Statistically inferred neuronal connections in subsampled neural networks strongly correlate with spike train covariances

Tong Liang[1,2] and Braden A. W. Brinkman[2,*]

[1]*Department of Physics and Astronomy, Stony Brook University, Stony Brook, New York 11794, USA*
[2]*Department of Neurobiology and Behavior, Stony Brook University, Stony Brook, New York 11794, USA*

(Received 8 November 2023; accepted 29 February 2024; published 5 April 2024)

Statistically inferred neuronal connections from observed spike train data are often skewed from ground truth by factors such as model mismatch, unobserved neurons, and limited data. Spike train covariances, sometimes referred to as "functional connections," are often used as a proxy for the connections between pairs of neurons, but reflect statistical relationships between neurons, not anatomical connections. Moreover, covariances are not causal: spiking activity is correlated in both the past and the future, whereas neurons respond only to synaptic inputs in the past. Connections inferred by maximum likelihood inference, however, can be constrained to be causal. However, we show in this work that the inferred connections in spontaneously active networks modeled by stochastic leaky integrate-and-fire networks strongly correlate with the covariances between neurons, and may reflect noncausal relationships, when many neurons are unobserved or when neurons are weakly coupled. This phenomenon occurs across different network structures, including random networks and balanced excitatory-inhibitory networks. We use a combination of simulations and a mean-field analysis with fluctuation corrections to elucidate the relationships between spike train covariances, inferred synaptic filters, and ground-truth connections in partially observed networks.

DOI: 10.1103/PhysRevE.109.044404

I. INTRODUCTION

Identifying the strength and timescales of synaptic transmission between neuron pairs is a major goal of neuroscience, as it would greatly facilitate our understanding of how a neural circuit's computational properties are shaped by its structure. It is now possible to record simultaneous activity from large populations of neurons, enabling the use of statistical methods to infer interactions between neuron pairs [1–4]. This has become a foundational tool for understanding the encoding and decoding properties of many biological neural networks. However, because no *in vivo* recording technique can record from all neurons in a circuit (Fig. 1) the inferred connections between neurons may only reflect statistical relationships between neurons, shaped by, but not necessarily representative of, the underlying anatomical connections [5–8].

To this end, it is useful to distinguish between two prominent measures of "functional" or "effective" interactions between neurons. "Functional" connections between neurons are estimated using pairwise crosscovariances between neuron spike trains, or other measures of neural activity, such as BOLD signals in fMRI [9]. Crosscovariances are not causal functions, as both the past and future of two spike trains are correlated. As a result, the covariance cannot generally predict whether the past spiking activity of one neuron drives the future activity of another neuron, and extracting information about causal circuit responses from functional connections is not always possible. In contrast, "effective" interactions obtained by performing, e.g., maximum likelihood inference on neural activity data, can be constrained to be causal functions, and could therefore represent actual causal responses of neurons to spikes from pre-synaptic partners. In principle, effective interactions should therefore be more useful for understanding how the underlying dynamics of neurons implement computations. However, in this work we use a combination of simulations and analytically tractable cases to show that when only a few neurons are recorded in a neural circuit—the typical case in any *in vivo* recording—the effective filters inferred from spontaneous neural activity correlate strongly with the corresponding half of the underlying spike train covariances. Relationships between covariances and the underlying synaptic connectivity of network models have previously been derived [10–12], but fewer studies have investigated the relationships to synaptic filters estimated by statistical inference. Other studies that have investigated this problem have focused on the effect of unobserved neurons in weakly coupled networks [5] or fully observed strongly coupled networks [13]. Our work will span these extremes.

We define the model and present our results in Sec. II and discuss the implications and directions for future work in Sec. III.

II. RESULTS

A. Model definition

Our goal is to understand how the circuit properties inferred using maximum likelihood inference are related to the ground-truth properties of the network or the statistics of spontaneous neural activity—that is, in this work we do not consider stimulus-driven activity. We require both a generative model and an inference model. We choose to use the

[*]braden.brinkman@stonybrook.edu

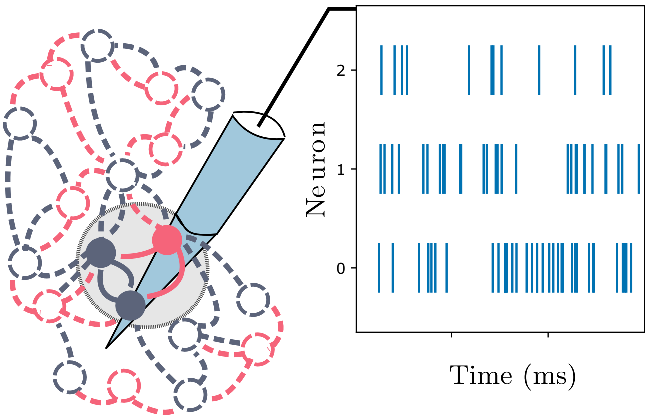

FIG. 1. Schematics of the hidden neuron problem and effective neuronal connection inference. Recording techniques are often only able to record from a subset of neurons in a neural circuit, especially in the tissue of a living organism. In this schematic three neurons' spike trains are shown to be recorded, while the activity of other nearby neurons remains unobserved. Statistical inference techniques applied to this activity data can predict effective connections between neurons, but these do not necessarily reflect the true anatomical connections.

same model family for both, a generalized linear point process model (GLM). As a generative model the GLM can be interpreted as a leaky integrate-and-fire model in which spike emission is stochastic [10,11,14,15]. As an inference model the GLM has been used to fit neural spiking data from many brain areas, including the retina [1] and the lateral intraparietal area of macaques [16].

The GLM models spike train emission as an inhomogeneous Poisson process in which the instantaneous firing rate is conditioned on the past history of neural activity in the network:

$$\dot{n}_i(t)dt \sim \text{Poiss}\left[\phi\left(\mu_i + \sum_{j=1}^{N}\int dt'\, J_{ij}(t-t')\dot{n}_j(t')\right)dt\right], \tag{1}$$

where $\dot{n}_i(t)dt$ is the number of spikes neuron i fires within a small window $[t, t+dt]$, $\phi(x)$ is a nonlinear activation function, μ_i is the baseline drive for neuron i that sets the baseline firing rate, and $J_{ij}(t)$ is the synaptic interaction or coupling filter from neuron j to neuron i; we will use the terms "interaction" and "coupling" interchangeably in this paper. We allow the "autocoupling" filter $J_{ii}(t)$, often called the self-history filter, to be nonzero and typically negative to allow neurons to suppress their own firing after spiking, mimicking the effects of membrane potential hyperpolarization observed experimentally. In this work we do not assign neurons spatial locations, so the indices correspond only to an arbitrary ordering of neurons, though the order may correspond to properties such as excitatory or inhibitory cell types. The parameters to be inferred from data are the baselines μ_i and synaptic filters $J_{ij}(t-t')$; the nonlinearity $\phi(x)$ is often fixed, the canonical choice being an exponential, $\phi(x) = \lambda_0 \exp(x)$ [1], where λ_0 is a constant that determines the units of time. We will adopt this choice for our inference model, as it offers several simplifications in both our statistical inference procedure and mathematical analysis of the maximum likelihood procedure (but see Appendix K for a brief discussion of the expected effects of nonlinearity mismatch.). It is important to stress that when the number of observed neurons N_{obs} is not equal to the number of neurons N in the generative model we are dealing with a model mismatch problem and we do not expect statistical inference to recover the parameters of the generative model [11,13]. To obtain the true generative model of just the observed neurons one should in principle marginalize out the unobserved neurons. This was done approximately by Ref. [11], but the resulting model is much more complex than the fully observed network model, and may not be suitable as a statistical inference model. It is therefore worth understanding how the inferred filters relate to the ground-truth circuit properties when Eq. (1) is used as the inference model with $N_{\text{obs}} < N$.

The maximum likelihood estimates (MLE) of circuit properties are the model parameters that render the observed data as probable as possible under the inference model. These parameters can be found by maximizing the (log)-likelihood. Because at each time step the GLM is Poisson conditioned on the spike train history, each neuron is conditionally independent and each has its own likelihood, which is a product of Poisson distributions at each time point:

$$\begin{aligned} L_i(\hat{\mu}, \hat{J}) &= \text{Prob}(\{\dot{n}_i(t)\}|\hat{\mu}_i, \hat{J}_{ij}) \\ &= \prod_t \frac{(\hat{\phi}_i(t)dt)^{\dot{n}_i(t)dt}}{(\dot{n}_i(t)dt)!} e^{-\hat{\phi}_i(t)dt}, \end{aligned} \tag{2}$$

where the equation should be interpreted in discrete time; we take the continuous time limit momentarily. The nonlinearity of the inference model is $\hat{\phi}_i(t) = \lambda_0 \exp(\hat{\mu}_i + \sum_j \int dt'\, \hat{J}_{ij}(t-t')\dot{n}_j(t'))$; we use hats to distinguish parameters of the inference model from their corresponding ground-truth counterparts. Note that the parameters $\hat{\mu}_i$ and $\hat{J}_{ij}(t-t')$ may be inferred independently for each neuron i. In the likelihood function the spike trains $\dot{n}_i(t)$ are considered to be observed from a single long trial.

After taking the logarithm we may properly take the continuum time limit to obtain the log-likelihood $\mathcal{L}_i(\{\mu\}, \{J\}) = T^{-1} \ln L_i(\hat{\mu}, \hat{J})$,

$$\mathcal{L}_i(\{\mu\}, \{J\}) = \frac{1}{T}\int_{t_0}^{t_0+T} dt\, \{\dot{n}_i(t) \ln \hat{\phi}_i(t) - \hat{\phi}_i(t)\}, \tag{3}$$

where we have dropped the parameter-independent term $-\ln(\dot{n}_i(t)dt)!$ and normalized the log-likelihood by the duration of the spike trains, T. We take the time window to be $t \in [t_0, t_0 + T]$ for some initial time t_0. The maximum likelihood estimates of the parameters $\hat{\mu}_i$ and $\hat{J}_{ij}(t)$ are those that maximize $\mathcal{L}_i(\{\mu\}, \{J\})$ for the observed data. Taking derivatives of the log-likelihood with respect to the parameters and equating the derivatives to zero yields implicit maximum likelihood estimation equations:

$$\frac{1}{T}\int_{t_0}^{t_0+T} dt\, \dot{n}_i(t) = \frac{1}{T}\int_{t_0}^{t_0+T} dt\, \hat{\phi}_i(t), \tag{4}$$

$$\frac{1}{T}\int_{t_0}^{t_0+T} dt\, \dot{n}_i(t)\dot{n}_j(t-\tau_{\text{lag}}) = \frac{1}{T}\int_{t_0}^{t_0+T} dt\, \hat{\phi}_i(t)\dot{n}_j(t-\tau_{\text{lag}}), \tag{5}$$

for a time-lag $\tau_{\rm lag}$, and where we have explicitly used the fact that we take the nonlinearity $\hat{\phi}$ to be exponential; see Appendix D for the derivation of the MLE equations. The left-hand sides of Eqs. (4) and (5) can be recognized as the empirical mean and covariance, respectively, of the observed spike trains.

We solve Eqs. (4) and (5) and investigate the inferred synaptic interaction filters in two ways. In Sec. II B we perform maximum likelihood estimation on simulated data, mimicking analysis of real data, and in Sec. II C we analyze the inference problem theoretically, deriving an approximate system of equations for the inferred filters that we can study analytically. In both cases we focus on how the inferred filters change as a function of the number of neurons recorded.

B. Simulations

We simulate two types of circuit networks in this study. The first network type comprises random Gaussian networks in which $J_{ij}(t) = \mathcal{J}_{ij} t e^{-t/\tau}/\tau^2 \Theta(t)$, where τ is the membrane time-constant and the synaptic weights $\mathcal{J}_{ij}$ are normally distributed with zero mean and variance $J_0^2/(pN)$: $\mathcal{J}_{ij} \sim \mathcal{N}(0, J_0^2/(pN))$. We set the baseline $\mu_i = -2$ for all neurons and set the network sparsity to $p = 50\%$ (i.e., each synaptic connection may be nonzero with a probability $p = 50\%$). For an exponential nonlinearity this allows us to tune the weight matrix coefficient J_0 up to an integer value of ~ 3 before the network becomes unstable. We do not need to restrict our simulations or analysis to the weak-coupling regime [5].

Though random networks are commonly used in theoretical studies, they lack biological realism because every neuron can make both excitatory (positive) and inhibitory (negative) synaptic connections. We therefore also consider balanced networks of excitatory and inhibitory (E-I) populations, for which each neuron is either excitatory or inhibitory and only makes connections of a single sign [17,18].

We simulate Eq. (1) by discretizing the spiking process into bins of size Δt. We determine the parameters that solve the discrete-time MLE Eqs. (4) and (5) by gradient ascent. For each network type we simulate the spiking activity of 64 neurons for 2 million time points. While simulating larger networks is possible, fitting to larger networks is very memory intensive. To fit the GLM to these simulated data sets it is necessary to parametrize the filters $\hat{J}_{ij}(t)$, either by inferring the value of the filter at each discrete time point (requiring as many parameters as the number of time-bins used to represent the filter) or by representing the filters as weighted sums of basis functions $\alpha_n(t)$, $\hat{J}_{ij}(t) = \sum_n w_{ij}^{(n)} \alpha_n(t)$, and inferring the unknown weights $w_{ij}^{(n)}$. The basis function approach reduces the number of unknowns to the number of weights, which requires less data than inferring each time point. The filter shapes that can be inferred are constrained by one's choice of basis functions $\alpha_n(t)$, whereas inferring each time-point can represent any function given enough temporal resolution and data, but often results in noisy estimates without adding fit penalties to impose smoothness. In most studies the basis function representation is preferred, but for our analyses the time-point inference will reveal interesting relationships between the inferred filters and statistics of the circuit activity. For this reason we also do not impose any regularization on the basis-less fit to smooth out the inferred filters. For additional details on our simulations, see Appendix A.

In Fig. 2 we show the results of statistical inference on $N_{\rm obs} = 3$ observed neurons out of 64 from the sparse random network, comparing the inferred filters and spike train covariances to the ground-truth filters. The results show that neither the inferred filters (red solid lines, basis fit; red points, basis-less fit) nor the empirical covariances (grey bars) match the ground-truth filters of the generative model (blue dashed lines). However, we observe that the basis-less inferred filters match quite well with the empirical covariances. Since the MLE inferred filters and spike train covariances are estimated using two independent methods, it is surprising to observe such a strong correlation between them. Elucidating this strong correlation between inferred filters and covariances, and their deviation from the ground-truth filters in subsampled networks, is the main goal of this work.

1. Only the strongest spike train covariances are indicative of true connections

First, we investigate the spike covariances in the random and E-I networks. In Fig. 3 we plot histograms of the directed magnitudes of the covariances, defined as $||C_{ij}|| \equiv \sqrt{\int_{0^+}^{\infty} dt\, C_{ij}(t)^2}$; note that we integrate over only positive lags, so the magnitude of $||C_{ij}|| \neq ||C_{ji}||$ and the δ function at zero-lag is excluded when calculating the directed magnitudes of the autocovariances $||C_{ii}||$. We condition the histograms by whether $\mathcal{J}_{ij} = 0$ or not. For sparse random networks we find that (i) the autocovariances have a strong positive peak near zero-lag, unlike the purely negative autocoupling filter that implements the soft-reset of the membrane potential, and (ii) the largest covariances between synaptically connected neurons are larger than the covariances between neurons that are not connected, though there is considerable overlap between the distributions. This means that a strong covariance may be indicative of a genuine synaptic connection between neurons, but weaker covariances may not distinguish between absence or presence of connections. The results hold in both sparse random networks and the E-I networks, even when separated by the types of neurons connected.

2. Correlations between spike train covariances, ground-truth filters, and inferred filters shift as fewer neurons are observed

Next, we investigate the apparent correlation between the maximum likelihood estimates of the filters with the spike train covariances. To verify the correlation is robust and not just a visual artifact, we calculate the linear Pearson correlation coefficient between the ground-truth filters, inferred filters, and the empirical spike train covariances, varying the number of observed neurons in both random and E-I networks.

We first look at the correlations between the positive-lag halves of the empirical covariances and the ground-truth filters, shown in Fig. 4. We see that there is generally a positive correlation between the two, though there is quite a bit of

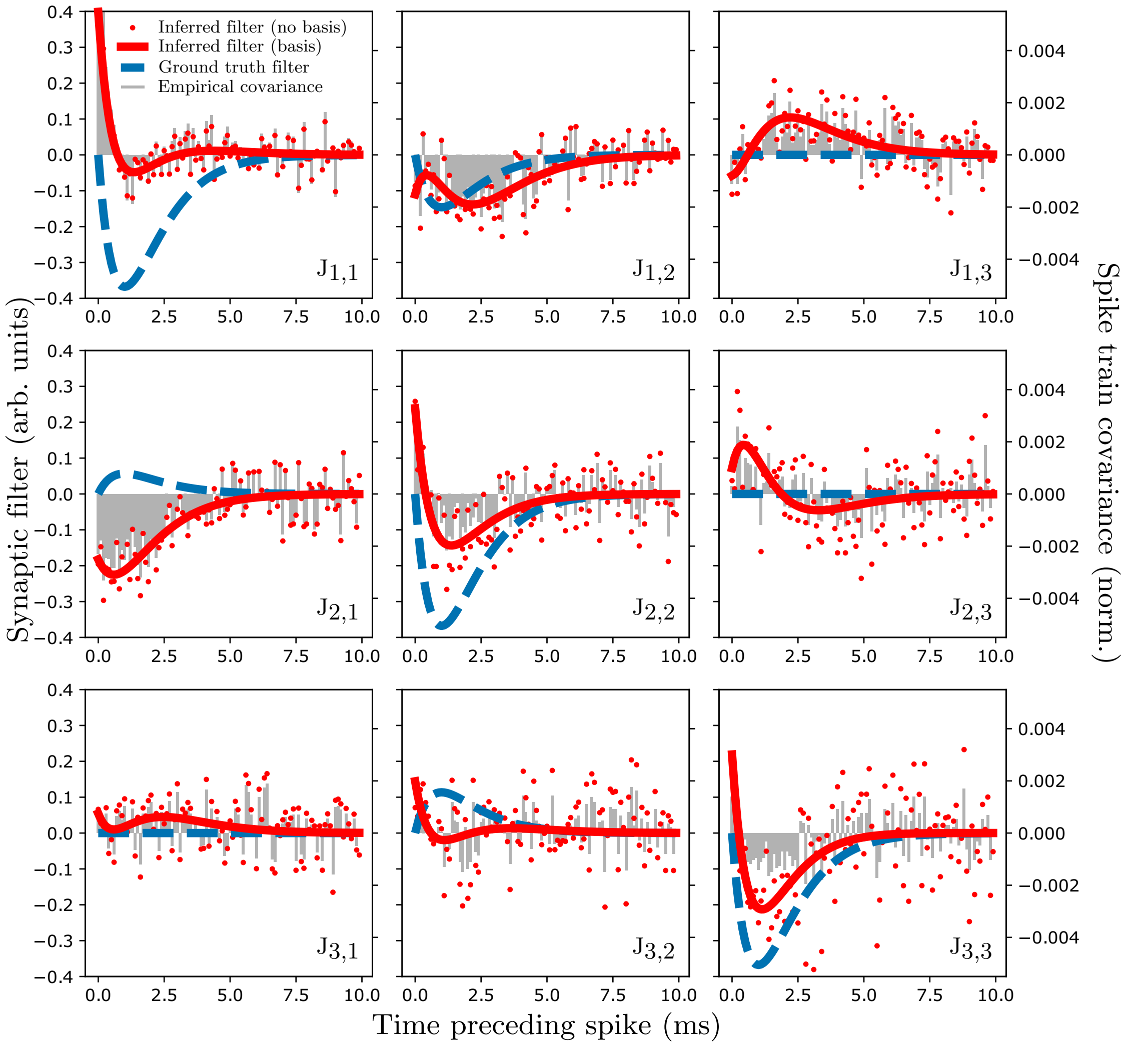

FIG. 2. Maximum-likelihood estimates (MLE) of coupling filters J_{ij} for $N_{\rm obs} = 3$ observed neurons out of 64 in a circuit with 50% sparsity and Gaussian weights. The ground-truth synaptic filters of the generative model are shown in blue dashed lines (values on left axis). Two types of inferred filters are shown: for each time point (red points) and using a basis expansion (red solid line). We also compare the filters to the empirical covariances, shown in grey bars (scaled up to be visible on the same plots as the filters; true values on right axes). The covariances correlate strongly with the filters inferred without using a basis expansion.

spread. This spread is due to two factors: the strength of the synaptic connections in these simulations and the finite-time estimates of the covariances (e.g., estimates of the covariance $C_{3,2}(t)$ in Fig. 2 are noisy and of either sign, while the synaptic connection is strictly positive, which will lead to a small estimate of the correlation coefficient). We explain later why strong connections will bias covariances away from the ground-truth filters.

The spike train covariances and ground-truth filters do not depend on the number of observed neurons, whereas the inferred filters do change with the number of neurons used in maximum likelihood estimation. In Fig. 5 we show the distribution of the correlation coefficients between covariances and inferred filters (estimated pointwise, not using basis functions) as a function of the fraction of neurons observed, for both sparse random networks [Fig. 5(a)] and E-I networks [Fig. 5(b)]. For both network types we find that the correlation between autocovariances and autocoupling filters is generally very close to 1 when $\lesssim 10\%$ of the network is observed, dropping to a median of ∼0.7 when the network is fully observed. The crosscovariances between different neurons are even more strongly correlated with their corresponding inferred crosscoupling filters.

We compare the covariance-inferred filter correlations with the correlations between the inferred filters and the nonzero ground-truth filters; for $J_{ij}(t) = 0$ the correlation is not defined. As shown in Fig. 6, we find that the correlations are strong when the network is fully observed, as expected, and decrease as fewer neurons are observed. In random networks we observe a considerable spread in the distribution of the covariances for the crosscouplings between neurons, with some inferred filters even being anticorrelated with the ground-truth filter (see, e.g., $J_{2,1}$ in Fig. 2)—this may be due to the relatively large distribution of synaptic weights in these networks. Variability in the E-I crosscouplings is due only to sparsity, and we observe smaller correlation distributions in these

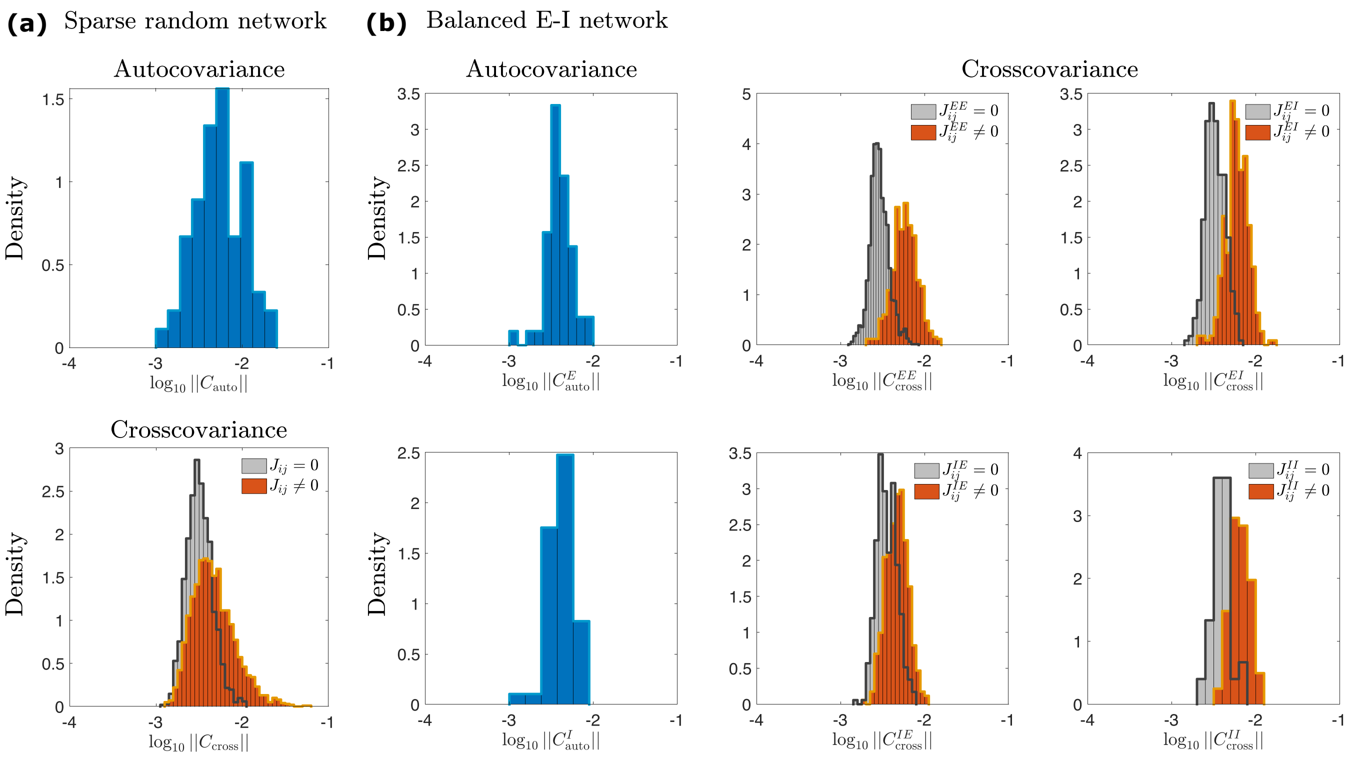

FIG. 3. Directed magnitudes $||C_{ij}|| = \sqrt{\int_{0^+}^{\infty} dt\, C_{ij}(t)^2}$ of the empirical spike train covariances estimated with the simulated spike trains. (a) (Log-)Magnitudes for a sparse random network. Top: Directed magnitude of the positive-lag of the autocovariance with zero-lag excluded; bottom: directed magnitudes of the crosscovariances. (b) (Log-)Magnitudes for a balanced excitatory-inhibitory network. Left-most column: directed magnitudes for the autocovariances with zero-lag excluded, separated by cell type (51 excitatory, 13 inhibitory). Middle- and right-most columns: directed magnitudes for the crosscovariances, separated by which pairs of cell types are connected. For the crosscovariances the histograms are separated also by whether the ground-truth connection is zero or not, showing that there is overlap between the covariances of connected and unconnected pairs of neurons.

networks. There is no variability in the autocouplings of either network.

The results of Figs. 5 and 6 are consistent with the covariance-ground-truth correlations shown in Fig. 4. For example, in random networks the strong correlations between covariances and inferred filters combined with the occasional anticorrelation between inferred filters and ground truth translates into the occasional anticorrelation between covariances and ground truth.

Our results therefore suggest that when the number of recorded neurons $N_{\rm obs}$ is a large fraction of the total number of neurons N in the network, the inferred filters correlate strongly with the ground-truth filters but less so with the spike train covariances. However, when the fraction of observed neurons is small, $N_{\rm obs}/N \ll 1$, the inferred filters correlate strongly with the positive-lag half of the covariances, but less so with the ground-truth filters.

3. Correlations become stronger in weakly coupled networks

So far we have shown results for strongly connected networks with synaptic weights just below the critical value for which the network activity would become unstable. We also investigated how the strength of these correlations depends on the synaptic weights. For random Gaussian networks we varied the standard deviation of the weights, $J_0 \in \{1, 2, 3\}$, while for balanced E-I networks we varied both excitatory and inhibitory weights by a multiplicative factor $J_0 \in \{1, 4, 7\}$. The results shown in Fig. 5 correspond to the strongest synaptic strengths we investigated, namely, $J_0 = 3$ for the random networks and $J_0 = 7$ for the E-I networks. We find that correlations between covariances grow even stronger when the synaptic connections are weaker, shown in Fig. 7. As shown in the right panel of Fig. 7(a), in the weak coupling limit the Pearson correlations between the MLE inferred filters and the spike train covariances are high even if all the neurons in the network are observed. Smaller J_0 confined the network to a noise-driven regime, where each neuron's firing rate is dominated by the same baseline drive μ set in the generative model, and a higher J_0 tunes the network into a strong coupling regime with more variable firing rates across neurons.

Because our simulation results demonstrate a high degree of correlation between the inferred synaptic filters (without using basis functions) and the empirically estimated spike train covariances, some property of the network statistics or maximum likelihood inference procedure must give rise to these strong correlations when the network is subsampled. To better understand this relationship, we turn to an analytic analysis of the maximum likelihood estimation of subsampled networks.

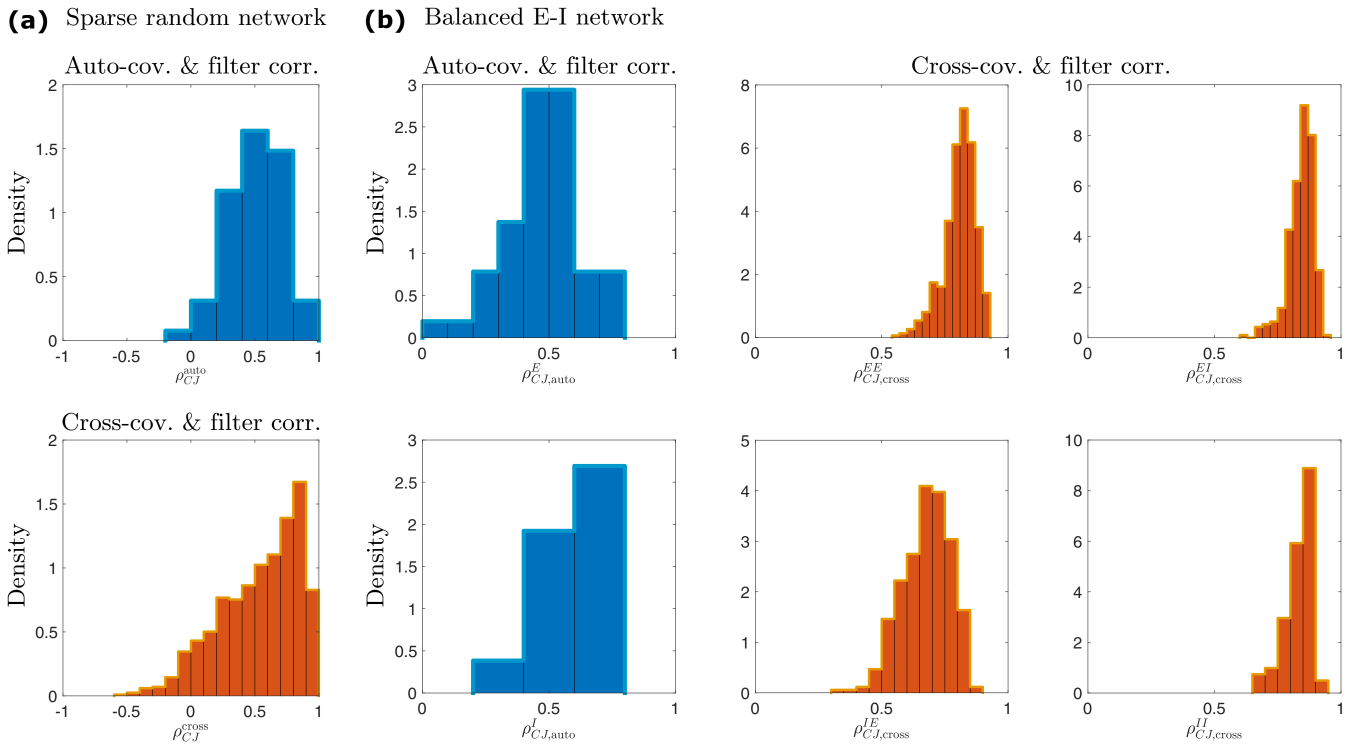

FIG. 4. Pearson correlation coefficient between the empirical spike train covariances and ground-truth synaptic filters. (a) Correlations for a sparse random network. Top: correlation between the positive-lag of the autocovariance and the autocoupling (self-history) filter; bottom: correlation between the crosscovariances and the crosscoupling filters between pairs of neurons. (b) Correlations for a balanced excitatory-inhibitory network. Left-most column: correlations between autocovariances and autocoupling, separated by cell type (51 excitatory, 13 inhibitory). Middle- and right-most columns: correlations between the crosscovariances and crosscouplings, separated by which pairs of cell types are connected. For the crosscovariances/couplings correlations are only computed for synaptic connections which are nonzero.

C. Theoretical analyses

To understand the observed correlations between spike train covariances, the underlying ground-truth synaptic connections, and the inferred synaptic filters, we analyze the spiking network models using a mean-field approximation with Gaussian fluctuation corrections. We analytically investigate the maximum-likelihood equations in the continuous time limit and the infinite data limit $T \to \infty$.

1. Mean-field approximation with Gaussian fluctuation corrections

We use a self-consistent mean-field approximation to estimate the firing rates, which consists of neglecting fluctuations in activity:

$$\begin{aligned} r_i \equiv \langle \dot{n}_i(t) \rangle &= \left\langle \phi\left(\mu_i + \sum_{j=1}^{N} \int dt'\, J_{ij}(t-t')\dot{n}_j(t') \right) \right\rangle \\ &\approx \phi\left(\mu_i + \sum_{j=1}^{N} \int dt'\, J_{ij}(t-t')\langle \dot{n}_j(t') \rangle \right) \\ &= \phi\left(\mu_i + \sum_{j=1}^{N} \mathcal{J}_{ij} r_j \right), \end{aligned} \tag{6}$$

where we assume the networks reach a time-independent steady state, $r_i = \langle \dot{n}_i(t) \rangle$. We then approximate the spiking process as Gaussian fluctuations around the mean-field estimates and calculate the covariances using the path integral formalism developed in Ref. [10]; see Appendix D for details. Within this approximation the covariances may be written

$$C_{ij}(t-t') = \sum_{k=1}^{N} \int_{-\infty}^{\infty} dt''\, \Delta_{ik}(t-t'')\Delta_{jk}(t'-t'')r_k, \tag{7}$$

where $\Delta_{ij}(t-t')$ are the linear response functions of the network: the average response of neuron i at time t to a forced spike from neuron j at time t'. The response functions are causal: $\Delta_{ij}(t-t') = 0$ for all $t - t' < 0$, reflecting the fact that neurons cannot respond to perturbations that occur in the future. While the response functions contain information about causality in the network, they are more difficult to measure because they cannot be directly inferred from ongoing activity, only from trial-averages of responses to forced-spike perturbations. We therefore focus on the covariance, and the extent to which it reflects the underlying causality of the synaptic connections $J_{ij}(t)$.

It is instructive to expand the covariance in a series for weak synaptic connections. In this regime the covariance may

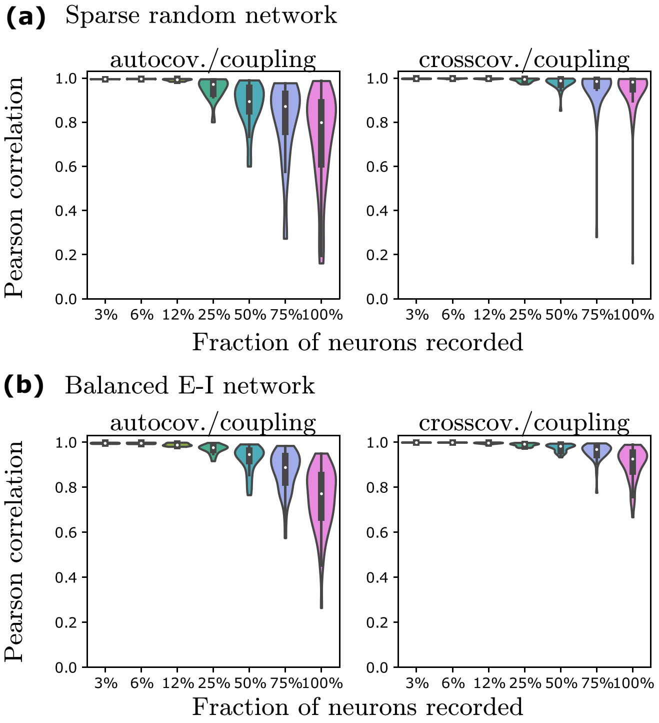

FIG. 5. Pearson correlation between the spike train covariances and inferred coupling filters in strongly coupled random networks and balanced E-I networks of 64 neurons. (a) Violin plots show how the correlations change when different fractions of neurons are observed in a network with sparse Gaussian synaptic connections, 3%, 6%, 12%, 25%, 50%, 75%, and 100%. In this 64 neuron network these fractions corresponds to 2, 4, 8, 16, 32, 48, and 64 observed neurons. The spike train covariance functions strongly correlate with the inferred filters inferred for each time point when fewer neurons are observed and the correlation decreases as more neurons are observed. The left panel shows the distribution of the Pearson correlation coefficients between the autocovariance and autocoupling (self-history) filters, while the right panel shows the distribution of the Pearson correlation coefficients between the crosscovariance and coupling filters between neuron pairs. (b) Same as panel (a) but for the balanced E-I networks.

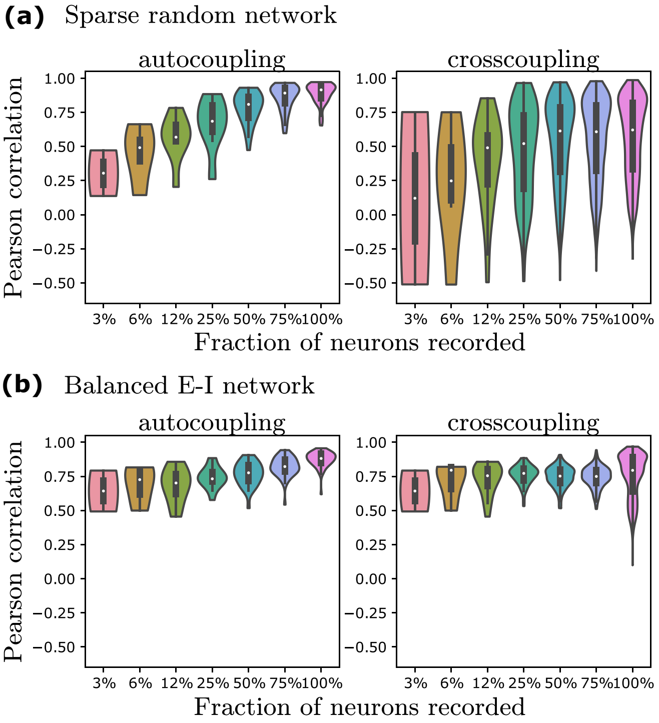

FIG. 6. Pearson correlation between the inferred spike train filters and the ground-truth filters in strongly coupled random networks and balanced E-I networks of 64 neurons. (a) For sparse random networks. (b) For balanced E-I networks. Plot details are the same as in Fig. 5. The correlation between ground-truth and inferred filters are highest when the network is fully observed, and are relatively high for the E-I network, but can be negative in the random network. The left panels show the distribution of the Pearson correlation coefficients between the autocouplings (self-connections implementing the soft reset), while the right panels show the distribution of correlation coefficients between the crosscouplings between neuron pairs.

be written (see Appendix F)

$$
\begin{aligned}
C_{ij}(t-t') \approx\ & r_i\delta_{ij}\delta(t-t') + g_iJ_{ij}(t-t')r_j + g_jJ_{ji}(t'-t)r_i \\
& + \sum_{\ell=1}^{N}\int_{-\infty}^{\infty} dt''\, [g_jJ_{j\ell}(t'-t'')g_\ell J_{\ell i}(t''-t)r_i \\
& + g_iJ_{i\ell}(t-t'')g_\ell J_{\ell j}(t''-t')r_j \\
& + g_iJ_{i\ell}(t-t'')g_jJ_{j\ell}(t'-t'')r_\ell], \qquad (8)
\end{aligned}
$$

where $g_i \equiv \phi'(\mu_i + \mathcal{J}_{ij}r_j)$ is gain of neuron i, equal to the firing rate r_i when $\phi(\cdot) = \exp(\cdot)$. We thus see that for sufficiently weak coupling the synaptic filters $J_{ij}(t)$ and $J_{ji}(t)$ can be estimated from the positive- and negative-lag halves of the covariances, respectively, up to constant factors.

For stronger synaptic connections feedback with the rest of the network, represented by the higher-order terms in the expansion (8), cannot be neglected. The shape of the covariances between pairs of neurons will be skewed away from the ground-truth synaptic filters in the strong coupling regime. Moreover, the last term in Eq. (8), $\sum_\ell \int dt''\, g_iJ_{i\ell}(t-t'')g_jJ_{j\ell}(t'-t'')r_\ell$, is not causal, as it is the contribution to the covariance of neurons i and j due to synaptic input from other neurons ℓ.

To investigate the covariances at strong coupling, for both the sparse random network and the E-I network, we solve for the mean-field firing rates and linear response functions numerically, from which we estimate the covariances using Eq. (7). The accuracy of the mean-field approximation is expected to be quantitatively inaccurate but remain qualitatively accurate as the coupling strength increases [11]. We find, as shown in Fig. 9, that many spike train covariances remain strongly correlated with the ground-truth filters, but many pairs display weaker correlations. Weak correlations are particularly notable between the autocoupling filters and the autocovariances.

Our mean-field analysis of the networks indeed yields qualitatively similar results to the simulations, though simulations show a wider spread of correlation values between the empirical covariances and the ground-truth filters. This is likely due to non-Gaussian contributions to fluctuations in spiking activity that are not negligible in strongly coupled networks.

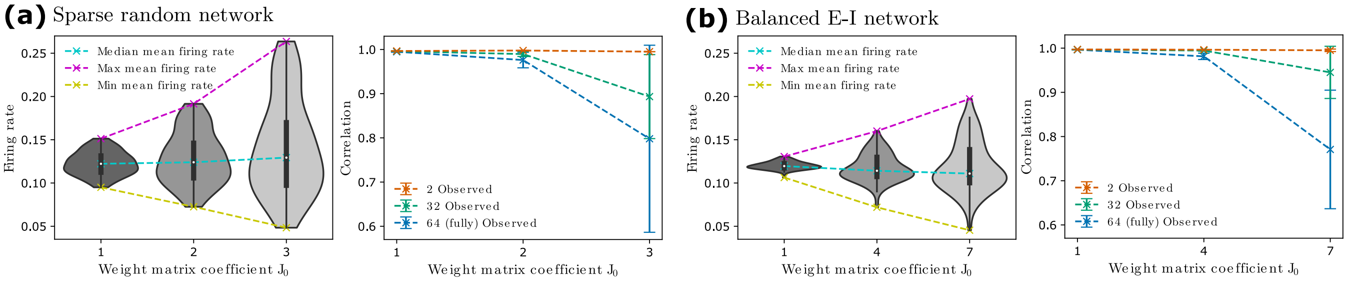

FIG. 7. Correlations are high for both random and balanced excitatory-inhibitory networks in weak coupling regimes. (a) Random network. When the weight matrix coefficient J_0 decreased from 3 to 1, the network transitioned into a noise-driven regime, and the mean firing rates of the neurons varied less and were driven by the same baseline drive in the generative model, as shown in the left panel. As shown in the right panel, in the weak coupling regime, such as when $J_0 = 1$, high correlations between the MLE inferred filters and spike train covariances were observed even when all the neurons in the network are observed, as compared to the network in a strong coupling regime where the high correlation between the MLE inferred filters and spike train covariances only happen in the sub-sampled network. (b) Similar results hold for a balanced excitatory-inhibitory network, where 20% of the neurons are inhibitory and 80% of them are excitatory. The weight matrix coefficient J_0 is tuned from 1 to 7, with $J_0 = 7$ the largest possible integer value for which the spiking process is still stable.

Nevertheless, the mean-field approximation remains qualitatively accurate, and we investigate the inferred filters within the scope of this approximation.

2. *Maximum-likelihood estimation equations*

Next, to understand the correlations with the inferred filters, we need to derive the maximum likelihood equations within this mean-field approximation. Because we are focusing on ongoing network activity, in this limit the integrals over time in Eqs. (4) and (5) become expectations over the spiking process:

$$\langle \dot{n}_i(t) \rangle = \langle \hat{\phi}_i(t) \rangle, \tag{9}$$

$$\langle \dot{n}_i(t)\dot{n}_j(t') \rangle = \langle \hat{\phi}_i(t)\dot{n}_j(t') \rangle, \tag{10}$$

where the angled brackets denote an expectation over spike trains of the *generative* circuit model, not the inference model.

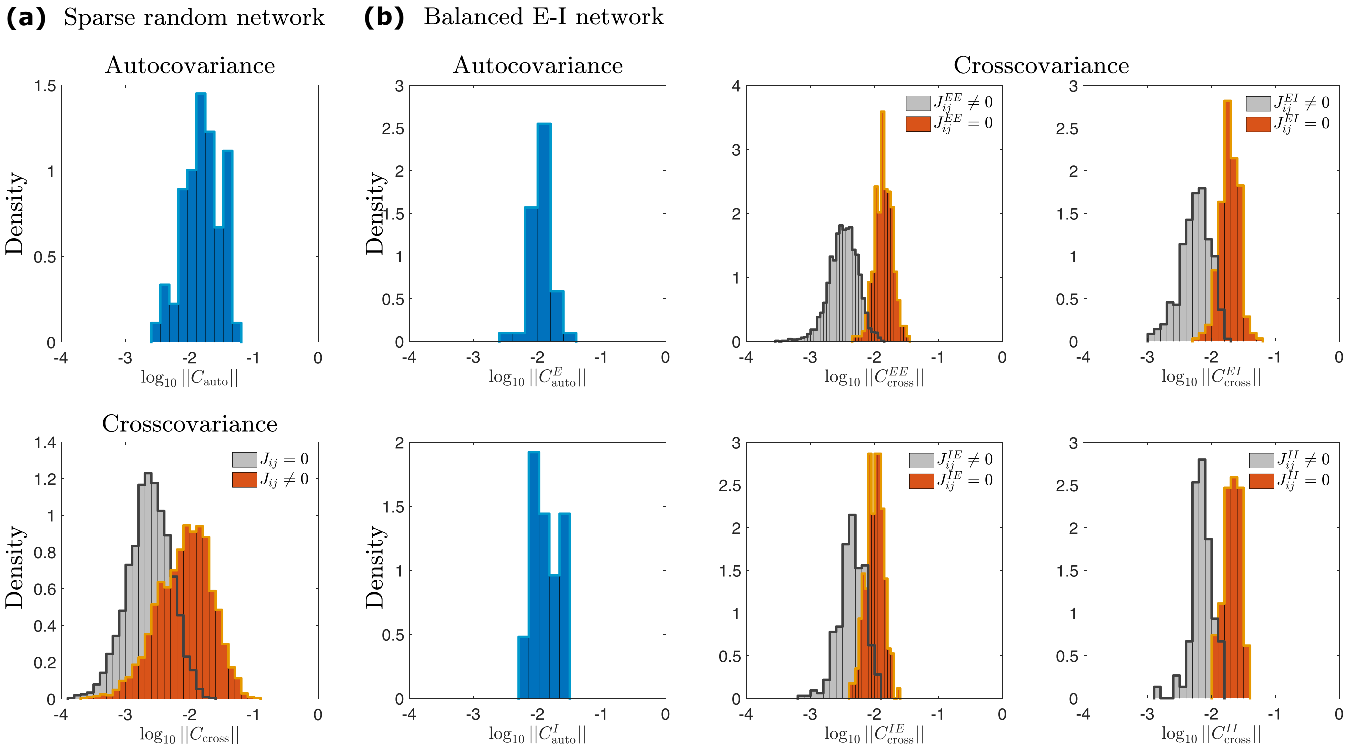

FIG. 8. Directed magnitudes $||C_{ij}|| = \sqrt{\int_{0^+}^{\infty} dt\, C_{ij}(t)^2}$ of the spike train covariances calculated using mean-field theory. (a) (Log-)Magnitudes for a sparse random network. Top: Autocovariance magnitudes; bottom: crosscovariance magnitudes. (b) (Log-)Magnitudes for a balanced excitatory-inhibitory network. Left-most column: Autocovariance magnitudes separated by cell type. Middle- and right-most columns: crosscovariance magnitudes separated by which pairs of cell types are connected. For the crosscovariances the histograms are separated also by whether the ground-truth connection is zero or not, showing that there is overlap between the covariances of connected and unconnected pairs of neurons.

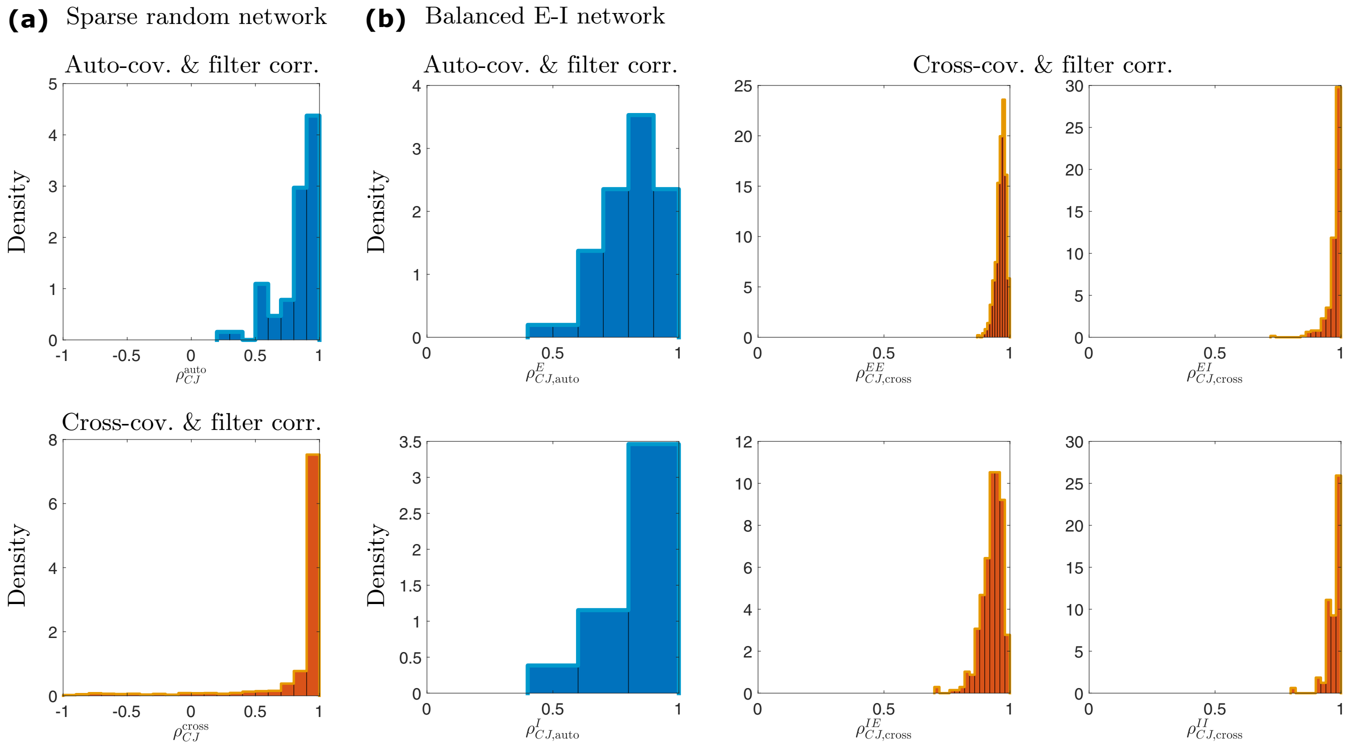

FIG. 9. Pearson correlation coefficient between the mean-field spike train covariance estimates and ground-truth synaptic filters. (a) Correlations for a sparse random network. Top: correlation between the positive-lag of the autocovariance and the autocoupling (self-history filter); bottom: correlation between the crosscovariances and the crosscoupling filters between pairs of neurons. (b) Correlations for a balanced excitatory-inhibitory network. Left-most column: correlations between autocovariances and autocoupling, separated by cell type. Middle- and right-most columns: correlations between the crosscovariances and crosscouplings, separated by which pairs of cell types are connected. For the crosscovariances/couplings correlations are only computed for synaptic connections which are nonzero in ground truth.

In a stationary steady-state Eq. (9) will be independent of time t and Eq. (10) will depend only on the time difference $t - t'$.

We exploit the choice of the exponential nonlinearity to relate the expectations in Eqs. (9) and (10) to the moment generating functional of the spike train process:

$$\begin{aligned}\langle\hat{\phi}_i(t)\rangle &= \lambda_0\left\langle e^{\hat{\mu}_i+\sum_j\int dt'\,\hat{J}_{ij}(t-t')\dot{n}_j(t')}\right\rangle\\ &= \lambda_0 e^{\hat{\mu}_i} Z[\tilde{j}_j(t') = \hat{J}_{ij}(t-t')],\end{aligned} \tag{11}$$

$$\langle\hat{\phi}_i(t)\dot{n}_j(t')\rangle = \lambda_0 e^{\hat{\mu}_i}\left.\frac{\delta Z[\tilde{j}]}{\delta\tilde{j}_j(t')}\right|_{\tilde{j}_j(t')=\hat{J}_{ij}(t-t')}, \tag{12}$$

where Z is the moment generating functional of the fluctuations, defined for an arbitrary "source" variable $\tilde{j}_i(t)$ as $Z[\tilde{j}] \equiv \langle\exp(\sum_i\int dt\,\tilde{j}_i(t)\dot{n}_i(t))\rangle$. The moment-generating functional for the spiking network cannot be evaluated exactly, hence the need for approximations. Within the mean-field and Gaussian fluctuation approximation the moment-generating functional can be evaluated explicitly [19]:

$$Z[\tilde{j}] = \exp\left(\sum_{i=1}^{N}\int_{-\infty}^{\infty} dt\,\tilde{j}_i(t)r_i + \frac{1}{2}\sum_{ij}\int dtdt'\,\tilde{j}_i(t)C_{ij}(t-t')\tilde{j}_i(t')\right). \tag{13}$$

An important implication of Eqs. (11) and (12) is that the moment generating functional contains only information about the noncausal statistical *moments* of the spike train process; they do not directly contain any information about the causal response functions. In a path integral formulation of this stochastic process, one can formulate a more general moment generating functional that contains information about both the statistical moments and the causal response functions of the process [19]. Crucially, however, the information about the response functions drops out of the expectations we have computed, meaning that the MLE equations do not directly contain any information about the response functions. The response functions enter only through their relationships to the covariances. In fully observed systems in steady-states it is often possible to derive fluctuation-dissipation relationships that linearly relate statistical moments and response functions [20]. While our results suggest that such a relationship may enable accurate inference of the ground-truth connections in a fully observed circuit, inference from subsets of neurons cannot recover information about circuit response functions, and the inferred connections may reflect only covariances in neural activity.

Evaluating Eqs. (9) and (10) using the Gaussian fluctuation approximation and using Eq. (11) to eliminate the dependence on $\hat{\mu}_i$ in Eq. (12) results in a system of Wiener-Hopf integral equations to solve for the filters $\hat{J}_{ij}(t)$:

$$C_{rr'}(t) = r_r \sum_{r''\in \text{obs.}} \int_{0^+}^{\infty} dt''\, \hat{J}_{rr''}(t'') C_{r''r'}(t-t''), \tag{14}$$

where $t > 0$ and the sum is only over the N_{obs} observed neurons. Although we expect $\hat{J}_{rr'}(t'') = 0$ for $t'' < 0$, which would let us extend the range of integration to the whole real line, the restriction to $t > 0$ prevents a straightforward solution via Fourier transform. If this restriction is neglected, then the resulting solutions $\hat{J}_{rr'}(t)$ may be noncausal. Analytically solving this system of equations while imposing causality is difficult and an area of active research [21], so here we pursue two analyses: an approximation via Picard iteration and a simple case that admits an analytic solution.

First, we use the fact that the spike train covariances can be separated into singular and regular terms, $C_{rr'}(t) = r_r\delta_{rr'}\delta(t) + \overline{C}_{rr'}(t)$, where the δ-function term is due to the conditionally Poisson nature of the spike trains. We can use this to rewrite the integral equation as

$$\hat{\mathbf{J}}(t) = \mathbf{R}^{-1}\overline{\mathbf{C}}(t)\mathbf{R}^{-1} - \int_{0^+}^{\infty} dt''\, \hat{\mathbf{J}}(t'')\overline{\mathbf{C}}(t-t'')\mathbf{R}^{-1}, \tag{15}$$

where we have written this using a matrix notation, where $\mathbf{R} = \text{diag}(\mathbf{r})$ is a diagonal matrix of the N_{obs} observed firing rates, and $\hat{\mathbf{J}}(t)$ and $\overline{\mathbf{C}}(t)$ are also $N_{\text{obs}} \times N_{\text{obs}}$ matrices. Equation (15) lends itself well to a numerical solution via Picard iteration, which consists of approximating $\hat{\mathbf{J}}(t)$ by an initial guess and updating this guess by repeatedly plugging it into the right hand side of Eq. (15) until it converges to within some numerical tolerance. We use this procedure to numerically estimate the inferred filters; we can also verify that when $N_{\text{obs}} = N$ the ground-truth filters satisfy the equation. We may then compute the correlations between the ground-truth filters, spike train covariances, and inferred filters within the mean-field approximation. We quantify the correlation by the overlap between functions. The overlap between two functions $A(t)$ and $B(t)$ is defined as

$$\rho_{AB} \equiv \frac{\int_{0^+}^{\infty} dt\, A(t)B(t)}{\sqrt{\int_{0^+}^{\infty} dt\, A(t)^2 \int_{0^+}^{\infty} dt\, B(t)^2}}. \tag{16}$$

The overlap is equivalent to the Pearson correlation coefficient estimated from infinite time-points. We can compute the overlaps between pairs of C, J, and $\hat{J}$ by numerically evaluating the integrals. The trends, shown in Figs. 10 and 11, qualitatively agree with the full simulations.

While Eq. (15) is useful for estimating the inferred filters within the mean-field approximation, the relationship to the ground-truth filters remains opaque. To this end, it is again useful to expand the covariance in a Taylor series in powers of $r_iJ_{ij}(t)$, similar to the investigation of subsampled inference in the weak coupling limit by Ref. [5]. We find, to quadratic order,

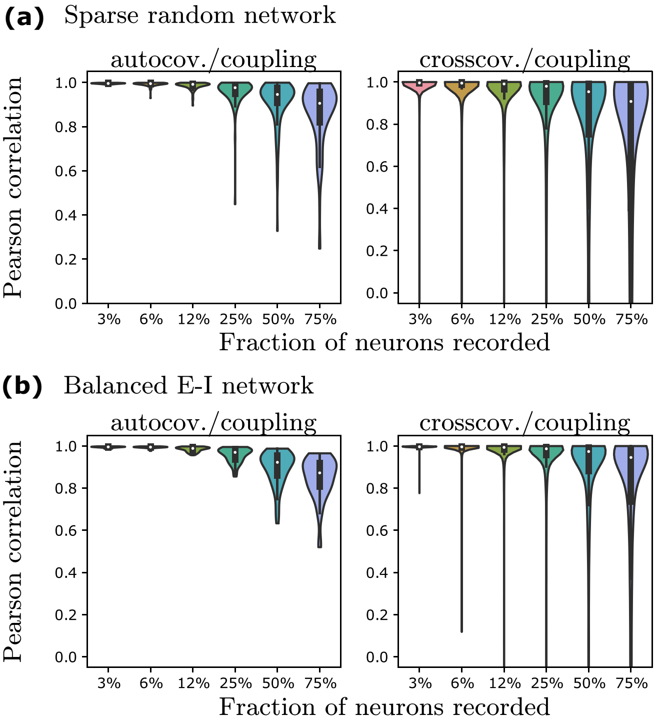

FIG. 10. Pearson correlation between the mean-field theory predictions of the spike train covariances and inferred coupling filters in strongly coupled random networks and balanced E-I networks of 64 neurons. (a) Results for sparse random networks. (b) Results for balanced E-I networks. The left panels show the correlation coefficients between the autocovariance and autocoupling (self-history) filters, while the right panels show the distribution of correlation coefficients between the crosscovariance and coupling filters. For both types of networks we use the same synaptic weight matrices as the simulations shown in Fig. 5. We estimate the covariances using Eq. (7) and the inferred filters using Eq. (15). Violin plots show how the correlations change when different numbers of neurons are observed in a network with sparse Gaussian synaptic connections; the percentages shown for this 64 neuron network correspond to 2, 4, 8, 16, 32, and 48, observed. For small N_{obs} several different pairs are randomly sampled. In the fully observed case, $N_{\text{obs}} = 64$, the inferred and ground-truth filters are perfectly correlated, and the corresponding correlation is shown in Fig. 9. The mean-field estimates qualitatively reproduce the trends seen in simulations, shown in Fig. 5, displaying large correlations when only a small fraction of the network is observed. The vertical axis has been truncated to [0,1] to show details more clearly, cutting off the long tails of the crosscovariance distributions, which extend close to -1. i.e., the mean-field theory estimates that several inferred filters are anticorrelated with the spike train covariances.

$$\hat{J}_{rr'}(t-t') \approx J_{rr'}(t-t') + \sum_{h\in\text{hidden}} \int_{-\infty}^{\infty} dt''\, [J_{rh}(t-t'')r_h J_{hr'}(t''-t') + J_{rh}(t-t'')J_{r'h}(t'-t'')r_h] + \dots, \tag{17}$$

where the sum is only over unobserved (or "hidden") neurons and $t - t' > 0$; $\hat{J}_{rr'}(t-t')$ is zero for negative arguments. Note that when all neurons are recorded–i.e., there are no hidden neurons—the nonlinear correction term vanishes. We

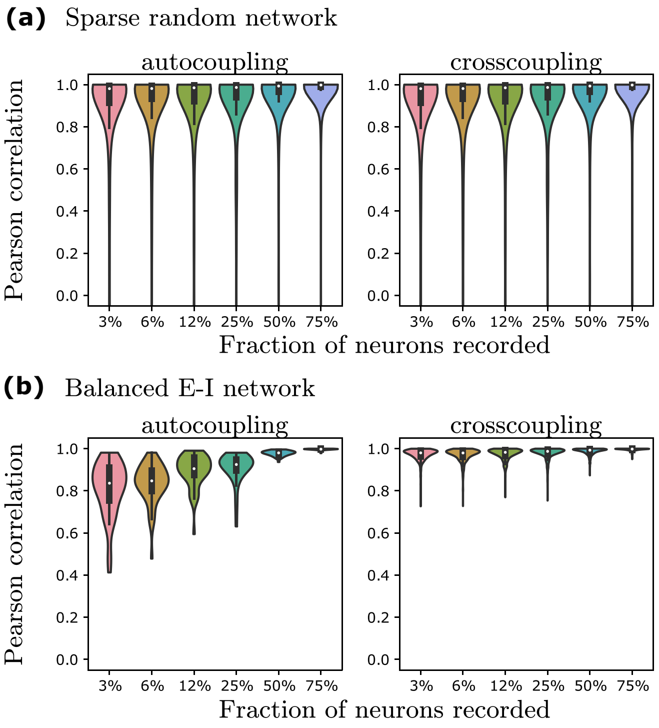

FIG. 11. Pearson correlation between the mean-field estimates of the inferred spike train filters and the ground-truth filters in strongly coupled random networks and balanced E-I networks of 64 neurons. (a) Results for sparse random networks. (b) Results for balanced E-I networks. The left panels show the correlation coefficients between the autocovariance and autocoupling (self-history) filters, while the right panels show the distribution of correlation coefficients between the crosscovariance and coupling filters. Other details are the same as Fig. 10. The results qualitatively reproduce the trends seen in simulations, shown in Fig. 6, albeit with a quantitatively weaker agreement, particularly for the sparse random network.

expect this to hold at all orders. This form of the solution highlights how the unobserved neurons distort the inferred filters: the inferred filters between recorded neurons must capture the variability caused by the hidden neurons. The first correction term in Eq. (17) captures the effects of a recorded neuron r' driving a hidden neuron h, which in turn drives the post-synaptic recorded neuron labeled r. The second correction term describes the activity of the two recorded neurons r and r' driven by common input from an unobserved neuron h. It is terms like this in particular that confound attempts to estimate causality from covariances, as $\int dt''\, J_{rh}(t-t'')J_{r''h}(t'-t'')$ is nonzero regardless of the sign of $t-t'$. Higher-order terms will involve similar contributions involving more hidden neurons.

When the synaptic connections or covariances are strong the iterative scheme considered above may take many iterations to converge, making it susceptible to numerical error. To gain more insight into the inferred filters, we turn to an analytically tractable special case.

3. Homogeneous networks

We consider a homogeneous network of all-to-all coupled neurons, which displays the key features observed in our simulations. We choose $J_{ij}(t) = Jt\exp(-t/\tau)\Theta(t)/\tau^2$, including the autocouplings $i = j$, and homogeneous baselines $\mu_i = \mu$. The self-consistently calculated mean firing rates can be solved in terms of the Lambert W function, $r = -(NJ)^{-1}W_{-1}(-NJ\lambda_0 e^{\mu})$, defined as the solution of the transcendental equation $z = W_{-1}(z)\exp(W_{-1}(z))$ for which $-1/e < z < 0$. This restriction defines the branch of the Lambert W function that we must use for excitatory $J > 0$. It follows that the network is only stable for $NJ\lambda_0 \exp(1+\mu) < 1$.

The mean-field approximation of the spike train covariance is

$$C_{ij}(t-t')$$
$$= r\left[\delta_{ij}\delta(t-t') + \frac{(a_-^2(1)-b_+^2)(b_+^2-a_+^2(1))}{(b_+-b_-)(b_++b_-)}\frac{e^{-b_+|t-t'|/\tau}}{2b_+\tau}\right.$$
$$\left. - \frac{(a_-^2(1)-b_-^2)(b_-^2-a_+^2(1))}{(b_+-b_-)(b_++b_-)}\frac{e^{-b_-|t-t'|/\tau}}{2b_-\tau}\right], \qquad (18)$$

where $a_\pm(n) = \sqrt{1+(N-n)Jr \pm \sqrt{(N-n)Jr(4-Jr)}}$ and $b_\pm = 1 \pm \sqrt{NJr}$. We plot $C(t>0)$ for several values of $NJ\lambda_0 e^{1+\mu} < 1$ in Fig. 12, where we see that the decay time of the covariances is much slower than the ground truth $J(t)$ as $NJ\lambda_0 e^{1+\mu} \to 1^-$. Because all synaptic filters are the same in this special case, we may assume the inferred filters will all be equal as well, which reduces Eq. (14) to a single equation. We can then use the scalar Wiener-Hopf procedure [21] to solve Eq. (14) for the inferred filters of N_{obs} observed neurons. We find

$$\hat{J}(t) = \frac{1}{r}[A_+(N_{\text{obs}})e^{-a_+(N_{\text{obs}})t/\tau}$$
$$- A_-(N_{\text{obs}})e^{-a_-(N_{\text{obs}})t/\tau}]\frac{\Theta(t)}{\tau}, \qquad (19)$$

where $A_\pm(N_{\text{obs}})$ depend on $a_\pm(N_{\text{obs}})$, $a_\pm(1)$, and $b_\pm$. Their full expressions are bulky and unenlightening, so we defer writing them out to Appendix I. For $N_{\text{obs}} = 1$ these coefficients reduce to $A_+(1) = \frac{(a_+-b_-)(a_+-b_+)}{a_+-a_-}$ and $A_-(1) = \frac{(a_--b_-)(a_--b_+)}{a_+-a_-}$. In the limit $N_{\text{obs}} \to N$ this filter recovers the ground-truth filter $J(t) = Jt\exp(-t/\tau)\Theta(t)/\tau^2$. The inferred filter differs from the effective filters of the model in which all but the recorded neuron are marginalized over [11], due to the fact that our inference model does not include an effective Gaussian noise generated by the fluctuations of the unobserved neurons' activity. Models with common Gaussian driving noise have been fit to data [22], but the fitting procedure is more involved than standard maximum likelihood inference, and we do not consider it here. In Fig. 12(a), we plot the filters for different numbers of observed neurons N_{obs}, observing that the amplitude and decay rate of the filters decreases as N_{obs} increases, for both fits to simulated data and our solutions of Eq. (14) (inset). The mean-field theory correctly captures the qualitative trend. In the homogeneous network the autocoupling (self-history) filters $\hat{J}_{ii}(t) \equiv \hat{J}_{\text{auto}}(t)$ and crosscoupling filters $\hat{J}_{i\neq j}(t) = \hat{J}_{\text{cross}}(t)$ are identical, though the inferred filters differ slightly due to finite data effects.

We can use Eq. (16) to compute the overlaps between pairs of C, J, and $\hat{J}$ analytically. The expressions are rather unwieldy, but can be expressed as a function of the combined

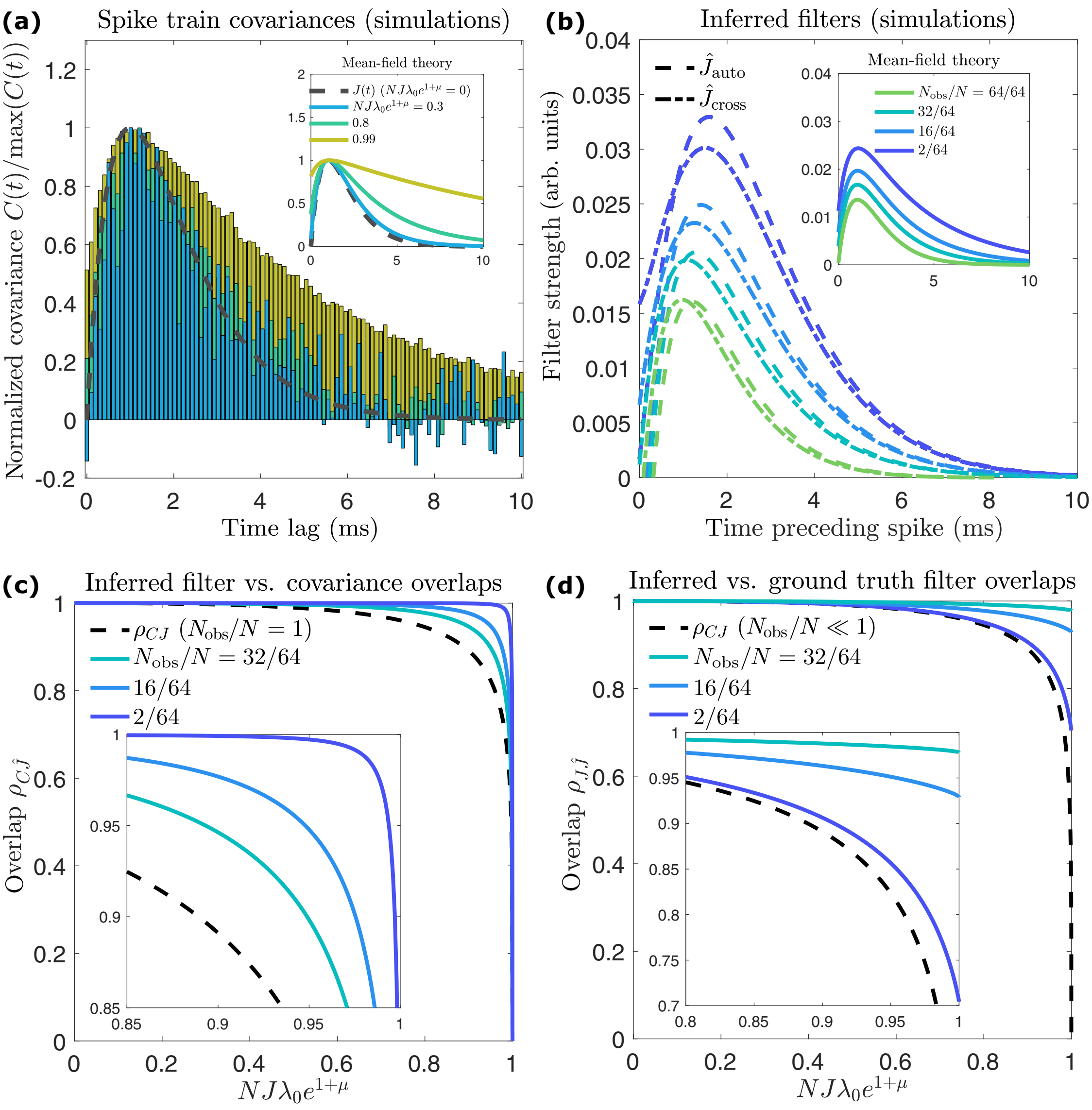

FIG. 12. Dependence of fits on the number of observed neurons in a homogeneous network. (a) The empirical covariance looks similar to the ground-truth filter, but shows clear differences in both simulations and the mean-field approximation (inset), especially for stronger networks. Simulation parameters: $N = 64$, $J = 0.00925$, 0.0185, 0.037, $\mu = -2$, $\lambda_0 = 1$, 4×10^6 time points. (b) For the strongest synaptic weight ($J = 0.037$), the inferred filters decrease in amplitude as the number of observed neurons $N_{\rm obs}$ increases to the total number of neurons in the network N, observed in fits to simulated data and our mean-field analysis (inset). In the ground-truth model the autocoupling (self-history) filters $\hat{J}_{\rm auto}(t)$ and crosscoupling filters $\hat{J}_{\rm cross}(t)$ are identical, as is the theoretical prediction from solving Eq. (14). The inferred auto- and crosscoupling filters differ due to finite data effects, but are close. (c) The predicted overlap $\rho_{C\hat{J}}$ drops as the synaptic weight J increases (i.e., $NJ\lambda_0 e^{1+\mu}$ increases), with varying speed depending on the ratio $N_{\rm obs}/N$. $N_{\rm obs}/N = 1$ corresponds to the fully observed case. The inset shows an enlarged view of the drop. Compare to Fig. 5. (d) The predicted overlap between inferred and ground-truth filters drops as the synaptic weight J increases (i.e., $NJ\lambda_0 e^{1+\mu}$ increases), approaching a finite limit of $\sim(N_{\rm obs}/N)^{1/4}$. The correlation between covariance and ground-truth filters, ρ_{CJ}, is included for comparison in both panels (c) and (d), being a limiting boundary in both cases.

parameter $NJ\lambda_0 e^{1+\mu} \leqslant 1$, where 1 is the edge of stability of the network, and the ratio $N_{\rm obs}/N$. We plot the overlaps between the covariance C and inferred filters $\hat{J}$ in Fig. 12(c), and the overlap between the inferred and ground-truth filters in Fig. 12(d). We see that the overlap between the covariance and ground truth [dashed line in Figs. 12(c) and 12(d)] is always less than the overlap between the inferred and ground-truth filters, as well as the overlap between the covariance and the inferred filters. In fact, ρ_{CJ} bounds the other overlaps in opposite limits: $\lim_{N_{\rm obs}/N\to 0} \rho_{J\hat{J}} = \lim_{N_{\rm obs}/N\to 1} \rho_{C\hat{J}} = \rho_{CJ}$. This reflects the fact that in a fully observed network $\hat{J}(t) = J(t)$, but when $N_{\rm obs}/N$ is small $\hat{J}(t) \approx C(t)$. Generally, we observe that the overlaps are generally very high away from the edge of stability of the network, but drop as the edge of stability is approached, capturing the behavior observed in Fig. 7.

In the three cases the overlaps are weakest close to the edge of network stability at $NJ\lambda_0 e^{1+\mu} = 1$ and for small $N_{\rm obs}/N$. In these limits we can derive the leading order behavior of the three overlaps. The overlap between the covariances and the ground-truth filters is independent of $N_{\rm obs}$, and scales near the edge of stability as

$$\rho_{CJ} \sim 2^{5/4}(1 - NJ\lambda_0 e^{1+\mu})^{1/4}; \tag{20}$$

i.e., the overlap between the covariance and the ground-truth filter vanishes as the network approaches its stability limit. The overlap between the ground-truth and inferred filters is finite as $NJ\lambda_0 e^{1+\mu} \to 1^-$, but scales as

$$\rho_{J\hat{J}} \sim 2(N_{\rm obs}/N)^{1/4} \tag{21}$$

for $N_{\rm obs}/N \ll 1$. Finally, the overlap between the covariances and inferred filters scales as

$$\rho_{C\hat{J}} \simeq 2^{5/4}(N_{\rm obs}/N)^{-1/4}(1 - NJ\lambda_0 e^{1+\mu})^{1/4}. \tag{22}$$

The window $1 - NJ\lambda_0 e^{1+\mu}$ over which the correlation between $\hat{J}(t)$ and $C(t)$ drops from $\rho = 1$ to 0 is $\mathcal{O}(N_{\rm obs}/N)$. Thus, even when the synaptic strength is quite strong, a heavily subsampled network must be tuned extremely close to the edge of stability before the inferred filters and spike train covariances differ appreciably.

In summary, the covariances correlate strongly with the ground-truth filters when connections are weak, independent of the number of neurons observed, while the inferred filters correlate slightly more strongly, tending to a finite limit for strongly coupled neurons, albeit a limit that is small for small numbers of observed neurons.

III. DISCUSSION AND FUTURE DIRECTIONS

The homogeneous network yields valuable insight into what is occurring in our simulations of random networks and balanced excitatory-inhibitory networks. The inferred filters obtained by maximum likelihood estimation are shaped by network responses through the causal halves (positive lag) of the otherwise noncausal spike train covariances. The causal halves of the spike train covariances tend to correlate with nonzero ground-truth synaptic connection between neurons, but this correlation is weak when the synaptic connections are strong. While the inferred filters tend to correlate with the ground-truth filters, this correlation becomes weak when the network is heavily subsampled and synaptic connections are strong. For weaker synaptic connections the inferred filters correlate more strongly with the spike train covariances than the ground-truth synaptic filters. Altogether, our results suggest that in heavily subsampled and weakly coupled networks the synaptic connections inferred from spontaneous activity data may offer only modest advantages over the cross-covariances ("functional connections"). In strongly coupled networks, neither the spike train covariances nor the inferred filters may be fully reflective of the underlying ground-truth synaptic connections. Interestingly, our simulation results show a strong correlation in the finite-data fluctuations of the empirical covariances and inferred filters (Fig. 2), suggesting a stronger relationship than we have been able to show in our analytic and numerical analyses of the mean-field network models.

Mapping out the network structure and inferring connections between neuron pairs from the recorded spike train data are challenging tasks, and understanding the results one obtains requires careful consideration of the assumptions underlying the statistical models fit to data. For example, Refs. [23] and [24] showed that it is possible to reconstruct neuronal connections from spike train covariances in certain sparse networks. Our results provide some support to this possibility: if one neuron synapses onto another then the directed crosscovariance magnitude is typically larger compared to neuron pairs with no connections (Figs. 3 and 8), but there is considerable overlap in the distribution of magnitudes, so perfect identification of ground-truth connections will not generally be possible, especially in subsampled circuitry. While some efforts have been made to infer neuronal connections between observed and hidden neurons, these methods must often make unrealistic assumptions, like allowing acausal connections between the observed and the hidden neurons [25], the number of hidden neurons is less than the observed neurons [26,27], or require careful modeling of the hidden neuron populations [28].

Other data-driven methods have emerged for inferring putative causal flows of information in neural circuitry, such as Granger causality, information-theoretic measures, or novel sampling paradigms [2,4,7,29–33]. However, even causality tests like Granger causality may not identify true causal influences between neurons due to unobserved neurons [34]. Theoretical and simulation-based analyses like Refs. [5,13], and this work are needed to understand the limits of statistical inference on subsampled neural data.

There are several natural directions our work on understanding the impact of unobserved neurons on statistical inference may continue in. In addition to unobserved neurons causing model mismatch, there are many other sources of model mismatch that could be treated with our formalism. Simple extensions would include investigating the impact of nonexponential nonlinearities in the generative and inference models. When the generative model has a nonexponential nonlinearity $\phi(\cdot)$ we expect, based on the calculations detailed in the appendices, that the synaptic connections are reweighted by the ratio of the gains to the firing rates, g_i/r_i. When the inference model's nonlinearity is nonexponential the MLE equations become nonlinear, even within the mean-field approximation we make in this work. This allows for the possibility of multiple solutions, reflecting the fact that the log-likelihood function is not concave for most nonlinearities [35]. We outline the form of these equations in Appendix K.

Another realistic source of model mismatch is that the form of the generative process that produces the data and the inference model will not match in reality, though one hopes the inference model is a reasonable approximation of the generative process. It is therefore valuable to understand the impact that using models with different features will have on inference. For instance, recently Ref. [36] has developed a path-integral formalism for spiking network models with hard resets—i.e., after a spike the membrane potential (which can be interpreted as the argument of the nonlinearity, $\mu_i + \int dt'\, J_{ij}(t-t')\dot{n}_j(t')$) is reset to 0, regardless of its value before the spike. In the "soft reset model" we use in this work the neurons inject a negative current into themselves after they spike, independent of the value of the current membrane potential. Because the membrane potential dynamics are explicit in such a model, the impact of nonlinearities in the membrane potential dynamics could also be studied. If we use the generalized linear model as our inference model, then such changes to the generative model would enter our formalism through the mean-field estimates of the firing rates, gains, and covariances of the network model.

Finally, in this work we focused on analyzing maximum likelihood inference applied to spontaneous activity data. However, applications of this method to real data often involve stimulus-driven activity [8]. One might wonder whether such input drives would result in maximum likelihood estimates that reflect the network response functions directly, rather than covariances, and thereby capture true causal activity within a circuit. The answer is no if the stimulus is provided as an input over a long single trial, or if it is chosen randomly across multiple trials, as in that case the stimulus can be treated as an additional stochastic process, and we expect the maximum likelihood estimates to reflect stimulus-spike covariances, not response functions. However, if an intra-cortical perturbation is delivered to a circuit and repeated several times after the activity returns to its steady state, then it becomes possible to align data to these events and estimate network response functions. These perturbations can be explicitly included in the likelihood of the model and may enable inference of synaptic filters $\hat{J}_{ij}(t)$ that directly reflect the truly causal response functions of the network, and hence causal flows of information through neural circuitry. Responses to perturbations have been gaining traction in neuroscience [37–39], and extensions of the analyses presented here will be valuable in guiding this next phase of probing neural circuitry.

ACKNOWLEDGMENTS

We thank Il Memming Park, Yuan Zhao, and Matthew Dowling for helpful discussions, and Siddharth Paliwal for code to generate the E-I networks. The authors thank Stony Brook University for financial support for this work.

APPENDIX A: SPIKE TRAIN SIMULATION

In this work, we use a generalized linear model (GLM) to simulate the neuron spike trains. In the GLM generative model, neurons emit spikes probabilistically following a Poisson process, with the rate given by $\phi(\mu_i + \sum_j \int dt' \, J_{ij}(t - t')\dot{n}_j(t'))$. Discrete spikes are generated for a small observation window $dt = 0.1$ ms. A first-order alpha function $J_{ij}(t) = \mathcal{J}_{ij}\, t\, e^{-t/\tau}/\tau^2\Theta(t)$ is used as the ground-truth interaction filters that govern the interaction of neuron j's spike train history on neuron i's instantaneous firing rate. Causality is imposed through the Heaviside step function $\Theta(t) = 1$ if $t > 0$ and 0 otherwise. The weight matrix $\mathbf{J}$ with entry $\mathcal{J}_{ij}$ sets the interaction strength of the filters and $\tau = 1$ sets the typical timescale of the decay of the response. The spike train is simulated by solving a second-order differential equation with a fourth-order Runge-Kutta method to conveniently track the spike train history, following the method used in previous works [10,11,40]. While simulating the spike train data, we run the simulation up to 2 million observation windows for the random and E-I networks.

1. Random network weight matrix generation

Following previous work [10,11], we generated a 64-neuron random network weight matrix with a sparsity $p = 50\%$, so that around half of the connections are nonzero. The nonzero synaptic weight strengths were drawn independently from a normal distribution with zero mean and standard deviation $J_0/\sqrt{pN}$, where J_0 is the weight matrix coefficient and N is the number of neurons in the network. The weight matrix coefficient J_0 takes three values 1, 2, 3, while we set the baseline drive $\mu_i = -2$ throughout the study. $J_0 = 3$ is the largest integer value we can set to still have a stable spike train process in the simulation, thus the network is considered to be in a strong coupling regime in that case. Importantly, the diagonal entries of the weight matrix were always set to -1 for different J_0 to simulate a soft refractory period for the neurons from their own spike history.

2. Balanced E-I network weight matrix generation

Following previous works [17,18], we generated a 64-neuron excitatory-inhibitory (E-I) network weight matrix with 20% (13) inhibitory neurons and 80% (51) excitatory neurons. Excitatory neurons make connections to excitatory neurons with a probability 20% and all other neuron pairs (E-I, I-E, and I-I) make connections with a probability 50%. The weight matrix coefficient J_0 was set to be a multiplicative factor of the base weight matrix. The base weight of the excitatory to excitatory connections was set to 0.0875, while the weight of excitatory to inhibitory connections was set 0.04125 and the weights of the inhibitory to excitatory and inhibitory to inhibitory connections were set to -0.16625 when $J_0 = 1$. Thus for the largest possible integer weight matrix coefficient $J_0 = 7$ in Fig. 7(b), the excitatory to excitatory, excitatory to inhibitory, and inhibitory to both excitatory and inhibitory weights were 0.6125, 0.28875, and -1.16375, respectively. The diagonal entries were always set to -1 for different J_0 to simulate a soft refractory period for the neurons from their own spike history.

APPENDIX B: NEURONAL CONNECTION INFERENCE WITH MAXIMUM LIKELIHOOD ESTIMATION

We infer the neuronal connections based on a generalized linear model with an exponential inverse-link function, which amounts to a model-matched inference given the same model used in generating the spike trains. The observed neuron spike train $\{\dot{n}_i(t)\}$ are assumed to follow $\dot{n}_i(t)dt \sim \text{Poiss}[\hat{\phi}(\hat{\mu}_i + \sum_{j\in obs} \int dt' \, \hat{J}_{ij}(t - t')\dot{n}_j(t'))dt]$, where $\hat{\mu}_i$ and $\hat{J}_{ij}$ are the inferred baseline drive and interaction filters to be determined and the observation window dt is set to 0.1 ms. The likelihood function is thus

$$L_i(\hat{\mu}, \hat{J}) = \text{Prob}(\{\dot{n}_i(t)\}|\hat{\mu}_i, \hat{J}_{ij})$$
$$= \prod_t \frac{(\hat{\phi}_i(t)dt)^{\dot{n}_i(t)dt}}{(\dot{n}_i(t)dt)!} e^{-\hat{\phi}_i(t)dt}. \quad \text{(B1)}$$

We use the Tweedie regressor with power 1 and log-link function in the scikit-learn (v.0.24.2) package to perform the inference [41], which minimizes the unit deviance and can be shown to be equivalent to maximizing the likelihood function in Eq. (B1). No regularization penalty is added for all the inference procedures used in this work.

For the inference of the filters with basis functions as shown in Fig. 1(c), we use basis functions of the form

$$\alpha_n(t) = t^n \exp(-t/\tau)\Theta(t)/\tau^n,$$

for $n = 0$, 1, and 2. In this scenario, the number of unknowns for inferring the filters decreases to 3, the same as the number of basis functions. The inferred neuronal connections are truncated at 100 observation windows, corresponding to 10 ms for the chosen time window $dt = 0.1$ ms. These basis functions are motivated by theoretical work that suggests the linear spike train filters in the true effective model for the observed neurons is a series of such functions [11].

For the inference of the filters without using basis functions, we use the same number of 100 observation windows, and thus 100 unknowns need to be inferred to determine the coupling filters at each time point preceding the spikes.

APPENDIX C: MEAN-FIELD ANALYSIS WITH GAUSSIAN FLUCTUATION CORRECTIONS

To estimate firing rates and covariances in the spiking network model we use the path integral formalism introduced for the stochastic spiking model [10,11,19,42]. Following Ocker *et al.* [10], we introduce an auxiliary variable $\tilde{n}$, called the “response variable,” and then the action of the spike train process under our neuron model becomes

$$S[\tilde{n}, \dot{n}] = \sum_{i=1}^{N} \int_{-\infty}^{\infty} dt \Bigg[\tilde{n}_i(t)\dot{n}_i(t) - (e^{\tilde{n}_i(t)} - 1) \times \phi\left(\mu_i + \sum_{j(=1}^{N} \int_{-\infty}^{t} dt'\, J_{ij}(t-t')n_j(t')\right)\Bigg], \tag{C1}$$

such that the joint probability distribution of the spike train and auxiliary variable follows

$$\text{Prob}[\tilde{n}, \dot{n}] \propto e^{-S[\tilde{n},\dot{n}]}. \tag{C2}$$

Going forward, we make a change of variables $\dot{n}_i = r_i + \delta n_i$ where $r_i = \langle \dot{n}_i \rangle$ is the mean firing rate of neuron i, so that the expansion below is around the first moment of the spike train process. Eq. (C1) can be split into free and interacting actions. We expand the action in powers of $\delta\dot{n}_i(t)$ and $\tilde{n}_i(t)$, keeping only terms to quadratic order, which amounts to the Gaussian process approximation,

$$S[\tilde{n}, \delta\dot{n}] \approx \sum_{ij} \int dt dt' \left\{ \tilde{n}_i (\Delta^{-1})_{ij}(t,t')\delta\dot{n}_j - \frac{1}{2}\tilde{n}_i(t)^2 \phi_i \right\}, \tag{C3}$$

where

$$(\Delta^{-1})_{ij}(t,t') \equiv \delta_{ij}\delta(t-t') - \phi_i^{(1)} J_{ij}(t-t') \tag{C4}$$

is the inverse of the linear response function and $\phi_i^{(n)} = \frac{d^n\phi(x)}{dx^n}|_{x=\mu_i+\sum_j \mathcal{J}_{ij}r_j}$ is the nth derivative of the nonlinear activation function evaluated at the mean firing rate r_i. The linear order terms have been eliminating by imposing that r_i satisfies

$$r_i = \phi\left(\mu_i + \sum_j \mathcal{J}_{ij} r_j\right), \tag{C5}$$

and assuming a time-independent solution. This is equivalent to calculating the average of the spike train $\langle \dot{n}_i(t) \rangle$ by neglecting fluctuations in spiking and self-consistently estimating the firing rate [Eq. (6) in the main text].

The quadratic action (C3) corresponds to a Gaussian distribution for the fluctuations $\delta\dot{n}_i(t)$, which have zero mean and covariance [10]

$$C_{ij}(t', t'') = \sum_k \int_{-\infty}^{\infty} dt\, \Delta_{ik}(t', t)\Delta_{jk}(t'', t) r_k. \tag{C6}$$

For steady-state networks the response function and covariances are time-translation invariant, $\Delta_{ij}(t,t') = \Delta_{ij}(t-t')$ and $C_{ij}(t,t') = C_{ij}(t-t')$. In this case we can compute these functions efficiently in the Fourier domain. The linear response function can be computed by a matrix inverse for each frequency ω:

$$\mathbf{\Delta}(\omega) = [\mathbb{I} - \text{diag}(\mathbf{g})\mathbf{J}(\omega)]^{-1}, \tag{C7}$$

where we have written the equation in matrix form, with $\mathbf{g}$ a vector of the mean-field estimates of the gain, $g_i = \phi'(\mu_i + \sum_j \mathcal{J}_{ij} r_j)$. In the case of exponential nonlinearity that we focus on in this work, $g_i = r_i$.

Given $\mathbf{\Delta}(\omega)$, the covariance can then be calculated as

$$\mathbf{C}(\omega) = \mathbf{\Delta}(\omega)\text{diag}(\mathbf{r})\mathbf{\Delta}^T(-\omega). \tag{C8}$$

To calculate the response functions and covariances numerically we discretize frequencies into steps of size $d\omega = 0.1$ and frequencies from $d\omega \times (-1000 : 1 : 1000)$ (i.e., 2001 frequency bins). We transform back into the time domain by numerically discretizing the integrals in the inverse Fourier transform,

$$\mathbf{\Delta}(t) = \int_{-\infty}^{\infty} dt e^{-i\omega t}\mathbf{\Delta}(\omega),$$

where time is discretized in steps of $dt = \pi/1000$ and the limits are taken to be $\pm 2\pi/d\omega$. To improve the numerical accuracy of the inverse Fourier transform, we find it is actually beneficial to subtract the identity matrix from $\mathbf{\Delta}(\omega)$, which removes a δ function in the time-domain; i.e., we inverse transform $\mathbf{\Delta}(\omega) - \mathbb{I}$ and similarly compute the covariance function with the δ function at zero-time removed.

APPENDIX D: MAXIMUM LIKELIHOOD ESTIMATE EQUATIONS USING THE PATH INTEGRAL FORMALISM

Maximizing the likelihood function in Eq. (B1) amounts to solve for the zero points of its derivatives with respect to the unknowns $\hat{J}_{ij}$ and $\hat{\mu}_i$ in $\hat{\phi}_i(t) = \lambda_0 \exp(\hat{\mu}_i + \sum_j \hat{J}_{ij}\dot{n}_j)$. For mathematical simplicity, we take the logarithm of the likelihood to get the log-likelihood function $\mathcal{L}_i = \log(L_i)$ and maximize the log-likelihood,

$$\begin{aligned}\frac{\partial \mathcal{L}_i}{\partial \hat{\mu}_i} &= \lim_{T\to\infty}\frac{1}{T}\int_{-T/2}^{T/2} dt\left(\frac{\dot{n}_i(t)}{\hat{\phi}_i(t)}-1\right)\partial_{\hat{\mu}_i}\hat{\phi}_i(t)\\ &= \lim_{T\to\infty}\frac{1}{T}\int_{-T/2}^{T/2} dt\left(\frac{\dot{n}_i(t)}{\hat{\phi}_i(t)}-1\right)\phi_i(t)\\ &= 0,\end{aligned}\tag{D1}$$

where we note that $\partial_{\hat{\mu}_i}\hat{\phi}_i(t)=\hat{\phi}_i(t)$ for the choice of exponential nonlinearity. For a stationary system the time average will tend to the expected value due to ergodicity, and is equivalent to forming the log-likelihood using a large number of independent trials, which limit to the expected value for infinitely many trials. Thus, Eq. (D2) can be simplified to

$$\langle \dot{n}_i(t)\rangle = \langle \hat{\phi}_i(t)\rangle,\tag{D2}$$

which will be independent of the time t for a stationary process, which we assume the steady state to be.

Similarly,

$$\begin{aligned}\frac{\delta \mathcal{L}_i}{\delta \hat{J}_{ij}(t)} &= \lim_{T\to\infty}\frac{1}{T}\int_{-T/2}^{T/2} dt'\left(\frac{\dot{n}_i(t')}{\hat{\phi}_i(t')}-1\right)\partial_{\hat{J}_{ij}(t)}\hat{\phi}_i(t')\\ &= \lim_{T\to\infty}\frac{1}{T}\int_{-T/2}^{T/2} dt'\left(\frac{\dot{n}_i(t')}{\hat{\phi}_i(t')}-1\right)\hat{\phi}_i(t')\dot{n}_j(t'-\tau)\\ &= 0,\end{aligned}\tag{D3}$$

where we note that $\partial_{\hat{J}_{ij}(t)}\hat{\phi}_i(t')=\hat{\phi}_i(t')\dot{n}_j(t'-t)$ for the choice of exponential nonlinearity $\hat{\phi}(x)=\lambda_0\exp(x)$ and thus the equation can be reduced to Eq. (10).

As explained in the main text, for an exponential nonlinearity we can relate the expectations over $\hat{\phi}_i(t)$ to the moment-generating functional of the spiking process,

$$\langle \hat{\phi}_i(t)\rangle = \lambda_0 e^{\hat{\mu}_i} Z[\tilde{j}_j(t') = \hat{J}_{ij}(t-t')],$$

$$\langle \hat{\phi}_i(t)\dot{n}_j(t')\rangle = \lambda_0 e^{\hat{\mu}_i}\left.\frac{\delta Z[\tilde{j}]}{\delta \tilde{j}_j(t')}\right|_{\tilde{j}_j(t')=\hat{J}_{ij}(t-t')}$$

[Eqs. (11) and (12) in the main text], where $Z[\tilde{j}]\equiv\langle\exp(\sum_i\int dt\,\tilde{j}_i(t)\dot{n}_i(t))\rangle$. The moment generating functional cannot generally be solved in closed form, so to make use of these equations we will need an approximation. We use the mean-field approximation with Gaussian fluctuation corrections described in Appendix C to approximate the spike trains as a Gaussian process and compute the moment generating functional is known in closed form.

For a Gaussian process the moment generating functional of the spike train is [19]

$$\begin{aligned}Z[\tilde{j}] &= \int \mathcal{D}\tilde{n}(t)\mathcal{D}\dot{n}(t)\, e^{\sum_i\int dt\,\tilde{j}_i(t)\dot{n}_i(t)}\, e^{-S[\tilde{n},\dot{n}]}\\ &\approx \exp\left(\sum_i\int dt\;\tilde{j}_i(t)r_i\right.\\ &\quad\left. + \frac{1}{2}\sum_{ij}\int dt'dt''\tilde{j}_i(t')C_{ij}(t',t'')\tilde{j}_j(t'')\right).\end{aligned}\tag{D4}$$

Combined with Eqs. (11) and (12), the derived moment generating functional can be used to evaluate the expectations in the MLE equations in Eqs. (9) and (10), leading to the closed maximum likelihood estimation equations,

$$C_{ij}(t-t') = r_i\sum_{k=1}^{N_{\rm obs}}\int_{-\infty}^{\infty} dt''C_{jk}(t'-t'')\hat{J}_{ik}(t-t'')$$

[Eq. (14) in the main text], which establishes the relationship between the spike train covariance function C_{ij} and the MLE inferred filters $\hat{J}_{ij}(t)$ in the Gaussian process approximation.

APPENDIX E: GENERAL SOLUTION OF THE INTEGRAL EQUATION

Equation (14) is an integral equation for the unknown filters $\hat{J}_{i\ell}(t')$. One might hope to be able to extend the limits of integration to the entire real line and take a Fourier transform to obtain a matrix system of equations that can be solved, but without explicitly imposing the causality constraint this procedure will generally yield a noncausal solution. To use the Fourier method, one first needs to generalize the equation to

$$G_{ij}(t) = r_i\sum_{\ell=1}^{N_{\rm obs}}\int_{-\infty}^{\infty} dt''\,\hat{J}_{i\ell}(t'')C_{j\ell}(t''-t),\tag{E1}$$

where

$$G_{ij}(t) = \begin{cases} C_{ij}(t), & t>0,\\ G^-_{ij}(t), & t\leqslant 0,\end{cases}\tag{E2}$$

for some unknown functions $G^-_{ij}(t)$ that must be determined as part of our solution. Although we have introduced an extra set of unknowns, once we solve for $\hat{J}_{i\ell}(t'')$ the $G^-_{ij}(t)$'s will be determined. This extra set of functions enables causal solutions for the filter by absorbing any noncausal pieces into them. We apply the Fourier transform to obtain

$$G^+_{ij}(\omega)+G^-_{ij}(\omega) = \sum_{\ell=1}^{N_{\rm obs}} r_i\hat{J}^+_{i\ell}(\omega)C_{\ell j}(\omega),$$

where we use $C_{j\ell}(t)=C_{\ell j}(-t)$ and defined the transforms

$$f^+(\omega) = \int_{0+}^{\infty} dt e^{-i\omega t}f(t),$$

$$f^-(\omega) = \int_{-\infty}^{0^+} dt e^{-i\omega t}f(t),$$

$$f(\omega) = \int_{-\infty}^{\infty} dt e^{-i\omega t}f(t).$$

We can write the equation to solve in matrix form,

$$\mathbf{G}^+(\omega)+\mathbf{G}^-(\omega) = \hat{\mathbf{J}}^+(\omega)\mathbf{C}(\omega).$$

Next, we assume we can decompose $\mathbf{C}(\omega)=\mathbf{S}_+(\omega)\mathbf{S}_-(\omega)$, where $\mathbf{S}_+(\omega)$ is analytic and nonvanishing in the upper half plane and $\mathbf{S}_-(\omega)$ is analytic and nonvanishing in the lower half plane.

Continuing, we assume $\mathbf{S}_-(\omega)$ has an inverse, such that we may write

$$\mathbf{G}^+(\omega)[\mathbf{S}_-(\omega)]^{-1}+\mathbf{G}^-(\omega)[\mathbf{S}_-(\omega)]^{-1} = \hat{\mathbf{J}}^+(\omega)\mathbf{S}_+(\omega).$$

Next, we split $\mathbf{G}^+(\omega)[\mathbf{S}_-(\omega)]^{-1} = (\mathcal{F}^{-1}[\mathbf{G}^+[\mathbf{S}_-]^{-1})^+(\omega) + (\mathcal{F}^{-1}[\mathbf{G}^+[\mathbf{S}_-]^{-1}])^-(\omega)$, where the two terms are defined by first taking the inverse Fourier transform of the left-hand side and then splitting the terms up in causal and acausal halves, which are then Fourier transformed. We can then rearrange our equation as

$$\begin{aligned}(\mathcal{F}^{-1}[\mathbf{G}^+[\mathbf{S}_-]^{-1}])^-(\omega) + \mathbf{G}^-(\omega)[\mathbf{S}_-(\omega)]^{-1}\\ = \hat{\mathbf{J}}^+(\omega)\mathbf{S}_+(\omega) - (\mathcal{F}^{-1}[\mathbf{G}^+[\mathbf{S}_-]^{-1})^+(\omega),\end{aligned}$$

where by construction the left-hand side has all of its poles in the lower half plane and the right hand side has all of its poles in the upper half plane. Because the two sides are analytic on different half-planes, the only possibility is that they are both equal to the same function, which must be polynomial of degree n if we require the growth at $|\omega| \to \infty$ to be less than $\mathcal{O}(\omega^n)$ [21]. If we demand that the filters decay as $|\omega| \to \infty$ (which excludes a δ-function component), then the only option is that the two sides must be equal to zero, and hence we arrive at the formal solution

$$\hat{\mathbf{J}}^+(\omega) = (\mathcal{F}^{-1}[\mathbf{G}^+[\mathbf{S}_-]^{-1})^+(\omega)[\mathbf{S}_+(\omega)]^{-1}. \tag{E3}$$

In practice, the primary obstacles in performing this procedure are finding a spectral decomposition of the kernel $\mathbf{C}(\omega)$ and then splitting up $\mathbf{G}^+(\omega)[\mathbf{S}_-(\omega)]^{-1}$ into its separate additive factors that are analytic on different half planes. For a one-dimensional system there is a general procedure for performing both of these steps, but for a system of equations the noncommutativity of matrices prevents the use of the scalar method.

APPENDIX F: SOLUTION FOR WEAK COUPLING

While the Weiner-Hopf integral equation cannot be solved easily in general, but we can derive an approximate solution in the case of weak synaptic coupling. In the weak coupling limit we can derive the linear response functions and spike train covariances to a desired order in the mean-field firing rates r_i. We will work to quadratic order here. To this end we introduce a bookkeeping parameter ε attached to the firing rates, which we will use to keep track of the order of terms in a series expansion. Introducing this parameter into the matrix expression for the response function $\mathbf{\Delta}(\omega)$ gives

$$\begin{aligned}\mathbf{\Delta}(\omega) &= [\mathbb{I} - \varepsilon\mathbf{A}(\omega)]^{-1}\\ &\approx \mathbb{I} + \varepsilon\mathbf{A}(\omega) + \varepsilon^2\mathbf{A}(\omega)^2 + \cdots,\end{aligned}$$

where $(\mathbf{A}(\omega))_{ij} = g_i J_{ij}(\omega)$ and $\mathbf{A}^\dagger(\omega) = \mathbf{A}^T(-\omega)$. The covariance is then

$$\begin{aligned}\mathbf{C}(\omega) &= \Delta(\omega)\mathrm{diag}(\mathbf{r})\Delta^\dagger(\omega)\\ &\approx [\mathbb{I} + \varepsilon\mathbf{A}(\omega) + \varepsilon^2\mathbf{A}(\omega)^2]\\ &\quad\times \mathrm{diag}(\mathbf{r})[\mathbb{I} + \varepsilon\mathbf{A}^\dagger(\omega) + \varepsilon^2\mathbf{A}^\dagger(\omega)^2]\\ &= \mathrm{diag}(\mathbf{r}) + \varepsilon[\mathbf{A}(\omega)\mathrm{diag}(\mathbf{r}) + \mathrm{diag}(\mathbf{r})\mathbf{A}^\dagger(\omega)]\\ &\quad+ \varepsilon^2[\mathbf{A}(\omega)^2\mathrm{diag}(\mathbf{r}) + \mathrm{diag}(\mathbf{r})\mathbf{A}^\dagger(\omega)^2\\ &\quad+ \mathbf{A}(\omega)\mathrm{diag}(\mathbf{r})\mathbf{A}^\dagger(\omega)].\end{aligned}$$

Working out the individual terms we find

$$\begin{aligned}(\mathbf{A}(\omega)\mathrm{diag}(\mathbf{r}))_{ij} &= g_i J_{ij}(\omega) r_j,\\ (\mathrm{diag}(\mathbf{r})\mathbf{A}^\dagger(\omega))_{ij} &= g_j J_{ji}(-\omega) r_i,\\ (\mathbf{A}(\omega)^2\mathrm{diag}(\mathbf{r}))_{ij} &= \sum_k g_i J_{ik}(\omega) g_k J_{kj}(\omega) r_j,\\ (\mathrm{diag}(\mathbf{r})\mathbf{A}^\dagger(\omega)^2)_{ij} &= \sum_\ell g_j J_{j\ell}(-\omega) g_\ell J_{\ell i}(-\omega) r_i,\\ (\mathbf{A}(\omega)\mathrm{diag}(\mathbf{r})\mathbf{A}^\dagger(\omega))_{ij} &= \sum_\ell g_i J_{i\ell}(\omega) r_\ell g_j J_{j\ell}(-\omega).\end{aligned}$$

Putting everything together and returning to the time-domain, we may write the covariance as

$$\begin{aligned}C_{ij}(t-t') = r_i\delta_{ij}\delta(t-t') + \varepsilon C^{(1)}_{ij}(t-t')\\ + \varepsilon^2 C^{(2)}_{ij}(t-t') + \ldots,\end{aligned}$$

where

$$C^{(1)}_{ij}(t-t') = g_i J_{ij}(t-t') r_j + g_j J_{ji}(t'-t) r_i \tag{F1}$$

and

$$\begin{aligned}C^{(2)}_{ij}(t-t') = \sum_\ell \int_{-\infty}^{\infty} dt''\, [g_j J_{j\ell}(t'-t'') g_\ell J_{\ell i}(t''-t) r_i\\ + g_i J_{i\ell}(t-t'') g_\ell J_{\ell j}(t''-t') r_j\\ + g_i J_{i\ell}(t-t'') g_j J_{j\ell}(t'-t'') r_\ell].\end{aligned} \tag{F2}$$

Note that this result gives Eq. (8) in the main text.

Next, we assume we can expand the filter as

$$\hat{J}_{rr'}(t) = \varepsilon\hat{J}^{(1)}_{rr'}(t) + \varepsilon^2\hat{J}^{(2)}_{rr'}(t) + \cdots.$$

Plugging this into the integral equation gives, for $t > 0$,

$$r_r r_{r'}\hat{J}^{(1)}_{rr'}(t) = C^{(1)}_{rr'}(t), \tag{F3}$$

$$r_r r_{r'}\hat{J}^{(2)}_{rr'}(t) = C^{(2)}_{rr'}(t) - r_r\sum_{r''}\int_{0^+}^{\infty} dt_1\, \hat{J}^{(1)}_{rr''}(t_1) C^{(1)}_{r''r'}(t_1 - t). \tag{F4}$$

Because $t > 0$ we have

$$\hat{J}^{(1)}_{rr'}(t) = \frac{g_r}{r_r} J_{rr'}(t),$$

where the second term in $C^{(1)}_{rr'}(t)$ vanishes for $t < 0$. The second term in $r_r r_{r'}\hat{J}^{(2)}_{rr'}(t)$ reduces to

$$-\sum_{r''}\int_0^\infty dt_1 g_r J_{rr''}(t_1)[g_{r''}J_{r''r'}(t_1 - t) r_{r'} + g_{r'}J_{r'r''}(t - t_1) r_{r''}].$$

We can now use the fact that $J_{rr''}(t_1)$ is zero for $t_1 < 0$ to extend the range of integration to the whole real line. This will allow us to rewrite this term as

$$\begin{aligned}-\sum_{r''}\int_{-\infty}^{\infty} dt_1 [g_r J_{rr''}(t-t'') g_{r''} J_{r''r'}(t'') r_{r'}\\ + g_r J_{rr''}(t-t'') g_{r'} J_{r'r''}(-t'') r_{r''}],\end{aligned}$$

where we made the change of variables $t_1 = t - t''$ for reasons that will become clear momentarily.

Now, looking at $C^{(2)}_{ij}(t)$, we see that the term $J_{j\ell}(-t'')J_{\ell i}(t''-t)$ must vanish, because the first filter is nonzero when $t''<0$, but the second is nonzero when $t''>t>0$, meaning the two intervals are mutually exclusive and the term must vanish. This leaves

$$C^{(2)}_{rr'}(t)=\sum_{\ell}\int_{-\infty}^{\infty}dt''[g_rJ_{r\ell}(t-t'')g_\ell J_{\ell r'}(t'')r_{r'}+g_rJ_{r\ell}(t-t'')g_{r'}J_{r'\ell}(-t'')r_\ell].$$

Putting our results together, we see that

$$r_rr_{r'}\hat{J}^{(2)}_{rr'}(t)=\sum_{\ell}\int_{-\infty}^{\infty}dt''[g_rJ_{r\ell}(t-t'')g_\ell J_{\ell r'}(t'')r_{r'}+g_rJ_{r\ell}(t-t'')g_{r'}J_{r'\ell}(-t'')r_\ell]$$
$$-\sum_{r''}\int_{-\infty}^{\infty}dt''[g_rJ_{rr''}(t-t'')g_{r''}J_{r''r'}(t'')r_{r'}+g_rJ_{rr''}(t-t'')g_{r'}J_{r'r''}(-t'')r_{r''}].$$

We see that both sums look similar, except for the fact that the first sum is over *all* neurons, observed or hidden, while the second sum is only over the observed neurons. Thus, when all neurons are observed these two terms cancel out and $\hat{J}_{rr'}(t)\approx J_{rr'}(t)$ to $\mathcal{O}(\varepsilon^3)$. As stated in the main text, We expect this result to hold to all orders so that the ground-truth filters are recovered in the limit of a fully observed network. When not all neurons are observed a sum over the unobserved, or "hidden" neurons remains. This is the result quoted in the main text [Eq. (17)].

APPENDIX G: ITERATIVE NUMERICAL SOLUTION

While an analytic solution is not generally tractable, the integral equation for the inferred filters is amenable to an iterative solution. We define $\mathbf{A}(t)=\mathbf{R}\hat{\mathbf{J}}(t)\mathbf{R}$, such that the equation to be solved is

$$\mathbf{A}(t)=\bar{\mathbf{C}}(t)-\int_{0^+}^{\infty}dt''\mathbf{A}(t'')\mathbf{R}^{-1}\bar{\mathbf{C}}(t-t''),t>0.$$

Solving for $\mathbf{A}(t)$ instead of $\hat{\mathbf{J}}(t)$ avoids multiplications by $\mathbf{R}^{-1}$ on each iteration.

We convert this to the following iterative equation:

$$\mathbf{A}^{(n+1)}(t)=\eta\mathbf{A}^{(n)}(t)+(1-\eta)\Theta(t)\times\left(\bar{\mathbf{C}}(t)-\int_{-\infty}^{\infty}\frac{d\omega}{2\pi}e^{-i\omega t}\mathbf{A}^{(n)}(\omega)\bar{\mathbf{C}}(\omega)\right),$$

where η is an "update" rate that controls how much to update the solution on each iteration. For $\eta=0$ the convergence oscillates around the eventual limit, and values $\eta=0.5$ or 0.8 work well in practice. The Fourier transform implements a convolution,

$$\int_{0^+}^{\infty}dt''\mathbf{A}^{(n)}(t'')\bar{\mathbf{C}}(t-t'');$$

we use the fact that $\mathbf{A}^{(n)}(t'')$ is causal at every step of the iteration (enforced by the Heaviside step function) to extend the range of integration to the whole real line.

In the limit $n\to\infty$ this equation should converge to the solution of the original integral equation. We choose as our initial guess $\mathbf{A}^{(0)}(t)=\bar{\mathbf{C}}(t)\Theta(t)$ and iterate until the error

$$\frac{\int_{0^+}^{\infty}dt\left(A^{(n+1)}_{ij}(t)-A^{(n)}_{ij}(t)\right)^2}{\int_{0^+}^{\infty}dt\left(A^{(0)}_{ij}(t)\right)^2+\int_{0^+}^{\infty}dt\left(A^{(n)}_{ij}(t)\right)^2}<10^{-5}$$

for every pair ij. The denominator interpolates the error between a relative error and an absolute error, as in some cases $A^{(\infty)}_{ij}(t)$ may be zero.

To solve the iterative equation numerically we discretize the integrals in time and frequency space, using a frequency step-size of $d\omega=0.1$ and a frequency range of $d\omega\times(-1000:1:1000)$. The corresponding time domain uses a step size of $dt=2\pi/1000$ and $\pi/d\omega$ bins.

APPENDIX H: EXAMPLE CASE: ALL-TO-ALL COUPLED NETWORK DRIVEN BY INDEPENDENT NOISE

To evaluate an explicit example that is valid in the strong coupling regime, we consider an all-to-all coupled network with $J_{ij}(t-t')=Jh(t-t')$, for some temporal profile $h(t)$. We evaluate the solutions explicitly for an exponential filter, which has simpler analytic expressions than the alpha functions we use in our simulations, and then give the corresponding results for α function filters.

For this all-to-all network the mean-field equation reduces to a single rate equation (due to homogeneity of the network),

$$r=\lambda_0\exp{(\mu+NJr)},$$

where $\sum_j\int dt'\,J_{ij}(t-t')r_j$ integrates to NJr for constant r. By manipulating this into the form of the Lambert transcendental equation, $W(z)e^{W(z)}=z$ we can write the solution in terms of the Lambert W function,

$$r=-\frac{W_{-1}(NJ\lambda_0e^{\mu})}{NJ}.$$

Next, we can calculate the linear response function by inverting Eq. (C4). This is easiest to do by first Fourier transforming the equation to turn the convolutions in time into multiplications in frequency space, and then solving the matrix equation

$$\sum_{k=1}^{N}[\delta_{ik}-gJh(\omega)]=\delta_{ij},$$

where $g \equiv \phi'(\mu + NJr)$ is the gain of the network in steady-state (equal to the firing rate r when ϕ is exponential), and $h(\omega)$ is the Fourier transform of $h(t)$. If we denote $\mathbb{I}$ as the identity and $\mathbf{P}$ as a matrix of all 1's, then the inverse is [43]

$$[a\mathbb{I} + b\mathbf{P}]^{-1} = \frac{1}{a}\mathbb{I} - \frac{b}{a(Nb+a)}\mathbf{P}.$$

In our case we have $a = 1$, $b = -gJh(\omega)$, giving

$$\Delta_{ij}(\omega) = \delta_{ij} + \frac{gJh(\omega)}{1 - NJgh(\omega)}.$$

Let's now assume an exponential filter $h(t) = \exp(-t/\tau)\Theta(t)/\tau$, which has Fourier transform $h(\omega) = 1/(1 + i\omega\tau)$ using the convention $h(\omega) = \int_{-\infty}^{\infty} dt e^{-i\omega t} h(t)$. Thus, in the time-domain $\Delta_{ij}(t)$ is given by

$$\begin{aligned}\Delta_{ij}(t) &= \int_{-\infty}^{\infty} \frac{d\omega}{2\pi} e^{i\omega t}\left(\delta_{ij} + \frac{J}{h(\omega)^{-1} - NJg}\right)\\ &= \int_{-\infty}^{\infty} \frac{d\omega}{2\pi} e^{i\omega t}\left(\delta_{ij} + \frac{gJ}{i\omega\tau + 1 - NJg}\right)\\ &= \delta_{ij}\delta(t) + gJ\exp(-(1 - NJg)t/\tau)\Theta(t)/\tau,\end{aligned}$$

where to evaluate the second term we used the residue theorem: factoring out a $i\tau$ from the denominator, we observe a pole at $\omega = i(1 - NJg)$ in the upper-half plane when $1 > NJg$. This restriction requires either $0 < Jg < 1/N$ or $J < 0$ for the process to be stable. For $t < 0$, $i\omega t = -iR|t|(\cos\theta + i\sin\theta)$ on a contour of radius R, and the real part of this, $+R|t|\sin\theta$ is only negative in the lower-half plane, so we must close the contour there and the integral evaluates to zero because there are no poles contained in the contour. For $t > 0$ the real part of the arc is $-R|t|\sin\theta$, and we must close the arc in the upper half plane, obtaining the contribution from the pole. Note that the response function is causal in time.

With the response function in hand we may use Eq. (7) to evaluate the covariance for this model. In the Fourier domain we have

$$\begin{aligned}C_{ij}(\omega) &= \sum_{k=1}^{N} \Delta_{ik}(\omega)\Delta_{jk}(-\omega)r\\ &= r\sum_{k=1}^{N}\left(\delta_{ik} + \frac{J}{i\omega\tau + 1 - NJg}\right)\left(\delta_{jk} + \frac{Jg}{-i\omega\tau + 1 - NJg}\right)\\ &= r\sum_{k=1}^{N}\left(\delta_{ik}\delta_{jk} + \frac{J\delta_{jk}}{i\omega\tau + 1 - NJg} + \frac{Jg\delta_{ik}}{-i\omega\tau + 1 - NJg} + \frac{J^2g^2}{|i\omega\tau + 1 - NJg|^2}\right)\\ &= r\left[\delta_{ij} + 2Jg\mathrm{Re}\left[\frac{1}{i\omega\tau + 1 - NJg}\right] + \frac{NJ^2}{(1 - NJg)^2 + (\omega\tau)^2}\right]\\ &= r\left[\delta_{ij} + \frac{2Jg(1 - NJg) + NJ^2g^2}{(i\omega\tau + 1 - NJg)(-i\omega\tau + 1 - NJg)}\right]\\ &= r\left[\delta_{ij} + \frac{Jg(2 - NJg)}{(i\omega\tau + 1 - NJg)(-i\omega\tau + 1 - NJg)}\right].\end{aligned}$$

We again evaluate the inverse Fourier transform by using the Residue theorem. There are now two symmetric poles at $\omega = \pm i(1 - NJg)/\tau$, so we get a contribution from both planes, as expected for a covariance. The result is

$$C_{ij}(t) = r\left[\delta_{ij}\delta(t) + \frac{Jg(2 - NJg)}{2(1 - NJg)}\frac{\exp(-(1 - NJg)|t|/\tau)}{\tau}\right].$$

We can use this result with Eq. (14) to solve for the inferred filters $\hat{J}(t)$.

APPENDIX I: SOLUTION FOR HOMOGENEOUS NETWORKS

Equation (14) is a Wiener-Hopf integral equation that is difficult to solve in the multivariate case, but is tractable in the scalar case, which corresponds to a single observed neuron in our context. We calculate this here for unit $i = 1$ in the all-to-all network. The equation to solve is

$$G^{+}(\omega) + G^{-}(\omega) = r\hat{J}^{+}(\omega)C(\omega),$$

where

$$\begin{aligned}G^{+}(\omega) &= r\int_{0^+}^{\infty} d\tau e^{-i\omega\tau}\left(\delta(\tau) + \frac{a}{2b\tau}e^{-b|t|/\tau}\right)\\ &= \frac{ar}{2b}\frac{1}{i\omega\tau + b},\end{aligned}$$

where we introduce $a = Jg(2 - NJg)$ and $b = 1 - NJg$ to simplify the upcoming formulas. The full Fourier transform of $C(\omega)$

$$C(\omega) = r\left[1 + \frac{a}{(i\omega\tau + b)(-i\omega\tau + b)}\right].$$

Therefore, we need to solve the equation

$$\begin{aligned}
&\frac{a}{2b}\frac{1}{i\omega\tau+b}+r^{-1}G^{-}(\omega)\\
&\quad= r\hat{J}^{+}(\omega)\left(1+\frac{a}{(i\omega\tau+b)(-i\omega\tau+b)}\right)\\
&\quad= r\hat{J}^{+}(\omega)\left(\frac{(i\omega\tau+b)(-i\omega\tau+b)+a}{(i\omega\tau+b)(-i\omega\tau+b)}\right)\\
&\quad= r\hat{J}^{+}(\omega)\left(\frac{-(i\omega\tau)^2+b^2+a}{(i\omega\tau+b)(-i\omega\tau+b)}\right)\\
&\quad= r\hat{J}^{+}(\omega)\left(\frac{(i\omega\tau+\sqrt{b^2+a})(-i\omega\tau+\sqrt{b^2+a})}{(i\omega\tau+b)(-i\omega\tau+b)}\right),
\end{aligned}$$

where we divided both sides by one factor of r. We now separate the factors that are analytic and nonvanishing on the lower-half-plane and the upper half-planes. We have

$$\begin{aligned}
&\frac{a}{2b}\frac{1}{i\omega\tau+b}\frac{-i\omega\tau+b}{-i\omega\tau+\sqrt{b^2+a}}+r^{-1}G^{-}(\omega)\frac{-i\omega\tau+b}{-i\omega\tau+\sqrt{b^2+a}}\\
&\quad= r\hat{J}^{+}(\omega)\left(\frac{i\omega\tau+\sqrt{b^2+a}}{i\omega\tau+b}\right).
\end{aligned}$$

We use partial fractions on the left-hand-side to write

$$\begin{aligned}
&\frac{a}{2b}\frac{1}{i\omega\tau+b}\frac{-i\omega\tau+b}{-i\omega\tau+\sqrt{b^2+a}}\\
&\quad= \frac{A}{i\omega\tau+b}+\frac{B}{-i\omega\tau+\sqrt{b^2+a}},
\end{aligned}$$

where

$$A=\frac{a}{b+\sqrt{b^2+a}},\quad B=-\frac{a}{2b}\frac{\sqrt{b^2+a}-b}{\sqrt{b^2+a}+b}.$$

Here we will only care about the filter $\hat{J}^{+}(\omega)$, so we only need the A term. After separating the terms analytic in the upper versus lower half planes, demanding that the filters decay at infinite ω means we must have

$$\begin{aligned}
r\hat{J}^{+}(\omega)\left(\frac{i\omega\tau+\sqrt{b^2+a}}{i\omega\tau+b}\right) &= \frac{a}{b+\sqrt{b^2+a}}\frac{1}{i\omega\tau+b}\\
\Rightarrow r\hat{J}^{+}(\omega) &= \frac{a}{b+\sqrt{b^2+a}}\frac{1}{i\omega\tau+\sqrt{b^2+a}}
\end{aligned}$$

Because $\hat{J}^{+}(\omega)$ only has poles in the lower half plane, as desired, we know it will be causal and we can use the regular Fourier transform to recover it in the time domain (as $\hat{J}^{+}(\omega)=\hat{J}(\omega)$). The result is

$$\hat{J}(t)=\frac{1}{r}\frac{a}{b+\sqrt{b^2+a}}\frac{e^{-\sqrt{b^2+a}t/\tau}}{\tau}\Theta(t).$$

Restoring $a=Jg(2-NJg)$ and $b=1-NJg$ gives

$$\begin{aligned}
\hat{J}(t) = \frac{g}{r}&\frac{J(2-NJg)}{1-NJg+\sqrt{(1-NJg)^2+Jg(2-NJg)}}\\
&\times\frac{e^{-\sqrt{(1-NJg)^2+Jg(2-NJg)}t/\tau}}{\tau}\Theta(t). \qquad \text{(I1)}
\end{aligned}$$

We can check the limit of the fully resolved case when $N=1$. We have $(1-Jg)^2+Jg(2-Jg)=1-2Jg+(Jg)^2+2Jg-(Jg)^2=1$, giving $\hat{J}(t)=g/rJe^{-t/\tau}\Theta(t)/\tau$. This shows that when the nonlinearity of the generative model is not exponential (a model-mismatch) the inferred filter is off from the ground truth by a multiplicative factor g/r. For the model-matched case where both the generative model and the inference model use an exponential nonlinearity the gain is equal to the rate, $g=r$, and we recover the true filter. We can also verify in Mathematica that this solution does satisfy the original integral equation.

Now that we have $\hat{J}(t)$ and $C(t)$ we can evaluate the normalized overlap between them,

$$\rho=\frac{\int_0^\infty dt\hat{J}(t)C(t)}{\sqrt{\int_0^\infty dt\hat{J}(t)^2\int_0^\infty dtC(t)^2}}.$$

Using Mathematica, this works out to

$$\rho=\frac{2\sqrt{1-NJr}\sqrt[4]{(1-NJr)^2+Jr(2-NJr)}}{1-NJr+\sqrt{(1-NJr)^2+Jr(2-NJr)}}$$

for the model-matched case $g=r$. Plotting this as a function of $x=NJr\in[0,1)$ for fixed N, we see that for small x $\rho\approx 1$, and rapidly approaches 0 as $x\to 1$ from below. As N increases the fraction of the range of x for which $\rho\approx 1$ increases.

1. N_{obs} observed neurons

For the homogeneous all-to-all network we can also exactly solve for the inferred synaptic connections, as all connection filters will be the same, $\hat{J}_{rr'}(t)=\hat{J}(t)$. The equation we must solve becomes

$$\begin{aligned}
&\frac{a}{2b}\frac{1}{i\omega\tau+b}+r^{-1}G^{-}(\omega)\\
&\quad= r\hat{J}^{+}(\omega)\left(1+\frac{N_{\mathrm{obs}}a}{(i\omega\tau+b)(-i\omega\tau+b)}\right).
\end{aligned}$$

Following the same steps as above, we obtain the inferred filters

$$\begin{aligned}
\hat{J}(t) = \frac{g}{r}&\frac{J(2-NJg)}{1-NJg+\sqrt{(1-NJg)^2+N_{\mathrm{obs}}Jg(2-NJg)}}\\
&\times\frac{e^{-\sqrt{(1-NJg)^2+N_{\mathrm{obs}}Jg(2-NJg)}t/\tau}}{\tau}\Theta(t). \qquad \text{(I2)}
\end{aligned}$$

The reader can check that we recover the ground-truth filter when $N_{\mathrm{obs}}\to N$ and $g\to r$.

APPENDIX J: RESULTS FOR α-FUNCTION FILTERS

The manipulations work similarly for an alpha function filter $h(t)=te^{-t/\tau}\Theta(t)/\tau^2$, which has Fourier transform

$h(\omega) = 1/(i\omega\tau + 1)^2$. This introduces more poles to deal with when using the residue theorem and partial fraction decomposition, but the calculations are tractable for the most part. We find the covariance of the units to be

$$C_{ij}(t-t') = r\left[\delta_{ij}\delta(t-t') + \frac{(a_-(1)^2 - b_+^2)(b_+^2 - a_+(1)^2)}{(b_+ - b_-)(b_+ + b_-)}\frac{e^{-b_+|t-t'|/\tau}}{2b_+\tau} - \frac{(a_-(1)^2 - b_-^2)(b_-^2 - a_+(1)^2)}{(b_+ - b_-)(b_+ + b_-)}\frac{e^{-b_-|t-t'|/\tau}}{2b_-\tau}\right], \tag{J1}$$

where $a_\pm(n) = \sqrt{1 + (N-n)Jg \pm \sqrt{(N-n)Jg(4 - Jg)}}$ and $b_\pm = 1 \pm \sqrt{NJg}$, and the effective autocoupling (self-history) filter to be

$$\hat{J}(t) = \frac{1}{r}[A_+(N_{\text{obs}})e^{-a_+(N_{\text{obs}})t/\tau} - A_-(N_{\text{obs}})e^{-a_-(N_{\text{obs}})t/\tau}]\frac{\Theta(t)}{\tau}. \tag{J2}$$

The full expressions for the amplitudes are

$$\begin{aligned} DA_+(N_{\text{obs}}) = {} & a_-(1)^2 a_+(1)^2 b_-^2 + a_-(1)^2 a_+(1)^2 b_- b_+ + a_-(1)^2 a_+(1)^2 b_+^2 - a_-(1)^2 b_-^2 b_+^2 - a_+(1)^2 b_-^2 b_+^2 - b_-^3 b_+^3 \\ & - a_+(N_{\text{obs}})^2\{a_-(1)^2 a_+(1)^2 - b_-^3 b_+ - b_-^2 b_+^2 + b_- b_+(a_-(1)^2 + a_+(1)^2 - b_+^2)\} \\ & - a_-(N_{\text{obs}})(b_- + b_+)\{-a_-(1)^2 a_+(1)^2 + b_-^2 b_+^2 + a_+(N_{\text{obs}})^2(a_-(1)^2 + a_+(1)^2 - b_-^2 - b_+^2)\}, \end{aligned} \tag{J3}$$

$$\begin{aligned} -DA_-(N_{\text{obs}}) = {} & -a_-(1)^2 a_+(1)^2 b_-^2 - a_-(1)^2 a_+(1)^2 b_- b_+ - a_-(1)^2 a_+(1)^2 b_+^2 + a_-(1)^2 b_-^2 b_+^2 + a_+(1)^2 b_-^2 b_+^2 + b_-^3 b_+^3 \\ & + a_+(N_{\text{obs}})(b_- + b_+)\{-a_-(1)^2 a_+(1)^2 + b_-^2 b_+^2 + a_-(1)^2 + a_+(1)^2 - b_-^2 - b_+^2\} \\ & + a_-(N_{\text{obs}})^2\{a_-(1)^2 a_+(1)^2 - b_-^3 b_+ - b_-^2 b_+^2 + b_- b_+(a_-(1)^2 + a_+(1)^2 - b_+^2)\}, \end{aligned} \tag{J4}$$

where $D = (b_- + a_-(N_{\text{obs}}))(b_- + a_+(N_{\text{obs}}))(b_+ + a_-(N_{\text{obs}}))(b_+ + a_+(N_{\text{obs}}))(a_-(N_{\text{obs}}) - a_+(N_{\text{obs}}))$. As stated in the main text, when $N_{\text{obs}} = 1$ these simplify to

$$A_+(1) = \frac{(a_+(1) - b_-)(a_+(1) - b_+)}{a_+(1) - a_-(1)}, \quad A_-(1) = \frac{(a_-(1) - b_-)(a_-(1) - b_+)}{a_+(1) - a_-(1)}.$$

APPENDIX K: INFERENCE WITH NONEXPONENTIAL NONLINEARITIES

In this work we have focused on generative and inference models that use exponential nonlinearities for the instantaneous firing rate, $\phi(x) = \exp(x)$. This is the canonical choice in performing statistical inference as it guarantees, for example, that the log-likelihood function is convex, ensuring a unique solution. It also offers several simplifications in our theoretical analyses. However, some behaviors may not be possible with an exponential nonlinearity—the networks considered here become unstable if the synaptic connection strengths are too large, for example. The derivations given in the above Appendices generally assume an arbitrary nonlinearity in the generative model, and suggest that inference with an exponential nonlinearity will scale the inferred nonlinearities by the ratio of the gain to the firing rate, g_i/r_i [see, e.g., Eqs. (I1) and (I2)], though we do not simulate this case in this work.

In this Appendix we briefly describe how the formalism could be adapted to study networks for which the nonlinearity in the inference model is also nonexponential, highlighting the difficulties that arise.

For nonexponential nonlinearities the maximum likelihood equations are more complicated compared to Eqs. (9) and (10) in the main text:

$$\left\langle \dot{n}_i(t)\frac{\hat{\phi}_i'(t)}{\hat{\phi}_i(t)}\right\rangle = \langle\hat{\phi}_i'(t)\rangle,$$

$$\left\langle \dot{n}_i(t)\dot{n}_j(t'-\tau)\frac{\hat{\phi}_i'(t)}{\hat{\phi}_i(t)}\right\rangle = \langle\hat{\phi}_i'(t)\dot{n}_j(t'-\tau)\rangle,$$

where $\hat{\phi}_i(t) = \hat{\phi}(\mu_i + \int\sum_j \hat{J}_{ij}(t-t')\dot{n}_j(t'))$ and the prime on $\hat{\phi}_i(t)$ indicates a derivative with respect to the argument of $\hat{\phi}(x)$, not t. We cannot exploit the definition of the moment generating functional in this case, so the simplest approximation one can imagine performing is to write $\dot{n}_i(t) = r_i + \delta\dot{n}_i(t)$, where r_i is the mean-field approximation of the firing rate and $\delta\dot{n}_i(t)$ are fluctuations, which we might assume to be small enough to expand the nonlinearities $\hat{\phi}_i$ to second order in the fluctuations. We obtain a superficially similar system of integral equations for the filters

$$C_{rr'}(t) = \left[\hat{\phi}_r - r_r + r_r\frac{(\hat{\phi}_r')^2}{\hat{\phi}_r''\hat{\phi}_r}\right] \times \sum_{r''}\int_0^\infty dt''\hat{J}_{rr''}(t'')C_{r''r'}(t-t''),$$

where the nonlinearities are evaluated at the mean-field values, e.g. $\hat{\phi}_i \equiv \hat{\phi}(\hat{\mu}_i + \sum_k \int dt' \hat{J}_{ik}(t) r_k)$. If the nonlinearity is exponential, then the equation for the rates (not shown) reduces to $r_i = \hat{\phi}_i$ and we recover Eq. (14)—however, note that in the exponential case in the main paper our derivation only assumes the fluctuations can be approximated as Gaussian, not that they are weak, and no expansion of the nonlinearities is necessary. This said, the Gaussian approximation of the fluctuations is expected to be better when the fluctuations are weak.

For general nonlinearity the system of equations is much more difficult to solve because both unknowns $\hat{J}$ and $\hat{\mu}$ appear inside the nonlinearities, rendering this a nonlinear integral equation which could potentially have multiple solutions. The possibility of multiple solutions makes sense from analyses of feedforward networks. In feedforward GLMs (i.e., single units lacking synaptic connections) the log-likelihood is provably concave for a class of functions including $\phi(x) = x$, $|x|$, and $\exp(x)$, yielding a unique solution to the maximum likelihood estimation problem [35]. Other nonlinearities violate this concavity, and there is no guarantee of a unique solution.

Nonetheless, the approximately linear form of the integral equation suggests we may expect linear correlations between the spike-spike covariances and inferred filters even when the nonlinearity is nonexponential, but this could break down if the variability of spiking is strong—i.e., if $\delta\dot{n}_i(t)$ cannot be approximated as Gaussian and small relative to the mean r_i.

[1] J. W. Pillow, J. Shlens, L. Paninski, A. Sher, A. M. Litke, E. Chichilnisky, and E. P. Simoncelli, Nature (London) **454**, 995 (2008).

[2] D. Soudry, S. Keshri, P. Stinson, M.-H. Oh, G. Iyengar, and L. Paninski, PLoS Comput. Biol. **11**, e1004464 (2015).

[3] Y. V. Zaytsev, A. Morrison, and M. Deger, J. Comput. Neurosci. **39**, 77 (2015).

[4] M. E. Lepperød, T. Stöber, T. Hafting, M. Fyhn, and K. P. Kording, PLoS Comput. Biol., **19**, e1011574 (2023).

[5] D. Q. Nykamp, Math. Biosci. **205**, 204 (2007).

[6] I. H. Stevenson, J. M. Rebesco, L. E. Miller, and K. P. Körding, Curr. Opin. Neurobiol. **18**, 582 (2008).

[7] F. Gerhard, T. Kispersky, G. J. Gutierrez, E. Marder, M. Kramer, and U. Eden, PLoS Comput. Biol. **9**, e1003138 (2013).

[8] M. Volgushev, V. Ilin, and I. H. Stevenson, PLoS Comput. Biol. **11**, e1004167 (2015).

[9] R. Mohanty, W. A. Sethares, V. A. Nair, and V. Prabhakaran, Sci. Rep. **10**, 1298 (2020).

[10] G. K. Ocker, K. Josić, E. Shea-Brown, and M. A. Buice, PLoS Comput. Biol. **13**, e1005583 (2017).

[11] B. A. Brinkman, F. Rieke, E. Shea-Brown, and M. A. Buice, PLoS Comput. Biol. **14**, e1006490 (2018).

[12] D. Dahmen, S. Grün, M. Diesmann, and M. Helias, Proc. Natl. Acad. Sci. USA **116**, 13051 (2019).

[13] A. Das and I. R. Fiete, Nat. Neurosci. **23**, 1286 (2020).

[14] S. Ostojic and N. Brunel, PLoS Comput. Biol. **7**, e1001056 (2011).

[15] A. I. Weber and J. W. Pillow, Neural Comput. **29**, 3260 (2017).

[16] I. M. Park, M. L. Meister, A. C. Huk, and J. W. Pillow, Nat. Neurosci. **17**, 1395 (2014).

[17] M. T. Schaub, Y. N. Billeh, C. A. Anastassiou, C. Koch, and M. Barahona, PLoS Comput. Biol. **11**, e1004196 (2015).

[18] T. Rost, M. Deger, and M. P. Nawrot, Biol. Cybern. **112**, 81 (2018).

[19] C. C. Chow and M. A. Buice, J. Math. Neurosc. **5**, 8 (2015).

[20] R. Kubo, Rep. Prog. Phys. **29**, 255 (1966).

[21] A. V. Kisil, I. D. Abrahams, G. Mishuris, and S. V. Rogosin, Proc. R. Soc. A. **477**, 20210533 (2021).

[22] M. Vidne, Y. Ahmadian, J. Shlens, J. W. Pillow, J. Kulkarni, A. M. Litke, E. Chichilnisky, E. Simoncelli, and L. Paninski, J. Comput. Neurosci. **33**, 97 (2012).

[23] V. Pernice and S. Rotter, J. Stat. Mech. (2013) P03008.

[24] J. Schiefer, A. Niederbühl, V. Pernice, C. Lennartz, J. Hennig, P. LeVan, and S. Rotter, PLoS Comput. Biol. **14**, e1006056 (2018).

[25] J. Pillow and P. Latham, Adv. Neural Info. Process. Syst. **20** (2007).

[26] B. Dunn and Y. Roudi, Phys. Rev. E **87**, 022127 (2013).

[27] J. Tyrcha and J. Hertz, Math. Biosci. Eng **11**, 149 (2014).

[28] S. Wang, V. Schmutz, G. Bellec, and W. Gerstner, Adv. Neural Inf. Proc. Syst. **35**, 23566 (2022).

[29] S. Kim, D. Putrino, S. Ghosh, and E. N. Brown, PLoS Comput. Biol. **7**, e1001110 (2011).

[30] C. J. Quinn, T. P. Coleman, N. Kiyavash, and N. G. Hatsopoulos, J. Comput. Neurosci. **30**, 17 (2011).

[31] D. Zhou, Y. Xiao, Y. Zhang, Z. Xu, and D. Cai, PLoS ONE **9**, e87636 (2014).

[32] R. Kobayashi, S. Kurita, A. Kurth, K. Kitano, K. Mizuseki, M. Diesmann, B. J. Richmond, and S. Shinomoto, Nat. Commun. **10**, 4468 (2019).

[33] R. Biswas and E. Shlizerman, Front. Syst. Neurosci. **16**, 817962 (2022).

[34] M. Maziarz, J. Philos. Econ.: Reflect. Econ. Soc. Issues **8**, 86 (2015).

[35] L. Paninski, Network **15**, 243 (2004).

[36] G. K. Ocker, Phys. Rev. X **13**, 041047 (2023).

[37] D. F. English, S. McKenzie, T. Evans, K. Kim, E. Yoon, and G. Buzsáki, Neuron **96**, 505 (2017).

[38] K. Kim, M. Vöröslakos, J. P. Seymour, K. D. Wise, G. Buzsáki, and E. Yoon, Nat. Commun. **11**, 2063 (2020).

[39] F. Randi, A. K. Sharma, S. Dvali, and A. M. Leifer, Nature (London) **623**, 406 (2023).

[40] G. B. Ermentrout and D. H. Terman, *Mathematical Foundations of Neuroscience* (Springer, Berlin, 2010), pp. 157–170.

[41] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, J. Mach. Learn. Res. **12**, 2825 (2011).

[42] M. Kordovan and S. Rotter, arXiv:2001.05057.

[43] Inverse of constant matrix plus diagonal matrix, StackExchange; accessed January 03, 2023, https://math.stackexchange.com/questions/840855/inverse-of-constant-matrix-plus-diagonal-matrix.