), both W and U were positive so that the result was ambiguous. In the common input network of Figure 6(B), only U was positive at the delay corresponding to the correlation, but U was barely above the noise, and the result was much weaker than in Figure 5(B). (If we quadrupled the simulation to 80 simulated minutes, then the network analysis was able to determine the connectivity even with ignoring history-dependent effects.)

4.4. Subpopulation ambiguity. In section 3.3.3, we described an assumption we made about the hidden nodes in order to complete our analysis. We briefly mentioned that this assumption resulted in a certain degree of ambiguity in the identity of causal connections. This ambiguity is described in detail in [14], where we refer to it as *subpopulation ambiguity.*

The nature of the subpopulation ambiguity is illustrated by Figure 7. Here we repeated the simulation of the common input network of Figure 4(B), except we

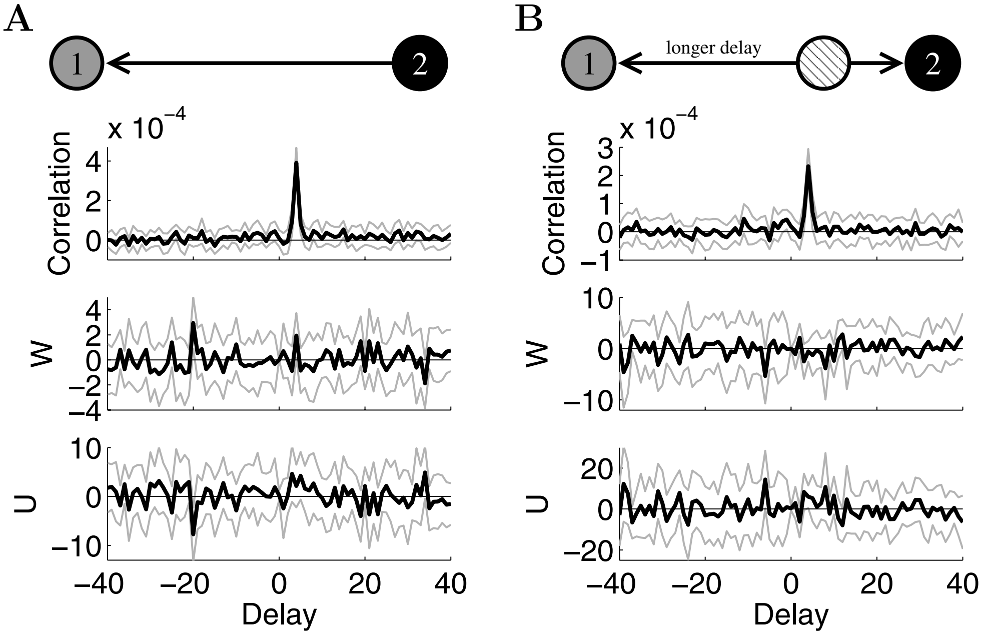

Fig. 6. *Reanalyzing the simulations of Figure* 5 *while ignoring all history-dependent effects. The history kernel* $\mathbf{h}_{\text{hist}}$ *of* (4.2) *was set to zero, so the analysis could not exploit the differences between causal connection and common input networks that are caused by history dependence. In this case, since the neural activity was only weakly related to the stimulus, the analysis failed to cleanly distinguish the circuitry. Panels as in Figure* 4. (A) *The causal connection measure* W *did have a (small) positive peak at the delay of the peak in the correlation. However, the common input measure* U *also had a positive peak at that delay, so that the identity of the causal connection could not be clearly determined. (At the peak* W *was three standard errors and* U *was over two standard errors from zero.)* (B) *Only the causal connection measure* U *had a peak at the delay of the correlation peak, so the results do correctly point to the presence of common input. However, the peak in* U *at that delay is small (though it was three standard errors from zero), especially compared to Figure* 5(B)*, indicating that ignoring history dependence hampered the ability to determine circuitry.*

changed the kernel $\bar{\mathbf{h}}_{\text{ext},3}$ of the unmeasured neuron first to match the kernel $\bar{\mathbf{h}}_{\text{ext},2}$ of neuron 2 and then to match the kernel $\bar{\mathbf{h}}_{\text{ext},1}$ of neuron 1. As shown in Figure 7(A), the analysis misidentifies the common input as a causal connection when the kernel of the unmeasured common input neuron matched neuron 2. The analysis does not have any trouble correctly identifying the common input when the kernel of the unmeasured common input neuron matched neuron 1, as shown in Figure 7(B).

We argue that the misidentification in Figure 7(A) merely introduces a relatively modest ambiguity into the interpretation of the results. Clearly, one cannot justify a strict interpretation that the peak in W always indicates a causal connection from neuron 2 itself onto neuron 1. However, note that in the network of Figure 7(A) there is a causal connection from the unmeasured neuron onto neuron 1 and that this unmeasured neuron has similar properties to neuron 2 (one might use the language that the unmeasured neuron has a *receptive field* that is similar to that of neuron 2). Hence, one can make a looser interpretation of the peak in W to indicate the presence of a causal connection onto neuron 1 from some neuron with properties (or receptive field) similar to neuron 2.

In experiments where one measures only the spike times of individual neurons, neurons are identified only by their properties, such as the relationship between their

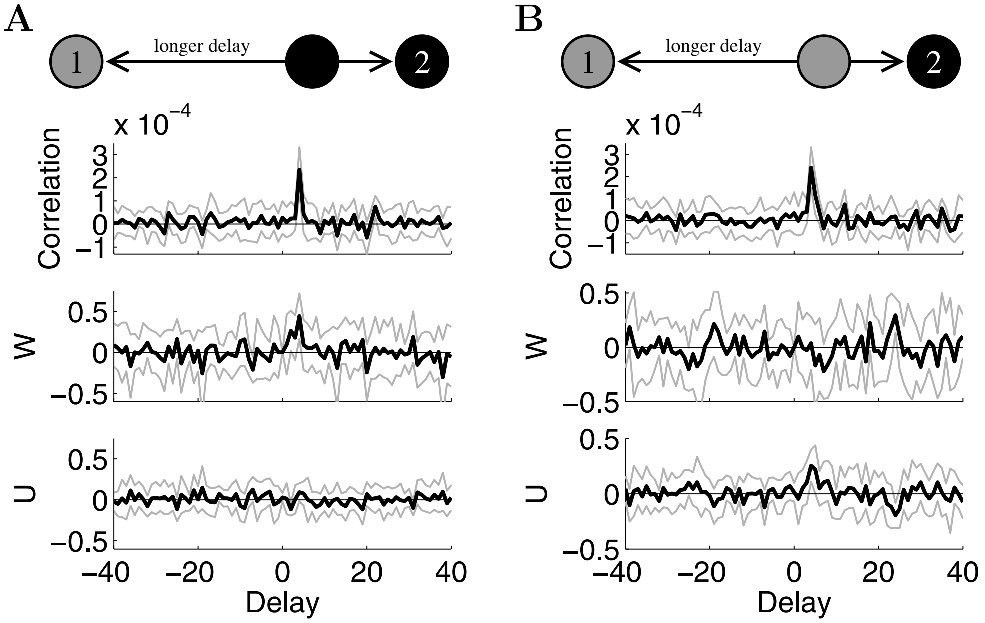

Fig. 7. *An illustration of the subpopulation ambiguity in the identification of the individual neurons involved in a connection. Note that, in both networks shown, the delays are set up so that the correlations mimic a connection from neuron* 2 *onto neuron* 1*. Hence neuron* 1 *and neuron* 2 *do not play symmetric roles. Panels as in Figure* 4. (A) *When the unmeasured neuron has similar properties as neuron* 2 *(as schematized by the black circles at top), the causal connection factor W has a peak at the delay of the correlation peak, incorrectly indicating a connection from neuron* 2 *onto neuron* 1*. (At the peak, W is four standard errors from zero.) However, there is a connection onto neuron* 1 *from a neuron similar to neuron* 2 *(a black neuron in the schematic). Hence, W must be interpreted as indicating a causal connection onto neuron* 1 *from a neuron with properties similar to those of neuron* 2*. Parameters as in Figure* 4(B)*, except that $\bar{y}_1 = 0.4$, $\bar{y}_2 = 0.5$, $\bar{y}_3 = 0.9$, $b_2 = 1$ ms, $\tau_{\text{ext},3} = 50$ ms, $\sigma_3 = 16$, $f_3 = 0.038$, and $\phi_3 = 2\pi/3$.* (B). *When the unmeasured neuron has properties similar to neuron* 1 *(as schematized by the gray circles at top), the results correctly indicate a common input connection as U has a positive peak at the delay of the correlation peak. (At the peak, U is five standard errors from zero.) In this case, it is important that the analysis obtained the correct results, as there is no connection from a neuron similar to neuron* 2 *(a black neuron in the schematic) onto a neuron similar to neuron* 1 *(a gray neuron). Parameters as in Figure* 4(B)*, except that $\bar{y}_1 = 0.4$, $\bar{y}_2 = 0.5$, $\bar{y}_3 = 0.7$, $b_1 = 5$ ms, $\tau_{\text{ext},3} = 40$ ms, $\sigma_3 = 11$, $f_3 = 0.078$, and $\phi_3 = 0$.*

spikes and external variables or stimuli. In this case, if two neurons had similar properties (such as the unmeasured neuron and neuron 2 in Figure 7(A)), those two neurons would be indistinguishable. Hence, it would not make a difference if one concluded that neuron 2 had a connection onto neuron 1 or concluded that an unmeasured neuron with similar properties had a connection onto neuron 1. In either case, the conclusion would be that a neuron with the properties of neuron 2 had a connection onto a neuron with the properties of neuron 1.

In Figure 7(B), there is no connection from a neuron with properties similar to neuron 2 onto a neuron with properties similar to neuron 1. Even with the looser interpretation of the causal connection W, this network cannot be identified as having a causal connection. It is critical that the analysis correctly identified the correlation as arising from common input.

In [14], we refer to a group of neurons with similar properties as a *subpopulation* of neurons. Since the identity of the presynaptic neuron involved in a connection is narrowed down only to an individual member within a supopulation, we use the language that our analysis can determine connectivity only with subpopulation ambiguity. In using such a term, one must be careful to recognize that one is not assuming connections between groups of neurons but only ambiguity in the identity of individual neurons. See [14] for more details, including more intuition behind the subpopulation ambiguity.

5. Discussion.

The present work represents a continuation of our development of methods to determine the pattern of causal connections among measured neurons while controlling for the effects of unmeasured neurons [14, 13, 12]. We have successfully eliminated the limitation of earlier versions that the activity of a neuron could depend only weakly on its history. In the process, we have discovered that one can exploit such history dependence to increase one's ability to distinguish common input from causal connections.

Although the analysis involved a fair number of technical manipulations, it turns out that the intuition developed in the introduction does hold for the class of models we consider. In the common input configuration (Figure 2(B)), but not in the causal connection configuration (Figure 2(A)), spikes that can be accounted for by the first neuron's history dependence do not influence the second neuron's spiking probability (see (3.17)). This difference is exploited by our analysis in order to distinguish common input from causal connections.

Successfully exploiting history dependence requires a strong dependence on history in a manner that one can capture by a model. In our simulations, we included such history dependence and demonstrated that we could use it to improve our estimates of connectivity. It is well known that the spike times of neurons are not well approximated by a Poisson process [23, 21] and hence contain history dependence. However, it remains unclear if this history dependence is sufficiently strong and if it can be sufficiently well modeled to aid in the determination of connectivity.

The analysis was justified by a weak coupling assumption (section 2.2) where the original coupling $\bar{W}$ was assumed to be a small parameter. However, even if the original coupling $\bar{W}$ were large and only the perturbation $\widetilde{W}$ due to coupling (see (3.15b)) were small, the analysis might still indicate the effective connectivity of the network. To interpret the analysis under these conditions, one could reinterpret the likelihood equation (3.15) as a perturbation off the effective models (2.3) rather than off the original network (2.2). In this case, one cannot assume that the causal connectivity obtained with W actually corresponds to the underlying connectivity of the network. Such a reinterpretation of W as an effective connectivity would allow application of the results to networks where the weak coupling assumption cannot be justified.

Although the analysis depends on selecting appropriate single-neuron models of the form (2.3), flexibility is given by the modular approach [14] employed in our analysis. One can develop additional single-neuron models and include them in the analysis without modification of the network analysis. In the simulation tests, we used only generalized linear models. Such a model of the dependence of neural activity on spiking history is, of course, only roughly approximated by such a model. One future goal is to implement more sophisticated models of history dependence, such as a stochastic integrate-and-fire model [18]. Paninski, Pillow, and Simoncelli have already developed efficient numerical schemes for determining the parameters of the stochastic integrate-and-fire model [18], and the model does fit into the formalism of (2.3). Such a model may more closely approximate history dependence observed in biological neurons.

Although there is a large literature focused on analyzing interactions among neurons [19, 1, 16, 2, 3, 11, 25, 15, 7, 20, 18, 24, 9], we are aware of only one other attempt to explicitly control for the effects of common input from *unmeasured* sources. Kulkarni and Paninski [9] have recently developed an expectation-maximization algorithm for fitting a neuronal model that contains a latent noise source that could correspond to such unmeasured common input. In their approach, the common input (i.e., latent noise) is assumed to be a Gaussian process (justified by thinking of the common input as a sum of a large number of small inputs). Hence, in place of a point process model (2.2) for a network containing unmeasured neurons, their model is a doubly stochastic process or Cox process [22]. As their approach differs significantly from ours, one future task will be to compare the results of the two methods to understand their relative strengths and weaknesses.

Earlier versions of our analysis relied exclusively on models of the relationship between neuron spikes and external variables such as stimuli. As we have demonstrated via simulations, modeling history-dependent effects may allow one to apply the analysis even in cases where the activity of neurons is not strongly related to external variables. Especially with the implementation of more sophisticated models of history dependence, our analysis may become applicable to a large variety of neuronal systems (or other networks), regardless of whether or not they are strongly linked to a stimulus or other external variable.

Appendix A. Calculations underlying analysis.

A.1. Averaging over hidden node activity. We outline how to simplify (3.4) for the probability distribution $\Pr(\mathbf{R}_{\mathcal{Q}}|\mathbf{X})$ of measured node activity $\mathbf{R}_{\mathcal{Q}}$ by explicitly computing the averages over hidden node activity $\mathbf{R}_{\mathcal{P}}$. We argued in the context of (3.4) that the value r_s^i of any random variable appears in (3.4) only as a polynomial in r_s^i times $\bar{P}_s$ (or times a derivative of $\bar{P}_s$).

Expressions involving the undifferentiated $\bar{P}_s$ are simple. Let a sum over $\mathbf{r}_s$ denote the sum over all possible values of the activity $\mathbf{r}_s$ (i.e., r_s^i for all i) of a given node s. Then, since $\bar{P}_s$ is shorthand for a probability distribution in the $\mathbf{r}_s$, we can conclude that

$$\sum_{\mathbf{r}_s} \bar{P}_s = 1, \qquad \sum_{\mathbf{r}_s} r_s^i \bar{P}_s = \bar{E}_0(R_s^i), \qquad \text{and} \qquad \sum_{\mathbf{r}_s} r_s^{i_1} r_s^{i_2} \bar{P}_s = \bar{E}_0(R_s^{i_1} R_s^{i_2}). \tag{A.1}$$

$\bar{E}_0(\cdot)$ denotes the expected value under the probability distribution defined by the P_s with W arguments set to zero, i.e., for any function g of the activity of nodes,

$$\bar{E}_0(g(\mathbf{R})) = \sum_{\mathbf{r}} g(\mathbf{r}) \prod_s \prod_i P_s\big(r_s^i, \mathbf{r}_s^{<i}, \mathbf{x}, 0; \bar{\theta}_s^i\big). \tag{A.2}$$

(The sum over $\mathbf{r}$ indicates the sum over all possible values of the activity of all nodes.) Note that $\bar{E}_0(R_s^i)$ is not the expected value of R_s^i under model (2.2); it is the expected value of R_s^i only if the coupling happened to be zero.

The expressions involving the derivatives of $\bar{P}_s$ are more subtle. First, note that, for any node s,

$$\sum_{\mathbf{r}_s} \prod_i P_s \left(r_s^i, \mathbf{r}_s^{<i}, \mathbf{x}, \sum_{\acute{s} \neq s} \sum_{\acute{\imath} < i} \bar{W}_{\acute{s},s}^{\acute{\imath},i} r_{\acute{s}}^{\acute{\imath}}; \bar{\theta}_s^i \right) = 1$$

independent of any value of the $r^{\tilde{\imath}}_{\tilde{s}}$ for $\tilde{s} \neq s$. (The case when all $r^{\tilde{\imath}}_{\tilde{s}} = 0$ for $\tilde{s} \neq s$ was the first identity of (A.1).) So, if we differentiate with respect to any $r^{\tilde{\imath}}_{\tilde{s}}$ with $\tilde{s} \neq s$, we will get zero:

$$\sum_{\mathbf{r}_s} \frac{\partial}{\partial r^{\tilde{\imath}}_{\tilde{s}}} \left(\prod_i P_s \left(r^i_s, \mathbf{r}^{<i}_s, \mathbf{x}, \sum_{\acute{s} \neq s} \sum_{\acute{\imath} < i} \bar{W}^{\acute{\imath},i}_{\acute{s},s} r^{\acute{\imath}}_{\acute{s}}; \bar{\theta}^i_s \right) \right) = 0.$$

In particular, this derivative is zero if we set all $r^{\tilde{\imath}}_{\tilde{s}} = 0$ for $\tilde{s} \neq s$, so that

$$\text{(A.3)} \qquad \sum_{\mathbf{r}_s} \frac{\partial \bar{P}_s}{\partial r^{\tilde{\imath}}_{\tilde{s}}} = 0, \qquad \text{and} \qquad \sum_{\mathbf{r}_s} \frac{\partial^2 \bar{P}_s}{\partial r^{\tilde{\imath}_1}_{\tilde{s}_1} \partial r^{\tilde{\imath}_2}_{\tilde{s}_1}} = 0.$$

On the other hand,

$$E(R^i_s | \{\mathbf{R}^{<i}_{\acute{s}} = \mathbf{r}^{<i}_{\acute{s}}\}_{\acute{s} \neq s}) = \sum_{\mathbf{r}_s} r^i_s \prod_{\acute{\imath}_1} P_s \left(r^{\acute{\imath}_1}_s, \mathbf{r}^{<\acute{\imath}_1}_s, \mathbf{x}, \sum_{\acute{s} \neq s} \sum_{\acute{\imath}_2 < \acute{\imath}_1} \bar{W}^{\acute{\imath}_2,\acute{\imath}_1}_{\acute{s},s} r^{\acute{\imath}_2}_{\acute{s}}; \bar{\theta}^{\acute{\imath}_1}_s \right)$$

does depend on the values of the $r^{\tilde{\imath}}_{\tilde{s}}$ for $\tilde{s} \neq s$ and $\tilde{\imath} < i$. Due to network connections, the expected value of R^i_s could indeed depend on the value of the past activity of another node. So, if we differentiate with respect to any $r^{\tilde{\imath}}_{\tilde{s}}$, we won't necessarily get zero. Denote this derivative, once we set all $r^{\tilde{\imath}}_{\tilde{s}} = 0$ for $\tilde{s} \neq s$, as

$$\text{(A.4)} \qquad \bar{E}_0 \left(\frac{\partial R^i_s}{\partial R^{\tilde{\imath}}_{\tilde{s}}} \right) = \frac{\partial}{\partial r^{\tilde{\imath}}_{\tilde{s}}} E(R^i_s | \{\mathbf{R}^{<i}_{\acute{s}} = \mathbf{r}^{<i}_{\acute{s}}\}_{\acute{s} \neq s}) \bigg|_{\{\mathbf{r}_{\acute{s}} = \mathbf{0} | \acute{s} \neq s\}} = \sum_{\mathbf{r}_s} r^i_s \frac{\partial \bar{P}_s}{\partial r^{\tilde{\imath}}_{\tilde{s}}}.$$

The notation captures that this expression represents how a change in the activity of node $\tilde{s}$ at time $\tilde{\imath}$ affects the average activity of node s at time i. This is nonzero, of course, only if $\tilde{\imath} < i$. Note that this expression doesn't depend on $R^{\tilde{\imath}}_{\tilde{s}}$, as the derivative is calculated around $R^{\tilde{\imath}}_{\tilde{s}} = 0$. Note also that this expression need not be zero even if $R^{\tilde{\imath}}_{\tilde{s}}$ does not directly influence R^i_s, i.e., if $\bar{W}^{\tilde{\imath},i}_{\tilde{s},s} = 0$. Because we have allowed R^i_s to depend arbitrarily on its history $R^{i_2}_s$ for $i_2 < i$, this derivative could be nonzero just because $\bar{W}^{\tilde{\imath},i_2}_{\tilde{s},s} \neq 0$.

We use the identities in (A.1), (A.3), and (A.4) to simplify all of the sums over $\mathbf{r}_{\mathcal{P}}$ in the marginal distribution of $\mathbf{R}_{\mathcal{Q}}$ given in (3.4). To use these identities, we need to distinguish all of the subsets of the various s indicies that could correspond to a hidden node. We do this by enumerating all of the possible ways in which each s index could be either a hidden or a measured node, as well as all of the possible ways in which hidden node indices in a given term could correspond to the same node. Hence each term in (3.4) will be expanded into many different terms.

However, due to the identities in (A.3), most terms involving derivatives of hidden nodes disappear. Recall that a derivative of $\bar{P}_s$ represents a connection onto node s. A connection onto a hidden node p should not directly affect the marginal distribution of the measured nodes $R_{\mathcal{Q}}$; such a connection should have an effect only through a connection from that hidden node onto a measured node. Indeed, the only place where derivatives of hidden nodes survive is in the last term:

$$\frac{1}{2} \sum_{\mathbf{r}_{\mathcal{P}}} \sum_{\substack{s_1, s_2, \tilde{s}_1, \tilde{s}_2 \\ s_2 \neq s_1, \tilde{s}_1 \neq s_1 \\ \tilde{s}_2 \neq s_2}} \sum_{\tilde{\imath}_1, \tilde{\imath}_2} \frac{\partial \bar{P}_{s_1}}{\partial r^{\tilde{\imath}_1}_{\tilde{s}_1}} \frac{\partial \bar{P}_{s_2}}{\partial r^{\tilde{\imath}_2}_{\tilde{s}_2}} r^{\tilde{\imath}_1}_{\tilde{s}_1} r^{\tilde{\imath}_2}_{\tilde{s}_2} \prod_{\substack{s_3 \\ s_3 \neq s_1 \\ s_3 \neq s_2}} \bar{P}_{s_3}.$$

If we set s_1 to a hidden node $s_1 = p_1$, then by identity (A.4), the term will survive only if $\tilde{s}_2$ also corresponds to the same hidden node $\tilde{s}_2 = p_1$ and if $\tilde{\imath}_2$ corresponds to an earlier time $\tilde{\imath}_2 < \tilde{\imath}_1$. If we then tried to set s_2 to another hidden node $s_2 = p_2$, we would need the contradictory condition of $\tilde{\imath}_2 > \tilde{\imath}_1$ for the term to survive. In this case, we must set s_2 to a measured node $s_2 = q_2$. Hence, with these substitutions, the term represents the cascade of the effect of a connection from node $\tilde{s}_2$ (which could be hidden or measured) onto hidden node p_1 combined with the effect of a connection from hidden node p_1 onto measured node q_2. (We must double the effect of this term because we could swap s_1 and s_2 and obtain the same result.)

In all other cases, only the effects of connections onto measured nodes survive. The connections from measured and hidden nodes must still be distinguished. We describe this process in [14] as identifying all possible subnetworks of two or fewer edges. When this process is completed, we end up with the following lengthy expression:

$$
\begin{aligned}
\Pr(\mathbf{R}_{\mathcal{Q}} = \mathbf{r}_{\mathcal{Q}} | \mathbf{X} = \mathbf{x}) = {} & \prod_{q} \bar{P}_q + \sum_{\substack{q_1, \tilde{q}_1 \\ q_1 \neq \tilde{q}_1}} \sum_{\tilde{\imath}_1} \frac{\partial \bar{P}_{q_1}}{\partial r_{\tilde{q}_1}^{\tilde{\imath}_1}} r_{\tilde{q}_1}^{\tilde{\imath}_1} \prod_{\substack{q_2 \\ q_2 \neq q_1}} \bar{P}_{q_2} \\
& + \sum_{q_1, \tilde{p}_1} \sum_{\tilde{\imath}_1} \frac{\partial \bar{P}_{q_1}}{\partial r_{\tilde{p}_1}^{\tilde{\imath}_1}} \bar{E}_0(R_{\tilde{p}_1}^{\tilde{\imath}_1}) \prod_{\substack{q_2 \\ q_2 \neq q_1}} \bar{P}_{q_2} \\
& + \frac{1}{2} \sum_{\substack{q_1, \tilde{q}_1, \tilde{q}_2 \\ q_1 \neq \tilde{q}_1, q_1 \neq \tilde{q}_2}} \sum_{\tilde{\imath}_1, \tilde{\imath}_2} \frac{\partial^2 \bar{P}_{q_1}}{\partial r_{\tilde{q}_1}^{\tilde{\imath}_1} \partial r_{\tilde{q}_2}^{\tilde{\imath}_2}} r_{\tilde{q}_1}^{\tilde{\imath}_1} r_{\tilde{q}_2}^{\tilde{\imath}_2} \prod_{\substack{q_2 \\ q_2 \neq q_1}} \bar{P}_{q_2} \\
& + \sum_{\substack{q_1, \tilde{q}_1, \tilde{p}_2 \\ q_1 \neq \tilde{q}_1}} \sum_{\tilde{\imath}_1, \tilde{\imath}_2} \frac{\partial^2 \bar{P}_{q_1}}{\partial r_{\tilde{q}_1}^{\tilde{\imath}_1} \partial r_{\tilde{p}_2}^{\tilde{\imath}_2}} r_{\tilde{q}_1}^{\tilde{\imath}_1} \bar{E}_0(R_{\tilde{p}_2}^{\tilde{\imath}_2}) \prod_{\substack{q_2 \\ q_2 \neq q_1}} \bar{P}_{q_2} \\
& + \frac{1}{2} \sum_{q_1, \tilde{p}_1, \tilde{p}_2} \sum_{\tilde{\imath}_1, \tilde{\imath}_2} \frac{\partial^2 \bar{P}_{q_1}}{\partial r_{\tilde{p}_1}^{\tilde{\imath}_1} \partial r_{\tilde{p}_2}^{\tilde{\imath}_2}} \bar{E}_0(R_{\tilde{p}_1}^{\tilde{\imath}_1} R_{\tilde{p}_2}^{\tilde{\imath}_2}) \prod_{\substack{q_2 \\ q_2 \neq q_1}} \bar{P}_{q_2} \\
& + \sum_{\substack{q_1, \tilde{p}_1, \tilde{p}_2 \\ \tilde{p}_1 \neq \tilde{p}_2}} \sum_{\substack{\tilde{\imath}_1, \tilde{\imath}_2 \\ \tilde{\imath}_2 < \tilde{\imath}_1}} \frac{\partial \bar{P}_{q_1}}{\partial r_{\tilde{p}_1}^{\tilde{\imath}_1}} \bar{E}_0 \left(\frac{\partial R_{\tilde{p}_1}^{\tilde{\imath}_1}}{\partial R_{\tilde{p}_2}^{\tilde{\imath}_2}} \right) \bar{E}_0(R_{\tilde{p}_2}^{\tilde{\imath}_2}) \prod_{\substack{q_2 \\ q_2 \neq q_1}} \bar{P}_{q_2} \\
& + \sum_{q_1, \tilde{p}_1, \tilde{q}_2} \sum_{\substack{\tilde{\imath}_1, \tilde{\imath}_2 \\ \tilde{\imath}_2 < \tilde{\imath}_1}} \frac{\partial \bar{P}_{q_1}}{\partial r_{\tilde{p}_1}^{\tilde{\imath}_1}} \bar{E}_0 \left(\frac{\partial R_{\tilde{p}_1}^{\tilde{\imath}_1}}{\partial R_{\tilde{q}_2}^{\tilde{\imath}_2}} \right) r_{\tilde{q}_2}^{\tilde{\imath}_2} \prod_{\substack{q_2 \\ q_2 \neq q_1}} \bar{P}_{q_2} \\
& + \frac{1}{2} \sum_{\substack{q_1, \tilde{p}_1, q_2, \tilde{p}_2 \\ q_2 \neq q_1}} \sum_{\tilde{\imath}_1, \tilde{\imath}_2} \frac{\partial \bar{P}_{q_1}}{\partial r_{\tilde{p}_1}^{\tilde{\imath}_1}} \frac{\partial \bar{P}_{q_2}}{\partial r_{\tilde{p}_2}^{\tilde{\imath}_2}} \bar{E}_0(R_{\tilde{p}_1}^{\tilde{\imath}_1} R_{\tilde{p}_2}^{\tilde{\imath}_2}) \prod_{\substack{q_3 \\ q_3 \neq q_1, q_3 \neq q_2}} \bar{P}_{q_3} \\
& + \sum_{\substack{q_1, \tilde{q}_1, q_2, \tilde{p}_2 \\ q_2 \neq q_1, q_1 \neq \tilde{q}_1}} \sum_{\tilde{\imath}_1, \tilde{\imath}_2} \frac{\partial \bar{P}_{q_1}}{\partial r_{\tilde{q}_1}^{\tilde{\imath}_1}} \frac{\partial \bar{P}_{q_2}}{\partial r_{\tilde{p}_2}^{\tilde{\imath}_2}} r_{\tilde{q}_1}^{\tilde{\imath}_1} \bar{E}_0(R_{\tilde{p}_2}^{\tilde{\imath}_2}) \prod_{\substack{q_3 \\ q_3 \neq q_1, q_3 \neq q_2}} \bar{P}_{q_3} \\
& + \frac{1}{2} \sum_{\substack{q_1, \tilde{q}_1, q_2, \tilde{q}_2 \\ q_2 \neq q_1, q_1 \neq \tilde{q}_1 \\ q_2 \neq \tilde{q}_2}} \sum_{\tilde{\imath}_1, \tilde{\imath}_2} \frac{\partial \bar{P}_{q_1}}{\partial r_{\tilde{q}_1}^{\tilde{\imath}_1}} \frac{\partial \bar{P}_{q_2}}{\partial r_{\tilde{q}_2}^{\tilde{\imath}_2}} r_{\tilde{q}_1}^{\tilde{\imath}_1} r_{\tilde{q}_2}^{\tilde{\imath}_2} \prod_{\substack{q_3 \\ q_3 \neq q_1, q_3 \neq q_2}} \bar{P}_{q_3} + O(\bar{W}^3).
\end{aligned}
\tag{A.5}
$$

A.2. Obtaining an effective parameter equation. We obtained (A.5) for $\Pr(\mathbf{R}_{\mathcal{Q}} | \mathbf{X})$ by averaging the full model (2.2) over the activity of all hidden nodes. The averaged model (2.3) is equivalent to the full model (2.2) averaged over the activity of all nodes except for a single node s. Hence, the averaged model (2.3) must be equal to (A.5) for $\Pr(\mathbf{R}_{\mathcal{Q}}|\mathbf{X})$ where the set $\mathcal{Q}$ of measured nodes is replaced by the single node s.

Given definition (3.5) for the effective probability distribution P_s, we obtain

$$\begin{aligned}
P_s &= \Pr(\mathbf{R}_s = \mathbf{r}_s|\mathbf{X} = \mathbf{x})\\
&= \bar{P}_s + \sum_{\substack{\tilde{s}_1\\ \tilde{s}_1\neq s}} \sum_{\tilde{\imath}_1} \frac{\partial \bar{P}_s}{\partial r^{\tilde{\imath}_1}_{\tilde{s}_1}} \bar{E}_0(R^{\tilde{\imath}_1}_{\tilde{s}_1})\\
&\quad + \frac{1}{2} \sum_{\substack{\tilde{s}_1,\tilde{s}_2\\ \tilde{s}_1\neq s,\tilde{s}_2\neq s}} \sum_{\tilde{\imath}_1,\tilde{\imath}_2} \frac{\partial^2 \bar{P}_s}{\partial r^{\tilde{\imath}_1}_{\tilde{s}_1} \partial r^{\tilde{\imath}_2}_{\tilde{s}_2}} \bar{E}_0(R^{\tilde{\imath}_1}_{\tilde{s}_1} R^{\tilde{\imath}_2}_{\tilde{s}_2})\\
&\quad + \sum_{\substack{\tilde{s}_1,\tilde{s}_2\\ \tilde{s}_1\neq s,\tilde{s}_2\neq s\\ \tilde{s}_1\neq\tilde{s}_2}} \sum_{\substack{\tilde{\imath}_1,\tilde{\imath}_2\\ \tilde{\imath}_2<\tilde{\imath}_1}} \frac{\partial \bar{P}_s}{\partial r^{\tilde{\imath}_1}_{\tilde{s}_1}} \bar{E}_0\left(\frac{\partial R^{\tilde{\imath}_1}_{\tilde{s}_1}}{\partial R^{\tilde{\imath}_2}_{\tilde{s}_2}}\right) \bar{E}_0(R^{\tilde{\imath}_2}_{\tilde{s}_2})\\
&\quad + \sum_{\substack{\tilde{s}_1\\ \tilde{s}_1\neq s}} \sum_{\substack{\tilde{\imath}_1,\tilde{\imath}_2\\ \tilde{\imath}_2<\tilde{\imath}_1}} \frac{\partial \bar{P}_s}{\partial r^{\tilde{\imath}_1}_{\tilde{s}_1}} \bar{E}_0\left(\frac{\partial R^{\tilde{\imath}_1}_{\tilde{s}_1}}{\partial R^{\tilde{\imath}_2}_{s}}\right) r^{\tilde{\imath}_2}_s + O(\bar{W}^3),
\end{aligned} \tag{A.6}$$

where we simply replaced all variations of q in (A.5) with s and replaced the $\tilde{p}$ in (A.5) with the corresponding $\tilde{s}$. Equation (A.6) relates the effective parameters θ (hidden in P) to the original model parameters $\bar{\theta}$ (hidden in $\bar{P}$). Since, by assumption, we have an algorithm to determine the effective parameters θ (at least for measured nodes), we want to be able to rewrite everything in terms of the effective parameters. To accomplish this, we need an expression for the original model parameters $\bar{\theta}$ in terms of the effective parameters θ.

Recall that each derivative with respect to r implicitly includes a factor of $\bar{W}$. Hence (A.6) shows that P_s deviates from $\bar{P}_s$ by an amount that is $O(\bar{W})$. Since we are computing only a second-order approximation in $\bar{W}$, we can replace $\bar{P}_s$ with P_s in any terms that are second-order in $\bar{W}$ (i.e., contain two derivatives with respect to r) without affecting the order of our approximation. Similarly expressions with E_0 differ by the equivalent expressions with $\bar{E}_0$ by an amount that is $O(\bar{W})$ (compare (3.7) with (A.2) and (A.4)), so we can also replace $\bar{E}_0$ with E_0 in terms that are second-order in $\bar{W}$. Then (A.6) becomes (after solving for $\bar{P}_s$)

$$\begin{aligned}
\bar{P}_s &= P_s - \sum_{\substack{\tilde{s}_1\\ \tilde{s}_1\neq s}} \sum_{\tilde{\imath}_1} \frac{\partial \bar{P}_s}{\partial r^{\tilde{\imath}_1}_{\tilde{s}_1}} \bar{E}_0(R^{\tilde{\imath}_1}_{\tilde{s}_1})\\
&\quad - \frac{1}{2} \sum_{\substack{\tilde{s}_1,\tilde{s}_2\\ \tilde{s}_1\neq s,\tilde{s}_2\neq s}} \sum_{\tilde{\imath}_1,\tilde{\imath}_2} \frac{\partial^2 P_s}{\partial r^{\tilde{\imath}_1}_{\tilde{s}_1} \partial r^{\tilde{\imath}_2}_{\tilde{s}_2}} E_0(R^{\tilde{\imath}_1}_{\tilde{s}_1} R^{\tilde{\imath}_2}_{\tilde{s}_2})\\
&\quad - \sum_{\substack{\tilde{s}_1,\tilde{s}_2\\ \tilde{s}_1\neq s,\tilde{s}_2\neq s\\ \tilde{s}_1\neq\tilde{s}_2}} \sum_{\substack{\tilde{\imath}_1,\tilde{\imath}_2\\ \tilde{\imath}_2<\tilde{\imath}_1}} \frac{\partial P_s}{\partial r^{\tilde{\imath}_1}_{\tilde{s}_1}} E_0\left(\frac{\partial R^{\tilde{\imath}_1}_{\tilde{s}_1}}{\partial R^{\tilde{\imath}_2}_{\tilde{s}_2}}\right) E_0(R^{\tilde{\imath}_2}_{\tilde{s}_2})\\
&\quad - \sum_{\substack{\tilde{s}_1\\ \tilde{s}_1\neq s}} \sum_{\substack{\tilde{\imath}_1,\tilde{\imath}_2\\ \tilde{\imath}_2<\tilde{\imath}_1}} \frac{\partial P_s}{\partial r^{\tilde{\imath}_1}_{\tilde{s}_1}} E_0\left(\frac{\partial R^{\tilde{\imath}_1}_{\tilde{s}_1}}{\partial R^{\tilde{\imath}_2}_{s}}\right) r^{\tilde{\imath}_2}_s + O(\bar{W}^3).
\end{aligned} \tag{A.7}$$

To write the right-hand side of (A.7) solely in terms of effective parameters θ, we need to change only the sum from the first line. Since this sum is $O(\bar{W})$, we need approximations to $\partial \bar{P}_s / \partial r^{\tilde{\imath}_1}_{\tilde{s}_1}$ and $\bar{E}_0(R^{\tilde{\imath}_1}_{\tilde{s}_1})$ that are accurate to first order in $\bar{W}$. We start with the first-order approximation of $\bar{P}_s$ (the first line of (A.7)):

$$\bar{P}_s = P_s - \sum_{\substack{\tilde{s}_1 \\ \tilde{s}_1 \neq s}} \sum_{\tilde{\imath}_1} \frac{\partial P_s}{\partial r^{\tilde{\imath}_1}_{\tilde{s}_1}} E_0(R^{\tilde{\imath}_1}_{\tilde{s}_1}) + O(\bar{W}^2). \tag{A.8}$$

Here we could replace $\bar{P}_s$ and $\bar{E}_0$ with P_s and E_0 in the terms that are first-order in $\bar{W}$, since we are computing only a first-order approximation.

When we differentiate $\bar{P}_s$ with respect to $r^{\tilde{\imath}_2}_{\tilde{s}_2}$, we are, by (3.6), essentially differentiating with respect to the $\bar{W}^{\tilde{\imath}_2,i}_{\tilde{s}_2,s}$. Hence, if we differentiate the left-hand side of (A.8) with respect to $r^{\tilde{\imath}_2}_{\tilde{s}_2}$, we need to differentiate only those terms on the right-hand side of (A.8) that are functions of P_s or $\bar{P}_s$. We obtain the following expression for the derivative $\partial \bar{P}_s / \partial r^{\tilde{\imath}_2}_{\tilde{s}_2}$ in terms of effective parameters:

$$\frac{\partial \bar{P}_s}{\partial r^{\tilde{\imath}_2}_{\tilde{s}_2}} = \frac{\partial P_s}{\partial r^{\tilde{\imath}_2}_{\tilde{s}_2}} - \sum_{\substack{\tilde{s}_1 \\ \tilde{s}_1 \neq s}} \sum_{\tilde{\imath}_1} \frac{\partial^2 P_s}{\partial r^{\tilde{\imath}_1}_{\tilde{s}_1} \partial r^{\tilde{\imath}_2}_{\tilde{s}_2}} E_0(R^{\tilde{\imath}_1}_{\tilde{s}_1}) + O(\bar{W}^2). \tag{A.9}$$

To find an expression for $\bar{E}_0(R^{\tilde{\imath}_1}_{\tilde{s}_1})$ in terms of effective parameters, we simplify its definition based on (A.2) to

$$\bar{E}_0(R^{\tilde{\imath}_1}_{\tilde{s}_1}) = \sum_{\mathbf{r}_{\tilde{s}_1}} r^{\tilde{\imath}_1}_{\tilde{s}_1} \bar{P}_{\tilde{s}_1}.$$

We similarly simplify the definition of $E_0(R^{\tilde{\imath}_1}_{\tilde{s}_1})$ (based on (3.7a)) to

$$E_0(R^{\tilde{\imath}_1}_{\tilde{s}_1}) = \sum_{\mathbf{r}_{\tilde{s}_1}} r^{\tilde{\imath}_1}_{\tilde{s}_1} P_{\tilde{s}_1}.$$

Then, by using (A.8) along with (3.7b), we can write $\bar{E}_0(R^{\tilde{\imath}_1}_{\tilde{s}_1})$ as

$$\begin{aligned}
\bar{E}_0(R^{\tilde{\imath}_1}_{\tilde{s}_1}) &= \sum_{\mathbf{r}_{\tilde{s}_1}} r^{\tilde{\imath}_1}_{\tilde{s}_1} \bar{P}_{\tilde{s}_1} \\
&= \sum_{\mathbf{r}_{\tilde{s}_1}} r^{\tilde{\imath}_1}_{\tilde{s}_1} P_{\tilde{s}_1} - \sum_{\mathbf{r}_{\tilde{s}_1}} \sum_{\substack{\tilde{s}_2 \\ \tilde{s}_2 \neq \tilde{s}_1}} \sum_{\tilde{\imath}_2} r^{\tilde{\imath}_1}_{\tilde{s}_1} \frac{\partial P_{\tilde{s}_1}}{\partial r^{\tilde{\imath}_2}_{\tilde{s}_2}} E_0(R^{\tilde{\imath}_2}_{\tilde{s}_2}) + O(\bar{W}^2) \\
&= E_0(R^{\tilde{\imath}_1}_{\tilde{s}_1}) - \sum_{\substack{\tilde{s}_2 \\ \tilde{s}_2 \neq \tilde{s}_1}} \sum_{\substack{\tilde{\imath}_2 \\ \tilde{\imath}_2 < \tilde{\imath}_1}} E_0\left(\frac{\partial R^{\tilde{\imath}_1}_{\tilde{s}_1}}{\partial R^{\tilde{\imath}_2}_{\tilde{s}_2}}\right) E_0(R^{\tilde{\imath}_2}_{\tilde{s}_2}) + O(\bar{W}^2).
\end{aligned} \tag{A.10}$$

We substitute (A.9) and (A.10) into the first line of (A.7) and obtain the following second-order expression of $\bar{P}_s$ in terms of effective parameters:

$$\begin{aligned}
\bar{P}_s = & P_s - \sum_{\substack{\tilde{s}_1 \\ \tilde{s}_1 \neq s}} \sum_{\tilde{\imath}_1} \frac{\partial P_s}{\partial r^{\tilde{\imath}_1}_{\tilde{s}_1}} E_0(R^{\tilde{\imath}_1}_{\tilde{s}_1}) \\
& - \frac{1}{2} \sum_{\substack{\tilde{s}_1, \tilde{s}_2 \\ \tilde{s}_1 \neq s, \tilde{s}_2 \neq s}} \sum_{\tilde{\imath}_1, \tilde{\imath}_2} \frac{\partial^2 P_s}{\partial r^{\tilde{\imath}_1}_{\tilde{s}_1} \partial r^{\tilde{\imath}_2}_{\tilde{s}_2}} [E_0(R^{\tilde{\imath}_1}_{\tilde{s}_1} R^{\tilde{\imath}_2}_{\tilde{s}_2}) - 2E_0(R^{\tilde{\imath}_1}_{\tilde{s}_1}) E_0(R^{\tilde{\imath}_2}_{\tilde{s}_2})] \\
& - \sum_{\substack{\hat{s}_2 \\ \hat{s}_2 \neq s}} \sum_{\substack{\tilde{\imath}_1, \tilde{\imath}_2 \\ \tilde{\imath}_2 < \tilde{\imath}_1}} \frac{\partial P_s}{\partial r^{\tilde{\imath}_1}_{\hat{s}_2}} E_0 \left(\frac{\partial R^{\tilde{\imath}_1}_{\hat{s}_2}}{\partial R^{\tilde{\imath}_2}_{s}} \right) [r^{\tilde{\imath}_2}_s - E_0(R^{\tilde{\imath}_2}_s)].
\end{aligned}$$

Since for $\tilde{s}_1 \neq \tilde{s}_2$, $E_0(R^{\tilde{\imath}_1}_{\tilde{s}_1} R^{\tilde{\imath}_2}_{\tilde{s}_2}) = E_0(R^{\tilde{\imath}_1}_{\tilde{s}_1}) E_0(R^{\tilde{\imath}_2}_{\tilde{s}_2})$, we can simplify this expression to obtain (3.8).

A.3. Measured node activity in terms of effective parameters. Equation (A.5) for $\Pr(\mathbf{R}_{\mathcal{Q}} | \mathbf{X})$, the probability distribution of the measured node activity, is given in terms of the original model parameters $\bar{\theta}$. Our next step is to use (3.8) to rewrite (A.5) in terms of effective parameters θ.

First, we rewrite (3.8) to replace the sums over all nodes in the network by two sums: one over the measured nodes and one over the hidden nodes. Recall that we use q (and its variants) to denote measured node indices and p (and its variants) to denote hidden node indices (i.e., implicitly restrict $q \in \mathcal{Q}$ and $p \in \mathcal{P}$).

$$\begin{aligned}
\bar{P}_s = {} & P_s - \sum_{\substack{\tilde{q}_1 \\ \tilde{q}_1 \neq s}} \sum_{\tilde{\imath}_1} \frac{\partial P_s}{\partial r^{\tilde{\imath}_1}_{\tilde{q}_1}} E_0(R^{\tilde{\imath}_1}_{\tilde{q}_1}) - \sum_{\substack{\tilde{p}_1 \\ \tilde{p}_1 \neq s}} \sum_{\tilde{\imath}_1} \frac{\partial P_s}{\partial r^{\tilde{\imath}_1}_{\tilde{p}_1}} E_0(R^{\tilde{\imath}_1}_{\tilde{p}_1}) \\
& - \frac{1}{2} \sum_{\substack{\tilde{q}_1 \\ \tilde{q}_1 \neq s}} \sum_{\tilde{\imath}_1, \tilde{\imath}_2} \frac{\partial^2 P_s}{\partial r^{\tilde{\imath}_1}_{\tilde{q}_1} \partial r^{\tilde{\imath}_2}_{\tilde{q}_1}} [E_0(R^{\tilde{\imath}_1}_{\tilde{q}_1} R^{\tilde{\imath}_2}_{\tilde{q}_1}) - E_0(R^{\tilde{\imath}_1}_{\tilde{q}_1}) E_0(R^{\tilde{\imath}_2}_{\tilde{q}_1})] \\
& - \frac{1}{2} \sum_{\substack{\tilde{p}_1 \\ \tilde{p}_1 \neq s}} \sum_{\tilde{\imath}_1, \tilde{\imath}_2} \frac{\partial^2 P_s}{\partial r^{\tilde{\imath}_1}_{\tilde{p}_1} \partial r^{\tilde{\imath}_2}_{\tilde{p}_1}} [E_0(R^{\tilde{\imath}_1}_{\tilde{p}_1} R^{\tilde{\imath}_2}_{\tilde{p}_1}) - E_0(R^{\tilde{\imath}_1}_{\tilde{p}_1}) E_0(R^{\tilde{\imath}_2}_{\tilde{p}_1})] \\
& - \sum_{\substack{\tilde{q}_1 \\ \tilde{q}_1 \neq s}} \sum_{\substack{\tilde{\imath}_1, \tilde{\imath}_2 \\ \tilde{\imath}_2 < \tilde{\imath}_1}} \frac{\partial P_s}{\partial r^{\tilde{\imath}_1}_{\tilde{q}_1}} E_0 \left(\frac{\partial R^{\tilde{\imath}_1}_{\tilde{q}_1}}{\partial R^{\tilde{\imath}_2}_{s}} \right) [r^{\tilde{\imath}_2}_s - E_0(R^{\tilde{\imath}_2}_s)] \\
& - \sum_{\substack{\tilde{p}_1 \\ \tilde{p}_1 \neq s}} \sum_{\substack{\tilde{\imath}_1, \tilde{\imath}_2 \\ \tilde{\imath}_2 < \tilde{\imath}_1}} \frac{\partial P_s}{\partial r^{\tilde{\imath}_1}_{\tilde{p}_1}} E_0 \left(\frac{\partial R^{\tilde{\imath}_1}_{\tilde{p}_1}}{\partial R^{\tilde{\imath}_2}_{s}} \right) [r^{\tilde{\imath}_2}_s - E_0(R^{\tilde{\imath}_2}_s)] \\
& + \frac{1}{2} \sum_{\substack{\tilde{q}_1, \tilde{q}_2 \\ \tilde{q}_1 \neq s, \tilde{q}_2 \neq s}} \sum_{\tilde{\imath}_1, \tilde{\imath}_2} \frac{\partial^2 P_s}{\partial r^{\tilde{\imath}_1}_{\tilde{q}_1} \partial r^{\tilde{\imath}_2}_{\tilde{q}_2}} E_0(R^{\tilde{\imath}_1}_{\tilde{q}_1}) E_0(R^{\tilde{\imath}_2}_{\tilde{q}_2}) \\
& + \sum_{\substack{\tilde{q}_1, \tilde{p}_2 \\ \tilde{q}_1 \neq s, \tilde{p}_2 \neq s}} \sum_{\tilde{\imath}_1, \tilde{\imath}_2} \frac{\partial^2 P_s}{\partial r^{\tilde{\imath}_1}_{\tilde{q}_1} \partial r^{\tilde{\imath}_2}_{\tilde{p}_2}} E_0(R^{\tilde{\imath}_1}_{\tilde{q}_1}) E_0(R^{\tilde{\imath}_2}_{\tilde{p}_2}) \\
& + \frac{1}{2} \sum_{\substack{\tilde{p}_1, \tilde{p}_2 \\ \tilde{p}_1 \neq s, \tilde{p}_2 \neq s}} \sum_{\tilde{\imath}_1, \tilde{\imath}_2} \frac{\partial^2 P_s}{\partial r^{\tilde{\imath}_1}_{\tilde{p}_1} \partial r^{\tilde{\imath}_2}_{\tilde{p}_2}} E_0(R^{\tilde{\imath}_1}_{\tilde{p}_1}) E_0(R^{\tilde{\imath}_2}_{\tilde{p}_2}) + O(\bar{W}^3).
\end{aligned} \tag{A.11}$$

The first term on the right-hand side of (A.5) for $\Pr(\mathbf{R}_{\mathcal{Q}}|\mathbf{X})$ is $\prod_q \bar{P}_q$. This is the only term that is zeroth-order in $\bar{W}$ and so is the only term where we need a second-order conversion from original paremeters $\bar{\theta}$ to effective parameters θ. We derive a second-order approximation of $\prod_q \bar{P}_q$ by taking the product of (A.11) (ignoring terms that are third- or higher-order in $\bar{W}$) and substitute this expression into (A.5). We use a first-order approximation of $\bar{P}_s$, $\bar{E}_0(R^i_s)$ (A.10), and $\partial \bar{P}_s / \partial r^{\tilde{i}}_{\tilde{s}}$ (A.9) to rewrite the first-order terms of (A.5) in terms of effective parameters. After simplification, (A.5) becomes

$$\begin{aligned}
\Pr(\mathbf{R}_{\mathcal{Q}} = \mathbf{r}_{\mathcal{Q}}|\mathbf{X} = \mathbf{x}) = {} & \prod_q P_q + \sum_{\substack{q_1,\tilde{q}_1\\ q_1\neq\tilde{q}_1}} \sum_{\tilde{i}_1} \frac{\partial P_{q_1}}{\partial r^{\tilde{i}_1}_{\tilde{q}_1}} [r^{\tilde{i}_1}_{\tilde{q}_1} - E_0(R^{\tilde{i}_1}_{\tilde{q}_1})] \prod_{\substack{q_2\\ q_2\neq q_1}} P_{q_2} \\
& + \sum_{\substack{q_1,\tilde{p}_1,\tilde{q}_2\\ \tilde{q}_2\neq q_1}} \sum_{\substack{\tilde{i}_1,\tilde{i}_2\\ \tilde{i}_1>\tilde{i}_2}} \frac{\partial P_{q_1}}{\partial r^{\tilde{i}_1}_{\tilde{p}_1}} E_0\left(\frac{\partial R^{\tilde{i}_1}_{\tilde{p}_1}}{\partial R^{\tilde{i}_2}_{\tilde{q}_2}}\right) [r^{\tilde{i}_2}_{\tilde{q}_2} - E_0(R^{\tilde{i}_2}_{\tilde{q}_2})] \prod_{\substack{q_2\\ q_2\neq q_1}} P_{q_2} \\
& + \frac{1}{2} \sum_{\substack{q_1,\tilde{p}_1,q_2\\ q_2\neq q_1}} \sum_{\tilde{i}_1,\tilde{i}_2} \frac{\partial P_{q_1}}{\partial r^{\tilde{i}_1}_{\tilde{p}_1}} \frac{\partial P_{q_2}}{\partial r^{\tilde{i}_2}_{\tilde{p}_1}} [E_0(R^{\tilde{i}_1}_{\tilde{p}_1} R^{\tilde{i}_2}_{\tilde{p}_1}) - E_0(R^{\tilde{i}_1}_{\tilde{p}_1}) E_0(R^{\tilde{i}_2}_{\tilde{p}_1})] \prod_{\substack{q_3\\ q_3\neq q_1, q_3\neq q_2}} P_{q_3} \\
& - \frac{1}{2} \sum_{\substack{q_1,\tilde{q}_1\\ q_1\neq\tilde{q}_1}} \sum_{\tilde{i}_1,\tilde{i}_2} \frac{\partial^2 P_{q_1}}{\partial r^{\tilde{i}_1}_{\tilde{q}_1} \partial r^{\tilde{i}_2}_{\tilde{q}_1}} [E_0(R^{\tilde{i}_1}_{\tilde{q}_1} R^{\tilde{i}_2}_{\tilde{q}_1}) - E_0(R^{\tilde{i}_1}_{\tilde{q}_1}) E_0(R^{\tilde{i}_2}_{\tilde{q}_1})] \prod_{\substack{q_2\\ q_2\neq q_1}} P_{q_2} \\
& - \sum_{\substack{q_1,\tilde{q}_1\\ \tilde{q}_1\neq q_1}} \sum_{\substack{\tilde{i}_1,\tilde{i}_2\\ \tilde{i}_1>\tilde{i}_2}} \frac{\partial P_{q_1}}{\partial r^{\tilde{i}_1}_{\tilde{q}_1}} E_0\left(\frac{\partial R^{\tilde{i}_1}_{\tilde{q}_1}}{\partial R^{\tilde{i}_2}_{q_1}}\right) [r^{\tilde{i}_2}_{q_1} - E_0(R^{\tilde{i}_2}_{q_1})] \prod_{\substack{q_2\\ q_2\neq q_1}} P_{q_2} \\
& + \frac{1}{2} \sum_{\substack{q_1,\tilde{q}_1,q_2,\tilde{q}_2\\ q_2\neq q_1, \tilde{q}_1\neq q_1\\ \tilde{q}_2\neq q_2}} \sum_{\tilde{i}_1,\tilde{i}_2} \frac{\partial P_{q_1}}{\partial r^{\tilde{i}_1}_{\tilde{q}_1}} \frac{\partial P_{q_2}}{\partial r^{\tilde{i}_2}_{\tilde{q}_2}} [r^{\tilde{i}_1}_{\tilde{q}_1} - E_0(R^{\tilde{i}_1}_{\tilde{q}_1})][r^{\tilde{i}_2}_{\tilde{q}_2} - E_0(R^{\tilde{i}_2}_{\tilde{q}_2})] \prod_{\substack{q_3\\ q_3\neq q_1, q_3\neq q_2}} P_{q_3} \\
& + \frac{1}{2} \sum_{\substack{q_1,\tilde{q}_1,\tilde{q}_2\\ q_1\neq\tilde{q}_1, q_1\neq\tilde{q}_2}} \sum_{\tilde{i}_1,\tilde{i}_2} \frac{\partial^2 P_{q_1}}{\partial r^{\tilde{i}_1}_{\tilde{q}_1} \partial r^{\tilde{i}_2}_{\tilde{q}_2}} [r^{\tilde{i}_1}_{\tilde{q}_1} - E_0(R^{\tilde{i}_1}_{\tilde{q}_1})][r^{\tilde{i}_2}_{\tilde{q}_2} - E_0(R^{\tilde{i}_2}_{\tilde{q}_2})] \prod_{\substack{q_2\\ q_2\neq q_1}} P_{q_2} + O(\bar{W}^3).
\end{aligned}$$

(A.12)

A.4. Transforming back to probability distribution. Equation (A.12) is a second-order approximation to a probability distribution, but it is not exactly a probability distribution. Since we wish to use $\Pr(\mathbf{R}_{\mathcal{Q}}|\mathbf{X})$ to compute maximum likelihood estimators of coupling parameters (i.e., find values of certain parameters that maximize $\Pr(\mathbf{R}_{\mathcal{Q}}|\mathbf{X})$), we need to use an expression for $\Pr(\mathbf{R}_{\mathcal{Q}}|\mathbf{X})$ that is a true probability distribution. For most terms of (A.12), one can simply reverse the Taylor expansion to pull terms back into the product of P_q.

However, one cannot simply reverse the Taylor expansion for the common input terms, i.e., the third line (common input from a hidden node onto two measured nodes) and the fourth line ("common input" from a measured node onto a single measured node). For those two terms, we'll need to tease apart the effects from different time points. We use the notation defined in (3.9) for P^i_s, the probability distribution at a single time point i (as well as its second derivative, defined analogously by (3.9)). We rewrite the derivatives with respect to r in terms of the P^i_s and its derivatives. We also separate out the common input effects at a single time point, rewriting the third and fourth lines of (A.12) as[9]

$$
\begin{aligned}
&+\frac{1}{2}\sum_{\substack{q_1,\tilde{p}_1,q_2\\ q_2\neq q_1}}\sum_{\tilde{\imath}_1,\tilde{\imath}_2}\frac{\partial P_{q_1}}{\partial r^{\tilde{\imath}_1}_{\tilde{p}_1}}\frac{\partial P_{q_2}}{\partial r^{\tilde{\imath}_2}_{\tilde{p}_1}}[E_0(R^{\tilde{\imath}_1}_{\tilde{p}_1}R^{\tilde{\imath}_2}_{\tilde{p}_1})-E_0(R^{\tilde{\imath}_1}_{\tilde{p}_1})E_0(R^{\tilde{\imath}_2}_{\tilde{p}_1})]\prod_{\substack{q_3\\ q_3\neq q_1,q_3\neq q_2}}P_{q_3}\\
&-\frac{1}{2}\sum_{\substack{q_1,\tilde{q}_1\\ q_1\neq\tilde{q}_1}}\sum_{\tilde{\imath}_1,\tilde{\imath}_2}\frac{\partial^2 P_{q_1}}{\partial r^{\tilde{\imath}_1}_{\tilde{q}_1}\partial r^{\tilde{\imath}_2}_{\tilde{q}_1}}[E_0(R^{\tilde{\imath}_1}_{\tilde{q}_1}R^{\tilde{\imath}_2}_{\tilde{q}_1})-E_0(R^{\tilde{\imath}_1}_{\tilde{q}_1})E_0(R^{\tilde{\imath}_2}_{\tilde{q}_1})]\prod_{\substack{q_2\\ q_2\neq q_1}}P_{q_2}\\
&=\sum_{\substack{q_1,\tilde{p}_1,q_2\\ q_2\neq q_1}}\sum_{\tilde{\imath}_1,\tilde{\imath}_2}\sum_{\substack{\hat{\imath}_1,\hat{\imath}_2\\ \hat{\imath}_2<\hat{\imath}_1}}\bar{W}^{\tilde{\imath}_1,\hat{\imath}_1}_{\tilde{p}_1,q_1}\bar{W}^{\tilde{\imath}_2,\hat{\imath}_2}_{\tilde{p}_1,q_2}\frac{\partial P^{\hat{\imath}_1}_{q_1}}{\partial w}\frac{\partial P^{\hat{\imath}_2}_{q_2}}{\partial w}[E_0(R^{\tilde{\imath}_1}_{\tilde{p}_1}R^{\tilde{\imath}_2}_{\tilde{p}_1})-E_0(R^{\tilde{\imath}_1}_{\tilde{p}_1})E_0(R^{\tilde{\imath}_2}_{\tilde{p}_1})]\frac{\prod_{q_3}P_{q_3}}{P^{\hat{\imath}_1}_{q_1}P^{\hat{\imath}_2}_{q_2}}\\
&+\frac{1}{2}\sum_{\substack{q_1,\tilde{p}_1,q_2\\ q_2\neq q_1}}\sum_{\tilde{\imath}_1,\tilde{\imath}_2}\sum_{\hat{\imath}_1}\bar{W}^{\tilde{\imath}_1,\hat{\imath}_1}_{\tilde{p}_1,q_1}\bar{W}^{\tilde{\imath}_2,\hat{\imath}_1}_{\tilde{p}_1,q_2}\frac{\partial P^{\hat{\imath}_1}_{q_1}}{\partial w}\frac{\partial P^{\hat{\imath}_1}_{q_2}}{\partial w}[E_0(R^{\tilde{\imath}_1}_{\tilde{p}_1}R^{\tilde{\imath}_2}_{\tilde{p}_1})-E_0(R^{\tilde{\imath}_1}_{\tilde{p}_1})E_0(R^{\tilde{\imath}_2}_{\tilde{p}_1})]\frac{\prod_{q_3}P_{q_3}}{P^{\hat{\imath}_1}_{q_1}P^{\hat{\imath}_1}_{q_2}}\\
&-\sum_{\substack{q_1,\tilde{q}_1\\ q_1\neq\tilde{q}_1}}\sum_{\tilde{\imath}_1,\tilde{\imath}_2}\sum_{\substack{\hat{\imath}_1,\hat{\imath}_2\\ \hat{\imath}_2<\hat{\imath}_1}}\bar{W}^{\tilde{\imath}_1,\hat{\imath}_1}_{\tilde{q}_1,q_1}\bar{W}^{\tilde{\imath}_2,\hat{\imath}_2}_{\tilde{q}_1,q_1}\frac{\partial P^{\hat{\imath}_1}_{q_1}}{\partial w}\frac{\partial P^{\hat{\imath}_2}_{q_1}}{\partial w}[E_0(R^{\tilde{\imath}_1}_{\tilde{q}_1}R^{\tilde{\imath}_2}_{\tilde{q}_1})-E_0(R^{\tilde{\imath}_1}_{\tilde{q}_1})E_0(R^{\tilde{\imath}_2}_{\tilde{q}_1})]\frac{\prod_{q_3}P_{q_3}}{P^{\hat{\imath}_1}_{q_1}P^{\hat{\imath}_2}_{q_1}}\\
&-\frac{1}{2}\sum_{\substack{q_1,\tilde{q}_1\\ q_1\neq\tilde{q}_1}}\sum_{\tilde{\imath}_1,\tilde{\imath}_2}\sum_{\hat{\imath}_1}\bar{W}^{\tilde{\imath}_1,\hat{\imath}_1}_{\tilde{q}_1,q_1}\bar{W}^{\tilde{\imath}_2,\hat{\imath}_1}_{\tilde{q}_1,q_1}\frac{\partial^2 P^{\hat{\imath}_1}_{q_1}}{\partial w^2}[E_0(R^{\tilde{\imath}_1}_{\tilde{q}_1}R^{\tilde{\imath}_2}_{\tilde{q}_1})-E_0(R^{\tilde{\imath}_1}_{\tilde{q}_1})E_0(R^{\tilde{\imath}_2}_{\tilde{q}_1})]\frac{\prod_{q_3}P_{q_3}}{P^{\hat{\imath}_1}_{q_1}}.
\end{aligned}
$$

The last line in the above equation corresponds to the second-order effect of a single connection between two measured nodes. For this term, we cannot reverse the Taylor expansion to fold the term back into the product of the P_q and create a probability distribution. However, this term represents a second-order effect that is not summed over all nodes of the network (it is simply summed over the measured nodes, which we view as a small subset). If we modify our weak coupling assumption to allow us to ignore second-order terms that are not summed over all nodes, we can neglect this last term. Since we no longer have exactly a second-order approximation in $\bar{W}$, we denote the approximation by $\approx$.

With this approximation, we can reverse the Taylor expansion of the remaining terms of (A.12) and obtain (3.10) for $\Pr(\mathbf{R}_{\mathcal{Q}}|\mathbf{X})$, which is written as a probability distribution in terms of effective parameters.

Appendix B. Estimation of single-node parameters. We sketch our algorithm for determining the single-node parameters θ^i_s of model (4.2) that we used to analyze the results of our simulations. The parameters θ^i_s correspond to A_s, y_s, $\mathbf{h}_{\text{hist},s}$, and $\mathbf{h}_{\text{ext},s}$. We calculated maximum likelihood estimators of these parameters from measurements of R^i_s, the spikes of neuron s, and the stimulus $\mathbf{X}$.

We chose our form (4.2) of $\lambda_s(\cdot)$ so that $\lambda_s(\cdot)$ is convex and $\log\lambda_s(\cdot)$ is concave as a function of y_s, $\mathbf{h}_{\text{hist},s}$, and $\mathbf{h}_{\text{ext},s}$. In this way, for a fixed A_s, the log-likelihood surface (logarithm of (2.3)) is free of nonglobal local maxima [17], and we could use standard gradient ascent algorithms to find the maximum, conditioned on a value of A_s. (We used the Polak–Ribiere conjugate gradient algorithm as implemented in the GNU Scientific Library [6].)

[9] Note that all of the probabilities P^i_q that appear in a denominator are also a factor in the corresponding numerator. If a P^i_q that appears in a denominator were to be zero, one could still define the ratio by canceling the factor in the numerator.

Before calculating these maximum likelihood estimators, we calculated the absolute refractory period τ_s^{absref} as the minimum number of $\Delta t = 1$ ms time bins observed between spikes. Then, so that our model predicts absolutely no firing for τ_s^{absref} time steps after each spike, we set $h^i_{\text{hist},s} = -10^{100}$ for $i \leq \tau_s^{\text{absref}}$. To reduce the dimension of the parameter space, we restricted the remainder of the history kernel $\mathbf{h}_{\text{hist},s}$ to be in the subspace spanned by the vectors

$$B^k_{s,1}(i) = \sin\left(\pi k \left[2\frac{i - \tau_s^{\text{absref}}}{\tau_{s,1}} - \left(\frac{i - \tau_s^{\text{absref}}}{\tau_{s,1}}\right)^2\right]\right)$$

for $0 < i - \tau_s^{\text{absref}} < \tau_{s,1}$ and $B^k_{s,1}(i) = 0$ otherwise. (These vector are not orthogonal, so we applied Gram–Schmidt orthonormalization to obtain basis vectors.) We set $\tau_{s,1} = 60 - \tau_s^{\text{absref}}$ time bins. These basis vectors are analagous to those used in [8]; they can represent fine temporal structure for the time immediately after the spike but are smoother for longer time scales. We used 29 basis vectors $1 \leq k \leq 29$ (viewing the 30th basis vector as capturing the absolute refractory period).

We similarly reduced the dimension of $\mathbf{h}_{\text{ext},s}$ by using basis vectors that are a product of a Hartley basis function in space (to match the stimulus) and temporal basis functions similar to the $B^k_{s,1}$. The basis function indexed by k and l evaluated at time bin i and space bin j was based on

$$B^{k,l}_{s,2}(i,j) = \text{cas}(2\pi l j/N_0) \sin\left(\pi k \left[2i/\tau_{s,2} - (i/\tau_{s,2})^2\right]\right)$$

for $0 < i < \tau_{s,2}$ and $B^{k,l}_{s,2}(i,j) = 0$ otherwise (again, we obtained orthogonal basis functions through Gram–Schmidt orthonormalization). As in the definition of the stimulus (section 4.1.1), $\text{cas}\, x = \cos x + \sin x$ and $N_0 = 100$. We set $\tau_{s,1} = 200$ time bins. We used the 210 basis vectors $-10 \leq l \leq 10$ and $1 \leq k \leq 10$.

As mentioned above, we calculated y_0 and the coefficients of the basis functions to maximize the log-likelihood, given a fixed value of A_s. This defines all parameters as a function of A_s. We then search for a value of A_s that maximizes the log-likelihood while keeping the other parameters set at this function of A_s. We use this procedure since the log-likelihood surface may not be well-behaved as a function of A_s.

Recall that the causal connection measure W and the common input measure U are maximum likelihood estimators based on (3.15), which depends on these values of θ^i_s. To reduce bias at this stage, we calculate the θ^i_s using cross-validation. We divided the data into 4 segments. For each time bin i from one of these segments, we calculated the parameters θ^i_q using only the data in the other 3 segments. (For computation efficiency, we don't recalculate A_s four times but base A_s from calculations using all of the data.)

Appendix C. Monte Carlo estimates of single-node expected values. The estimation of connectivity parameters is based on (3.15) for $\Pr(\mathbf{R}_\mathcal{Q}|\mathbf{X})$, the probability distribution of measured node activity. Once the effective parameters θ_q of the measured nodes have been estimated, the only unknown quantities in (3.15) are the causal connection W and common input U parameters. However, some of the known quantities are given as expected values of functions of the measured node activities as predicted by the averaged model (2.3). Although these expected values are completely determined by the averaged model and the known effective parameters θ_q, computing them explicitly would be impractical, as one would need to enumerate all possible sequences of the history of each node and average over them all.[10] Instead, for each measured node, we use the averaged model (2.3) to randomly generate sequences of activity in order to estimate these expected values using Monte Carlo.

There are three different expected values that appear in (3.15b). They are the average activity $E_0(R_q^i)$ at a given time bin, the second moment $E_0(R_q^i R_q^{i-j})$, and the expected value involving the derivative $E_0(R_q^i(\partial P_q^{i-j}/\partial w)/P_q^{i-j})$. To estimate these expected values via Monte Carlo, we randomly generate a sequence $\mathbf{R}_q$ of the activity of the node from the averaged model (2.3). Then, at each time point i (ignoring initial time points for which we don't have enough preceeding history), we make a sample estimate of each expected value, as described below. We repeat this process 1000 times, setting our final estimates to be averages of these 1000 samples.

To compute the average activity $E_0(R_q^i)$, we could simply record the sampled R_q^i and average these. However, we improve our estimate by taking advantage of the fact that we have an analytic expression for the mean of R_q^i conditioned on the history $\mathbf{R}_q^{<i}$ (for the Poisson distribution it is simply $\lambda_q(\mathbf{R}_q^{<i}, \mathbf{x}, 0; \theta_q^i)$). Our estimate of $E_0(R_q^i)$ is the average of such conditioned means.

In our examples, we use the Poisson distribution (section 3.4) for the probability distribution $P_q(R_q^i, \mathbf{R}_q^{<i}, \cdot)$ of R_q^i conditioned on the history $\mathbf{R}_q^{<i}$. However, one must remember that R_q^i no longer has a probability distribution of the form $P_q(R_q^i, \mathbf{R}_q^{<i}, \cdot)$ once one averages over all possible histories. Since R_q^i does not have a Poisson distribution, one must resist the urge to estimate the variance $E_0((R_q^i)^2) - E_0(R_q^i)E_0(R_q^i)$ as being equal to the mean $E_0(R_q^i)$. Instead, one must calculate $E_0((R_q^i)^2)$ in the same manner as that described above for calculating $E_0(R_q^i)$. Since we have an analytic formula for the second moment of R_q^i conditioned on this history $\mathbf{R}_q^{<i}$ (for the Poisson distribution, it is $\lambda_q^2 + \lambda_q$), we can estimate $E_0((R_q^i)^2)$ as the average of such conditioned second moments. To estimate $E_0(R_q^i R_q^{i-j})$ (for $j > 0$), we take our analytic expression for the average of R_q^i conditioned the history $\mathbf{R}_q^{<i}$, multiply it by the sampled value of R_q^{i-j}, and average over all samples.

For the derivative term, $E_0(R_q^i(\partial P_q^{i-j}/\partial w)/P_q^{i-j})$, we first look at the $j = 0$ case. We can rewrite it as

$$E_0\left(R_q^i \frac{\partial P_q^i}{\partial w} \frac{1}{P_q^i}\right) = \sum_{\mathbf{r}_q^{<i+1}} r_q^i \frac{\partial P_q^i}{\partial w} \frac{1}{P_q^i} \prod_{\tilde{i} \le i} P_q^{\tilde{i}} = \sum_{\mathbf{r}_q^{<i+1}} r_q^i \frac{\partial P_q^i}{\partial w} \prod_{\tilde{i} < i} P_q^{\tilde{i}}, \tag{C.1}$$

where the sum is over all possible values of the r_q^k for $k \le i$. At least for the Poisson distribution, we can calculate an analytic expression[11] for $\sum_{r_q^i} r_q^i \partial P_q^i / \partial w$, and we take the average of that quantity over all samples. For $j > 0$, the term is

$$E_0\left(R_q^i \frac{\partial P_q^{i-j}}{\partial w} \frac{1}{P_q^{i-j}}\right) = \sum_{\mathbf{r}_q^{<i+1}} r_q^i \frac{\partial P_q^{i-j}}{\partial w} \frac{1}{P_q^{i-j}} \prod_{\tilde{i} \le i} P_q^{\tilde{i}}. \tag{C.2}$$

[10] Hence, the computational cost would increase exponentially in length of the history that could affect the activity.

[11] For the Poisson distribution, $\sum_{r_q^i} r_q^i \partial P_q^i / \partial w = \partial_w \lambda_q(\mathbf{r}_q^{<i}, \mathbf{x}, 0; \theta_q^i)$.

In this case, we take the average value of R_q^i conditioned on the sampled history and multiply it by $(\partial P_q^{i-j}/\partial w)/P_q^{i-j}$. We average this quantity over all samples.[12]

Acknowledgments. The author thanks Liam Paninski, Jayant Kulkarni, Jonathan Pillow, and an anonymous reviewer for helpful input.

REFERENCES

[1] A. M. H. J. Aertsen, G. L. Gerstein, M. K. Habib, and G. Palm, *Dynamics of neuronal firing correlation: Modulation of "effective connectivity"*, J. Neurophysiol., 61 (1989), pp. 900–917.

[2] E. N. Brown, R. E. Kass, and P. P. Mitra, *Multiple neural spike train data analysis: State-of-the-art and future challenges*, Nat. Neurosci., 7 (2004), pp. 456–461.

[3] E. S. Chornoboy, L. P. Schramm, and A. F. Karr, *Maximum likelihood identification of neural point process systems*, Biol. Cybern., 59 (1988), pp. 265–275.

[4] D. R. Cox and V. Isham, *Point Processes*, Chapman and Hall, New York, 1980.

[5] D. J. Daley and D. Vere-Jones, *An Introduction to the Theory of Point Processes*, Springer-Verlag, New York, 1988.

[6] M. Galassi, J. Davies, J. Theiler, B. Gough, G. Jungman, M. Booth, and F. Rossi, *GNU Scientific Library Reference Manual*, 2nd ed., Network Theory Ltd., Bristol, United Kingdom, 2003; also available online from http://www.gnu.org/software/gsl/.

[7] K. D. Harris, J. Csicsvari, H. Hirase, G. Dragoi, and G. Buzsáki, *Organization of cell assemblies in the hippocampus*, Nature, 424 (2003), pp. 552–556.

[8] J. Keat, P. Reinagel, R. C. Reid, and M. Meister, *Predicting every spike: A model for the responses of visual neurons*, Neuron, 30 (2001), pp. 803–817.

[9] J. E. Kulkarni and L. Paninski, *Common-Input Models for Multiple Neural Spike-Train Data*, Network: Comput. Neural Syst., to appear.

[10] S. Marcelja, *Mathematical description of the responses of simple cortical cells*, J. Opt. Soc. Amer., 70 (1980), pp. 1297–1300.

[11] L. Martignon, G. Deco, K. Laskey, M. Diamond, W. Freiwald, and E. Vaadia, *Neural coding: Higher-order temporal patterns in the neurostatistics of cell assemblies*, Neural Comp., 12 (2000), pp. 2621–2653.

[12] D. Q. Nykamp, *Reconstructing Stimulus-Driven Neural Networks from Spike Times*, in Advances in Neural Information Processing Systems 15, S. Becker, S. Thrun, and K. Obermayer, eds., MIT Press, Cambridge, MA, 2003, pp. 309–316.

[13] D. Q. Nykamp, *Revealing pairwise coupling in linear-nonlinear networks*, SIAM J. Appl. Math., 65 (2005), pp. 2005–2032.

[14] D. Q. Nykamp, *A mathematical framework for inferring connectivity in probabilistic neuronal networks*, Math. Biosci., 205 (2007), pp. 204–251.

[15] M. Okatan, M. A. Wilson, and E. N. Brown, *Analyzing functional connectivity using a network likelihood model of ensemble neural spiking activity*, Neural Comp., 17 (2005), pp. 1927–1961.

[16] G. Palm, A. M. H. J. Aertsen, and G. L. Gerstein, *On the significance of correlations among neuronal spike trains*, Biol. Cybern., 59 (1988), pp. 1–11.

[17] L. Paninski, *Maximum likelihood estimation of cascade point-process neural encoding models*, Network: Comput. Neural Syst., 15 (2004), pp. 243–262.

[18] L. Paninski, J. W. Pillow, and E. P. Simoncelli, *Maximum likelihood estimation of a stochastic integrate-and-fire neural encoding model*, Neural Comp., 16 (2004), pp. 2533–2561.

[19] D. H. Perkel, G. L. Gerstein, and G. P. Moore, *Neuronal spike trains and stochastic point processes*. II. *Simultaneous spike trains*, Biophys. J., 7 (1967), pp. 419–440.

[20] J. R. Rosenberg, A. M. Amjad, P. Breeze, D. R. Brillinger, and D. M. Halliday, *The Fourier approach to the identification of functional coupling between neuronal spike trains*, Prog. Biophys. Mol. Biol., 53 (1989), pp. 1–31.

[12] Here, we couldn't avoid explicitly dividing by P_q^{i-j} because the conditioned expected value of R_q^i depended on the particular value of R_q^{i-j} that we randomly generated. Note that P_q^{i-j} must be greater than zero for this value of R_q^{i-j} because R_q^{i-j} was randomly generated with probabilty P_q^{i-j}.

[21] M. N. Shadlen and W. T. Newsome, *The variable discharge of cortical neurons: implications for connectivity, computation, and information coding*, J. Neurosci., 18 (1998), pp. 3870–3896.

[22] D. Snyder and M. Miller, *Random Point Processes in Time and Space*, Springer-Verlag, Berlin, 1991.

[23] C. F. Stevens and A. M. Zador, *Input synchrony and the irregular firing of cortical neurons*, Nat. Neurosci., 1 (1998), pp. 210–217.

[24] L. Stuart, M. Walter, and R. Borisyuk, *The correlation grid: Analysis of synchronous spiking in multi-dimensional spike train data and identification of feasible connection architectures.*, Biosystems, 79 (2005), pp. 223–234.

[25] W. Truccolo, U. T. Eden, M. R. Fellows, J. P. Donoghue, and E. N. Brown, *A point process framework for relating neural spiking activity to spiking history, neural ensemble, and extrinsic covariate effects*, J. Neurophysiol., 93 (2005), pp. 1074–1089.