Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Medicine and Health Sciences
Kavli Institute for Systems Neuroscience

Valdemar Kargård Olsen

Evaluating the quality of pairwise maximum entropy models in large neural datasets

Master's thesis in Neuroscience
Supervisor: Yasser Roudi
Co-supervisor: Benjamin Adric Dunn
May 2023

NTNU
Norwegian University of
Science and Technology

Valdemar Kargård Olsen

Evaluating the quality of pairwise maximum entropy models in large neural datasets

Master's thesis in Neuroscience
Supervisor: Yasser Roudi
Co-supervisor: Benjamin Adric Dunn
May 2023

Norwegian University of Science and Technology
Faculty of Medicine and Health Sciences
Kavli Institute for Systems Neuroscience

NTNU
Norwegian University of
Science and Technology

Preface

This master's thesis was written as part of the *NEVR3901 - Thesis in Neuroscience* course at the Norwegian University of Science and Technology (NTNU) during the academic year of 2022/2023. The code used in the analyses is available at https://github.com/MarvelousTurtle/Master_thesis.git.

I want thank Yasser Roudi and Benjamin Adric Dunn for their supervision and patience. I'm especially grateful to Yasser for the freedom to experiment and try all sorts of stuff that didn't work, and for the timely nudges in a more useful direction. I'm also thankful to Benjamin for including me in group meetings and social events. Additionally, my fellow master's students have provided valuable emotional support and sporadically fruitful discussions. Finally, I want to thank Elise Eknes for proofreading and being accepting of my funny working hours.

Abstract

An intuitive and attractive model for describing multi-neuronal activity is the pairwise maximum entropy model. This model has shown particular promise in capturing the experimentally observed probabilities of activity patterns, at least for few ($\sim$ 10) neurons N. The model was initially applied to retinal ganglion cell activity, but have later shown equally promising results in the cortex. However, whether this good performance for small N generalizes to larger N is unclear. Previous work has suggested that the quality of the pairwise model should be linear in $N\bar{v}\delta t$ when $N\bar{v}\delta t \ll 1$, the so-called perturbative regime, regardless of what the true probabilities of activity patterns are, where $\bar{v}$ is the mean firing rate and δt is the binsize. Here, we analysed data from the rat visual and auditory cortex, using various measures performance. We find that the performance of the pairwise model decays with $N\bar{v}\delta t$, although the model performs well in terms of predicting the third order correlations even for large $N\bar{v}\delta t$.

Sammendrag

En intuitiv og tiltrekkende modell for å beskrive aktiviteten i en populasjon av nevroner er den parevise maksimalentropi-modellen. Denne modellen har vist seg å være god til å finne de eksperimentelt observerte sannsynlighetene for forskjellige aktivitetsmønstre, i hvert fall for få (~ 10) nevroner N. Modellen ble først anvendt på aktivitet i ganglionceller i netthinnen, men har senere vist like lovende resultater i hjernebarken. Det er imidlertid uklart om denne gode ytelsen for liten N også holder for større N. Tidligere arbeid har antydet at ytelsen til den parvise modellen burde være lineær i $N\bar{\nu}\delta t$ når $N\bar{\nu}\delta t \ll 1$, det såkalte perturbasjonsregimet, uavhengig av hva den sanne sannsynligheten for forskjellige aktivitetsmønstre er, der $\bar{\nu}$ er gjennomsnittlig avfyringsfrekvens og δt er tidsintervallet. Her analyserte vi data fra den visuelle og auditive hjernebarken hos rotter, ved å bruke ulike mål for ytelse. Vi finner at ytelsen til parevise modellen avtar med $N\bar{\nu}\delta t$, selv om modellen predikerer tredjeordens korrelasjoner godt selv for store verdier av $N\bar{\nu}\delta t$.

Table of Contents

1 Why Care About Probabilities?

Hopefully, neuronal activity is not random. If it was, it would be difficult for the brain to generate useful behavior. An abundance of different ideas have been explored to find patterns in this non-random activity (Stephens et al., 2011. Among those are a range of dimensionality reduction techniques (Cunningham and Yu, 2014). If the activity of N different neurons is related in some way, one might be able to describe their activity with less than N numbers. This can be done through a number of methods, importantly those that take into account the stochastic nature of neural spiking. This might be useful for several reasons. First, the brain must deal with probabilities on some level due to the noisiness of individual neurons. For example, there is likely a probability distribution over activity patterns that correspond to a given head direction, and this distribution must be known to the brain somehow. Second, a description in terms of probability distributions may aid in the discovery of computational algorithms, for example by probabilistic examining dependencies between different neurons (Savin and Tkačik, 2017; Schneidman, 2016). Third, sampling from such distributions could produce brain-like synthetic data, which has a variety of purposes (Betzel and Bassett, 2017). Finally, probability distributions formalize the problem in a way that allows for the use of well-developed methods from other fields, such as statistical mechanics (e.g., Sompolinsky, 1988) or information theory (e.g., Timme and Lapish, 2018).

One way of defining discrete "brain states" is to let a state be the presence or absence of an action potential in tiny slivers of time, called timebins, for each recorded neuron. That is, a state is an activity pattern described by a binary vector. Ultimately, we may care more about states that are meaningfully different to the brain (whatever that means), and it seems unlikely that the brain distinguishes equally between all patterns of action potentials (see Section 6.5 and e.g., Ganmor et al., 2015). While we may eventually want to define brain states differently, the binary vector is a good starting point.

The simplest way of building an approximate probability distribution over these states from some recording of neuronal activity is to simply count the number of times each state is observed and dividing by the total number of observed states. For this to be a good approximation, one requires many samples/observations per state. When we care about few neurons at a time this is achievable in a typical experiment, but it rapidly becomes unfeasible because the number of possible states increases as 2^N with the number of neurons N. Thus, the number of possible state quickly becomes so big that we could not write it down even if we had enough data to reliably estimate the $2^N - 1$ parameters (probability of each state, in this case). So, as recording techniques improve and more neurons are recorded simultaneously (Gao and Ganguli, 2015; Stevenson and Kording, 2011; Yuste, 2015), the distribution over states have to be approximated by something other than counting. Preferably by finding some model that requires fewer than $2^N - 1$ parameters to be fit to the data. This nicely reflects our desire to find patterns in the data that allows for a simple description. A reasonable approach could be to make sure some feature(s) of the data match the model, while introducing as little structure as possible. That is, to make the approximated distribution as flat as possible given some constraints. This is exactly what maximum entropy models does (Savin and Tkačik, 2017). In particular, maximum entropy models that make all firing rates and pairwise correlations of the model match the data look promising as they only have $N + N(N-1)/2$ parameters and have been shown to approximate the true (counted) distribution well for few neurons (Schneidman et al., 2006; Shlens et al., 2006). This is known as the pairwise maximum entropy model (pairwise model, for brevity), or the Sherrington-Kirkpatrick (SK) Ising

model in the context of statistical mechanics (Sherrington and Kirkpatrick, 1975). Intuitively speaking, this is similar to fitting a multi-variate Guassian to a number of real valued observations, but it is done for binary variable.

Following these promising results, the pairwise model has been applied to different types of neural recordings, such as calcium imaging (Meshulam et al., 2017, 2021; Wolf et al., 2023), intracranial electroencephalography (iEEG; Ashourvan et al., 2021), and functional magnetic resonance imaging (fMRI; Ezaki et al., 2017; Watanabe et al., 2013) data. Note that instead of considering neurons, iEEG data typically considers electrodes while fMRI data typically considers regions of interest. Some have used inferred pairwise models to estimate functional connectivity between neurons (Kadirvelu et al., 2017) or brain areas (Ashourvan et al., 2021; Watanabe et al., 2013), to look for signatures of neuropsychiatric disorders (Ezaki et al., 2017), and for decoding of a binary stimulus (Posani et al., 2017). Beyond neuroscience, pairwise models have also seen use in a wide variety of biological systems ranging from amino acid interactions in proteins to voting interactions in the US supreme court (Cofré et al., 2019).

Despite this widespread use, we still don't know whether the good model fit we see for small N generalizes to large N. This prevents us from drawing strong conclusions about a system from a fitted pairwise model as N increases. However, it is difficult to evaluate how close the inferred model is to the data (i.e., how similar the probability of the different states are) for large N because the number of states grows exponentially with the number of neurons. This makes it challenging to estimate the true distribution from the data and to calculate the probability of each state in the model because it involves a sum over all states. Nevertheless, we here attempt to evaluate the performance of the pairwise model for large N. In order to express this aim precisely, we must first have a better understanding of pairwise maximum entropy models and how to evaluate them.

2 Maximum Entropy Models

Our goal is to make a probability distribution over some states $\mathbf{s}$. Here, these states represent activity patterns of a population of neurons. They are constructed by converting the spike times of a collection of N recorded neurons to spike trains by binning the spikes in time with a binsize of δt. That is, each sample is described by a vector of length N, $\mathbf{s}(t) = [s_1(t), s_2(t), ..., s_N(t)]$, where $s_i(t) = 1$ when neuron i spikes at least once in bin t and $s_i(t) = -1$ when it does not. Here, we will use a dataset recorded from the visual, auditory, somatosensory, and motor cortices of freely moving rats (Mimica et al., 2022), described in more detail in Section 4.1.

One way to describe the distribution over all states with less than $2^N - 1$ parameters is to only consider some features of the data. Apart from these features, the distribution is made as unstructured or uniform as possible; the entropy is maximised. This corresponds to adding no more information than that contained in the features. In the general case (Jaynes, 1957), the distribution that achieves this is called the (Gibbs-)Boltzmann distribution, and is given by

$$p(\mathbf{s}) \equiv \frac{1}{Z(\{g_\mu\})} \exp\left(\sum_\mu g_\mu f_\mu(\mathbf{s})\right). \tag{1}$$

The parameters g_μ are fit such that the expectation values $\langle f_\mu(\mathbf{s})\rangle$ of some features $f_\mu(\mathbf{s})$ in the model match the observed average values of those features in the data. The partition function

$Z(\{g_\mu\})$ normalize the distribution and is thus given by

$$Z \equiv \sum_{\mathbf{s}} \exp\left(\sum_{\mu} g_\mu f_\mu(\mathbf{s})\right). \tag{2}$$

2.1 Pairwise Maximum Entropy Models

In the pairwise maximum entropy model (pairwise model, for brevity), the features are simply the means (i.e., firing rates) and pairwise correlations of every neuron and pair of neurons, respectively. That is, the parameters are fit such that the means $\langle s_i \rangle_{\text{pair}}$ and correlations $\langle s_i s_j \rangle_{\text{pair}}$ of the pairwise model match the means $\langle s_i \rangle_{\text{data}}$ and correlations $\langle s_i s_j \rangle_{\text{data}}$ in the data. The model is given by

$$p_{\text{pair}}(\mathbf{s}) \equiv \frac{1}{Z} \exp\left(\sum_i h_i s_i + \sum_{i<j} J_{ij} s_i s_j\right), \tag{3}$$

where the biases (or external fields) h_i and couplings J_{ij} are parameters fit to the data. Note that the couplings are symmetric ($J_{ij} = J_{ji}$) and that self-connections are omitted ($J_{ii} = 0$), resulting in $N + N(N-1)/2$ parameters.

2.2 Independent Models

To evaluate the performance of a fitted pairwise model, one may want to compare it to another distribution fit to the data. A common choice here is the maximum entropy independent model in which neurons are assumed to spike independently of one another. By doing this comparison, one is measuring the importance of the pairwise correlations in accounting for the data. In the maximum entropy independent model, only the means $\langle s_i \rangle$ are matched, giving

$$p_{\text{ind}}(\mathbf{s}) \equiv \frac{1}{Z} \exp\left(\sum_i h_i s_i\right). \tag{4}$$

Now, only the biases h_i have to be fit to the data such that $\langle s_i \rangle_{\text{ind}}$ match $\langle s_i \rangle_{\text{data}}$. In the absence of J_{ij}s this is simple, as the biases are given by (Roudi, Aurell et al., 2009)

$$h_i^{\text{ind}} = \operatorname{arctanh} \langle s_i \rangle_{\text{data}}. \tag{5}$$

Another way of constructing an independent model to compare the pairwise model to, is simply remove the J_{ij} from the a pairwise model fitted to the data. By doing the comparison between a pairwise model and a pairwise model where the J_{ij}s are set to zero manually, we evaluate how much the pairwise model relies on its couplings in achieving whatever performance it achieves.

Here, we use both of these independent models, defined by h_{ind} or h, as a yardstick to measure the performance of the pairwise model, in addition to evaluating its reliance on the couplings J.

2.3 Approximating h and J

2.3.1 Boltzmann learning

Inferring the biases h_i and couplings J_{ij} of the pairwise model from data is more elaborate than the Independent max ent model. The simplest and most reliable way of doing so is to, starting from some initial value, iteratively change the parameters until the means $\langle s_i \rangle$ and correlations $\langle s_i s_j \rangle$ of the pairwise model match those in the data. This procedure is called Boltzmann learning, and the update rules are given by

$$\delta h_i = \eta(\langle s_i \rangle_{\text{data}} - \langle s_i \rangle_{\text{pair}}) \tag{6a}$$

$$\delta J_{ij} = \eta(\langle s_i s_j \rangle_{\text{data}} - \langle s_i s_j \rangle_{\text{pair}}) \tag{6b}$$

where η is the learning rate chosen such that h_i and J_{ij} converges (Ackley et al., 1985). At convergence the solutions will be that of the maximum likelihood estimates of h and J, given the data.

The means and correlations from the data can be calculated directly. However, the means and correlations from the pairwise model can only be calculated exactly for small N, when the sum over all states (in Z) is small enough. One solution is to sample the pairwise model using Monte Carlo sampling, with the current values of h and J, at each iteration and calculate the means and correlations from that. We start from some initial state, then pick one random neuron i and flip it such that $s_i \to -s_i$. Now, we compare the probability of the old unflipped state $\mathbf{s}_{\text{old}}$ with the new flipped state $\mathbf{s}_{\text{new}}$. We want to sample more high-probability than low-probability states, so we want to accept the new state as a sample if $p_{\text{pair}}(\mathbf{s}_{\text{new}})/p_{\text{pair}}(\mathbf{s}_{\text{old}})$ is larger than one and sometimes accept the new state if $p_{\text{pair}}(\mathbf{s}_{\text{new}})/p_{\text{pair}}(\mathbf{s}_{\text{old}})$ is less than one. More precisely, we accept the new state as a sample with probability

$$p = \min\left\{1, \frac{p_{\text{pair}}(\mathbf{s}_{\text{new}})}{p_{\text{pair}}(\mathbf{s}_{\text{old}})}\right\} = \min\left\{1, \exp\left(H[\mathbf{s}_{\text{old}}] - H[\mathbf{s}_{\text{new}}]\right)\right\}, \tag{7}$$

where $H(\mathbf{s}) = \sum_i h_i s_i + \sum_{i<j} J_{ij} s_i s_j$ is the "energy" or Hamiltonian of state $\mathbf{s}$. To reduce the influence of the arbitrary initial state, a burn-in period in which the first samples are discarded is often included. This algorithm is a special case of the Metropolis-Hastings algorithm (Ghojogh et al., 2020; Hastings, 1970). During Boltzmann learning, a given number of states are sampled at each iteration using this algorithm, and then used to approximate the means $\langle s_i \rangle_{\text{pair}}$ and correlations $\langle s_i s_j \rangle_{\text{pair}}$ of the pairwise model in Eq. (6). So, there are three tunable parameters in Boltzmann learning: the learning rate η, the number of iterations, and the number of samples per iteration. Biases and couplings obtained from Boltzmann learning are denoted as h^{boltz} and J^{boltz}.

Boltzmann learning is reliable in the sense that it will eventually converge to the maximum likelihood values of h and J, but this may take a long time. To address this, several approximate closed-form solutions have been developed. These approximations include the naive mean field (nMF; e.g., Roudi, Aurell et al., 2009; Roudi, Tyrcha et al., 2009), Thouless-Anderson-Palmer (TAP; Thouless et al., 1977), Independent Pair (IP; e.g., Roudi, Aurell et al., 2009; Roudi, Tyrcha et al., 2009), and Sessak-Monasson (SM; Sessak and Monasson, 2009) approximation. However, these often make assumptions that may be violated in neuronal data.

2.3.2 Naive mean-field approximation

In mean field theory generally, we have a system of many interacting elements that we simplify by averaging over degrees of freedom. In our case, we average neural activity over time so that neurons are approximated to only interact with each other via their mean activity. We can derive the mean field equation by applying this approximation to the mean $\langle s_i \rangle$ of neuron i. We start by rewriting $\langle s_i \rangle$ using the probability of $s_i = \pm 1$ given the state of all the other neurons $s_{/i}$,

$$\langle s_i \rangle = p(s_i = 1|s_{/i}) - p(s_i = -1|s_{/i}) = \frac{\exp\left(h_i + \sum_j J_{ij} s_j\right) - \exp\left(-h_i - \sum_j J_{ij} s_j\right)}{\exp\left(h_i + \sum_j J_{ij} s_j\right) + \exp\left(-h_i - \sum_j J_{ij} s_j\right)}$$
$$= \tanh\left(h_i + \sum_j J_{ij} s_j\right), \quad (8)$$

where the second equality follow from Z only being a sum over $s_i = \pm 1$. Now, we can apply the mean field approximation by replacing s_j with its mean $\langle s_j \rangle$ (Hertz et al., 2011), giving the mean field equation

$$\langle s_i \rangle = \tanh\left(h_i + \sum_j J_{ij} \langle s_j \rangle\right). \quad (9)$$

Another way to understand this approximation is that we assume that the fluctuations around the means are small or have a small effect. Eq. (9) is readily solved for h_i because we can calculate the means $\langle s_i \rangle$ and $\langle s_j \rangle$ from data. To find J_{ij} we take the derivative of Eq. (9) with respect to $\langle s_j \rangle$, obtaining the inverse susceptibility matrix

$$\chi^{-1}_{ij} = -J_{ij}, \quad (10)$$

which is equal to the inverse (connected) correlation matrix $\widetilde{C}_{ij} = \langle (s_i - \langle s_i \rangle)(s_j - \langle s_j \rangle) \rangle$. The naive Mean Field (nMF) approximation is then given by

$$J^{\mathrm{nMF}}_{ij} \equiv -(\widetilde{\mathbf{C}}^{-1})_{ij}. \quad (11a)$$

$$h^{\mathrm{nMF}}_i \equiv \tanh^{-1}\langle s_i \rangle - \sum_j J_{ij} \langle s_j \rangle \quad (11b)$$

To get some notion of the accuracy of the closed-form approximations, we plot their inferred parameters against those obtained from Boltzmann learning. For illustration purposes, we use the same $N = 20$ and $N = 100$ random neurons from the dataset (Mimica et al., 2022) in all of these comparisons (Figure 2.1-2.4). In Figure 2.1 we see that the nMF approximation is better for fewer neurons and typically overestimate the real parameters, in agreement with previous findings (Roudi, Tyrcha et al., 2009).

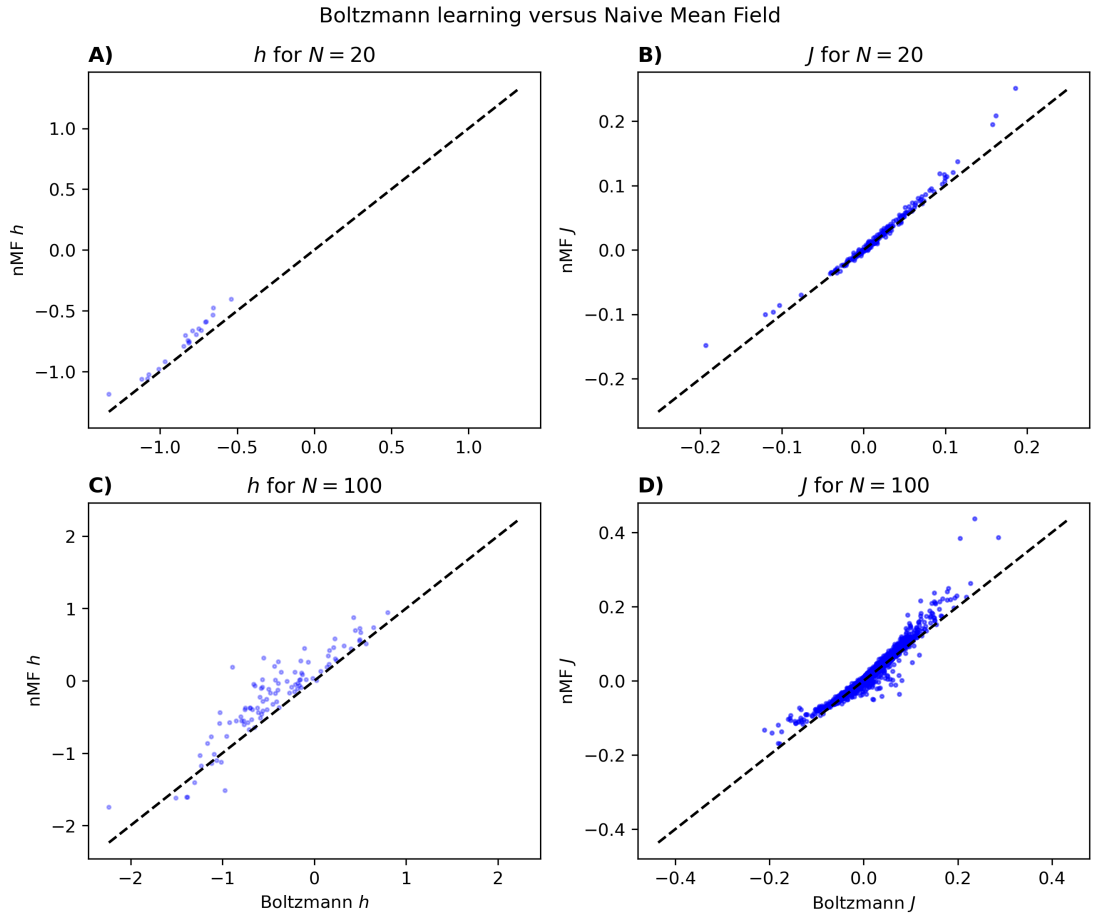


Figure 2.1: Boltzmann learning against nMF for neural data. (A-B) A random subpopulation of 20 out of the 495 neurons were chosen. The nMF parameters h_i^{nMF} and J_{ij}^{nMF} were plotted against the Boltzmann learning parameters h_i^{boltz} and J_{ij}^{boltz}, which used a learning rate of $\eta = 0.01$, 40000 iterations, and 10000 samples per iteration. (C-D) A random subpopulation of 100 out of the 495 neurons were chosen. The nMF parameters h_i^{nMF} and J_{ij}^{nMF} were plotted against the Boltzmann learning parameters h_i^{boltz} and J_{ij}^{boltz}, which used a learning rate of $\eta = 0.01$, 80000 iterations, and 50000 samples per iteration. This figure shows that the nMF approximation finds somewhat larger parameters than Boltzmann learning, especially for larger N.

2.3.3 Thouless-Anderson-Palmer approximation

The Thouless-Anderson-Palmer (TAP) approximation is an extension of the nMF approximation, where the effect of neuron i on its own mean, via $\langle s_j \rangle$ in Eq. (9), is corrected for (Thouless et al., 1977). This results in

$$\langle s_i \rangle = \tanh\left(h_i + \sum_j J_{ij}\langle s_j \rangle - \langle s_i \rangle \sum_j J_{ij}^2 \left(1 - \langle s_j \rangle^2\right) \right) \tag{12}$$

replacing Eq. (9). The new term is often called the Onsager correction term. One could continue to add new correction terms, getting a sequence of progressively better approximations, of which the nMF and TAP approximations are the first two (Plefka, 1982). Like in the nMF approximation, one can solve Eq. (12) for h_i and obtain J_{ij} by taking the derivative of Eq. (12) with respect to

$\langle s_j \rangle$. This results in the inverse susceptibility (i.e., connected correlation) matrix

$$\chi_{ij}^{-1} = (\widetilde{\mathbf{C}}^{-1})_{ij} = -J_{ij} - 2J_{ij}^2 \langle s_i \rangle \langle s_j \rangle, \tag{13}$$

which can be solved for J_{ij} (taking the root closest to $(\widetilde{\mathbf{C}}^{-1})_{ij}$). Thus, the TAP approximation is given by

$$J_{ij}^{\mathrm{TAP}} \equiv \frac{-1 + \sqrt{1 - 8\langle s_i \rangle \langle s_j \rangle (\widetilde{\mathbf{C}}^{-1})_{ij}}}{4\langle s_i \rangle \langle s_j \rangle}. \tag{14a}$$

$$h_i^{\mathrm{TAP}} \equiv \tanh^{-1} \langle s_i \rangle - \sum_j J_{ij} \langle s_j \rangle + 2J_{ij}^2 \langle s_i \rangle \langle s_j \rangle \tag{14b}$$

In general, both the nMF and TAP approximation performs best when the couplings J_{ij} are small (Plefka, 1982; Roudi, Aurell et al., 2009; Roudi, Tyrcha et al., 2009). In Figure 2.2 we see that the TAP approximation is considerably better than the nMF approximation. However, the overestimation observed previously (Roudi, Aurell et al., 2009; Roudi, Tyrcha et al., 2009) is less conspicuous here.

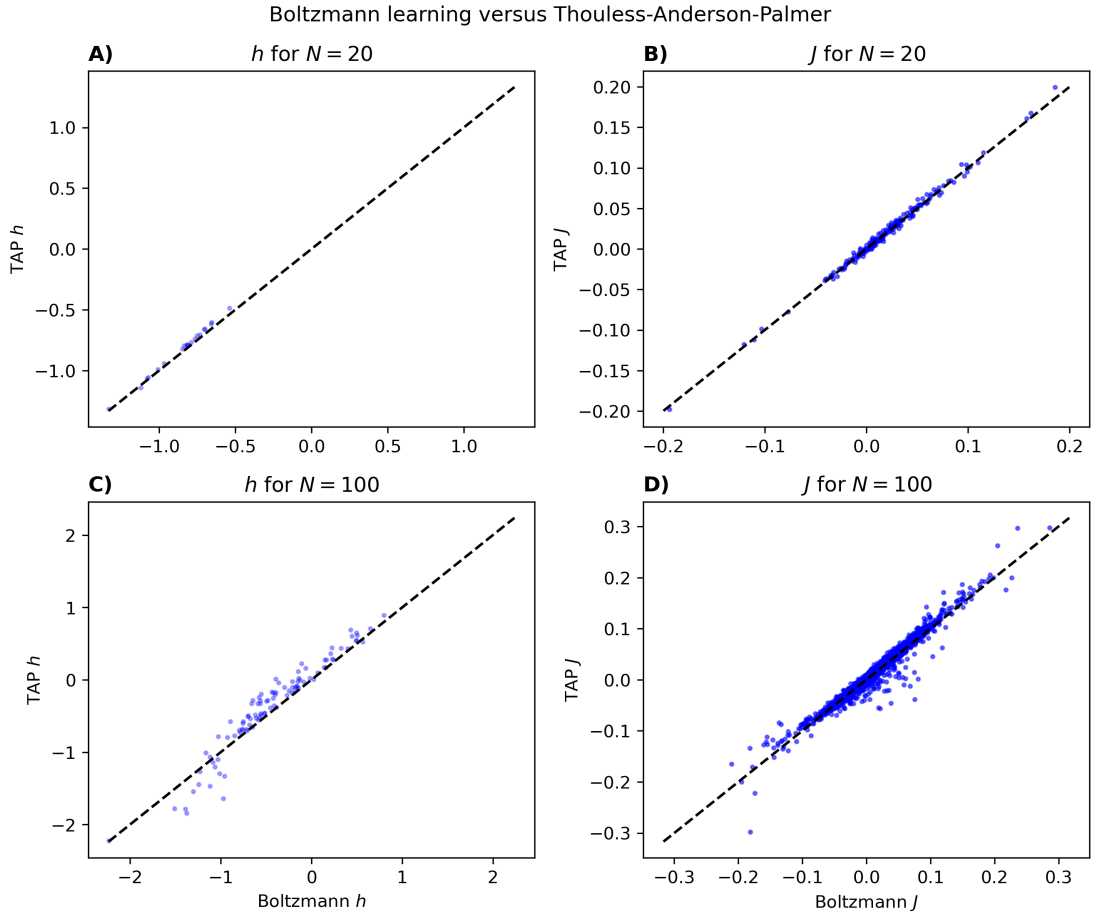


Figure 2.2: Boltzmann learning against TAP for neural data. (A-B) A random subpopulation of 20 out of the 495 neurons were chosen. The TAP parameters h_i^{TAP} and J_{ij}^{TAP} were plotted against the Boltzmann learning parameters h_i^{boltz} and J_{ij}^{boltz}, which used a learning rate of $\eta = 0.01$, 40000 iterations, and 10000 samples per iteration. (C-D) A random subpopulation of 100 out of the 495 neurons were chosen. The TAP parameters h_i^{TAP} and J_{ij}^{TAP} were plotted against the Boltzmann learning parameters h_i^{boltz} and J_{ij}^{boltz}, which used a learning rate of $\eta = 0.01$, 80000 iterations, and 50000 samples per iteration. This figure shows that Boltzmann learning and TAP find fairly similar parameters for a population of 20 and 100 neurons.

2.3.4 Independent pair approximation

In the Independent Pair (IP) approximation, we instead simplify the inference problem by considering subnetworks of only two neurons at a time. In the subnetwork consisting of neuron i and neuron j, defined by $p(\mathbf{s}) = \exp\left(h_i^j s_i + h_j^i s_j + J_{ij} s_i s_j\right)/Z$ where h_i^j is the bias on neuron i paired with neuron j, the parameters are given by (Roudi, Aurell et al., 2009; Roudi, Tyrcha et al., 2009)

$$J_{ij}^{\mathrm{IP}} \equiv \frac{1}{4}\log\left[\frac{((1+\langle s_i\rangle)(1+\langle s_j\rangle)+\widetilde{C}_{ij})((1-\langle s_i\rangle)(1-\langle s_j\rangle)+\widetilde{C}_{ij})}{((1-\langle s_i\rangle)(1+\langle s_j\rangle)-\widetilde{C}_{ij})((1+\langle s_i\rangle)(1-\langle s_j\rangle)-\widetilde{C}_{ij})}\right] \tag{15a}$$

$$h_i^j = \frac{1}{2}\log\left[\frac{(1+\langle s_i\rangle)(1-\langle s_j\rangle)-\widetilde{C}_{ij}}{(1-\langle s_i\rangle)(1-\langle s_j\rangle)+\widetilde{C}_{ij}}\right] + J_{ij}. \tag{15b}$$

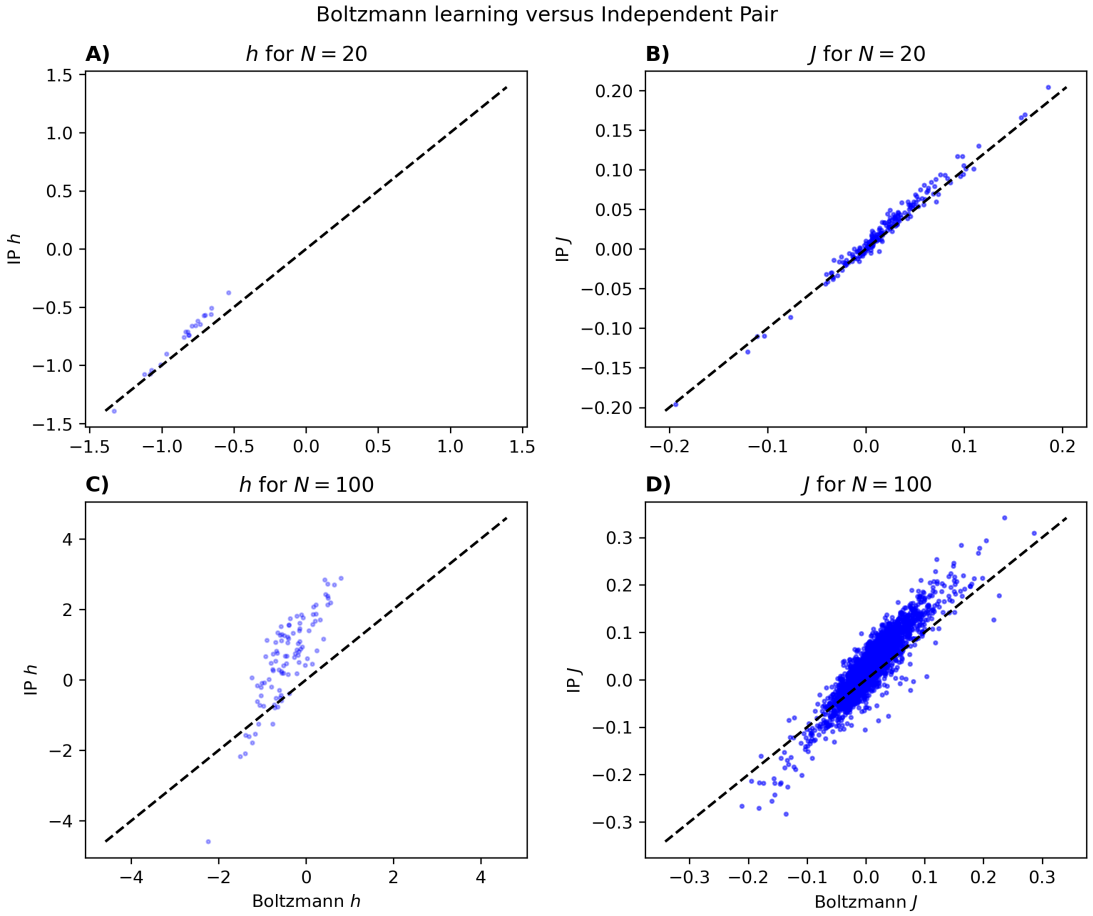


Figure 2.3: Boltzmann learning against IP for neural data. (A-B) A random subpopulation of 20 out of the 495 neurons were chosen. The IP parameters h_i^{IP} and J_{ij}^{IP} were plotted against the Boltzmann learning parameters h_i^{boltz} and J_{ij}^{boltz}, which used a learning rate of $\eta = 0.01$, 40000 iterations, and 10000 samples per iteration. (C-D) A random subpopulation of 100 out of the 495 neurons were chosen. The IP parameters h_i^{IP} and J_{ij}^{IP} were plotted against the Boltzmann learning parameters h_i^{boltz} and J_{ij}^{boltz}, which used a learning rate of $\eta = 0.01$, 80000 iterations, and 50000 samples per iteration. This figure shows that the IP approximation finds parameters similar to Boltzmann learning for a population of 20 neurons, but overestimate the magnitude for a population of 100 neurons.

The inferred couplings can be used directly, but the biases needs to account for the interaction

between neuron i and all other neurons, not just neuron j. The idea is to sum h_i^j over j, excluding $j = i$ (Roudi, Aurell et al., 2009). To do this, we decompose Eq. (15b) into terms that include j and terms that don't, and then sum the latter over all pairs where $j \neq i$, obtaining

$$h_i^{\mathrm{IP}} \equiv \frac{1}{2}\log\left[\frac{1+\langle s_i\rangle}{1-\langle s_i\rangle}\right] + \frac{1}{2}\sum_{j\neq i}\log\left[\frac{\frac{1-\langle s_j\rangle-\widetilde{C}_{ij}+\langle s_i\rangle-\langle s_i\rangle\langle s_j\rangle}{1+\langle s_i\rangle}}{\frac{1-\langle s_j\rangle+\widetilde{C}_{ij}-\langle s_i\rangle+\langle s_i\rangle\langle s_j\rangle}{1-\langle s_i\rangle}}\right] + \sum_{j\neq i} J_{ij}. \tag{16}$$

The IP approximation is thus given by Eq. (15a) and (16). The accuracy of both nMF and IP generally decrease with increased N and δt. It is conceptually simple to extend the IP approximation by considering sets of more than two neurons. However, this is computationally expensive and typically gives better approximations only for very small binsizes δt (Roudi, Aurell et al., 2009). Figure 2.3 shows that the accuracy of the IP approximation drops considerably as N increases from 20 to 100. We also see the previously observed (Roudi, Tyrcha et al., 2009) overestimation of the magnitude of the parameters for large N.

2.3.5 Sessak-Monasson approximation

Sessak and Monasson (2009) derived an approximation of the biases and couplings by performing a perturbative expansion in the (connected) correlations $\widetilde{C}_{ij}$. The Sessak-Monasson (SM) approximation is given by (Sessak and Monasson, 2009; Roudi, Tyrcha et al., 2009)

$$J_{ij}^{\mathrm{SM}} \equiv J_{ij}^{\mathrm{nMF}} + J_i j^{\mathrm{IP}} - \frac{\widetilde{C}_{ij}}{(1-\langle s_i^2\rangle)(1-\langle s_j^2\rangle) - \widetilde{C}_{ij}^2} \tag{17a}$$

$$\begin{aligned} h_i^{\mathrm{SM}} \equiv{}& \frac{1}{2}\log\left[\frac{1+\langle s_i\rangle}{1-\langle s_i\rangle}\right] - \sum_j J_{ij}^{\mathrm{SM}}\langle s_j\rangle + \sum_{j\neq i} K_{ij}\langle s_i\rangle L_j \\ &- \frac{2}{3}(1+3\langle s_i\rangle^2)\sum_{j\neq i} K_{ij}^3\langle s_j\rangle L_j - 2\langle m_i\rangle\sum_{j<k} K_{ij}K_{jk}K_{ki}L_jL_k \\ &+ 2\langle s_i\rangle\sum_{l<j}\sum_k K_{lk}K_{kj}K_{ji}K_{il}L_lL_jL_k \\ &+ \langle s_i\rangle\sum_j K_{ij}^4 L_j(1+\langle s_i\rangle^2 + 3\langle s_j\rangle^2 + 3\langle s_i\rangle^2\langle s_j\rangle^2) \\ &+ \langle s_i\rangle\sum_{l\neq i}\sum_j K_{lj}^2K_{ji}^2L_lL_j^2, \end{aligned} \tag{17b}$$

where $L_i = 1-\langle s_i\rangle$ and $K_{ij} = \widetilde{C}_{ij}/L_iL_j$. Note that this approximation assumes small (connected) correlations, which may or may not be the case in neural data. In our testing, h^{SM} is substantially more sensitive to this assumption than J^{SM}. Still, the TAP and SM approximations, or their mean (Roudi, Aurell et al., 2009; Roudi, Tyrcha et al., 2009), generally give the best approximations (i.e., closest to Boltzmann learning). Figure 2.4 displays an example of SM versus Boltzmann learning parameters, illustrating the expected (Roudi, Aurell et al., 2009; Roudi, Tyrcha et al., 2009) underestimation for large N.

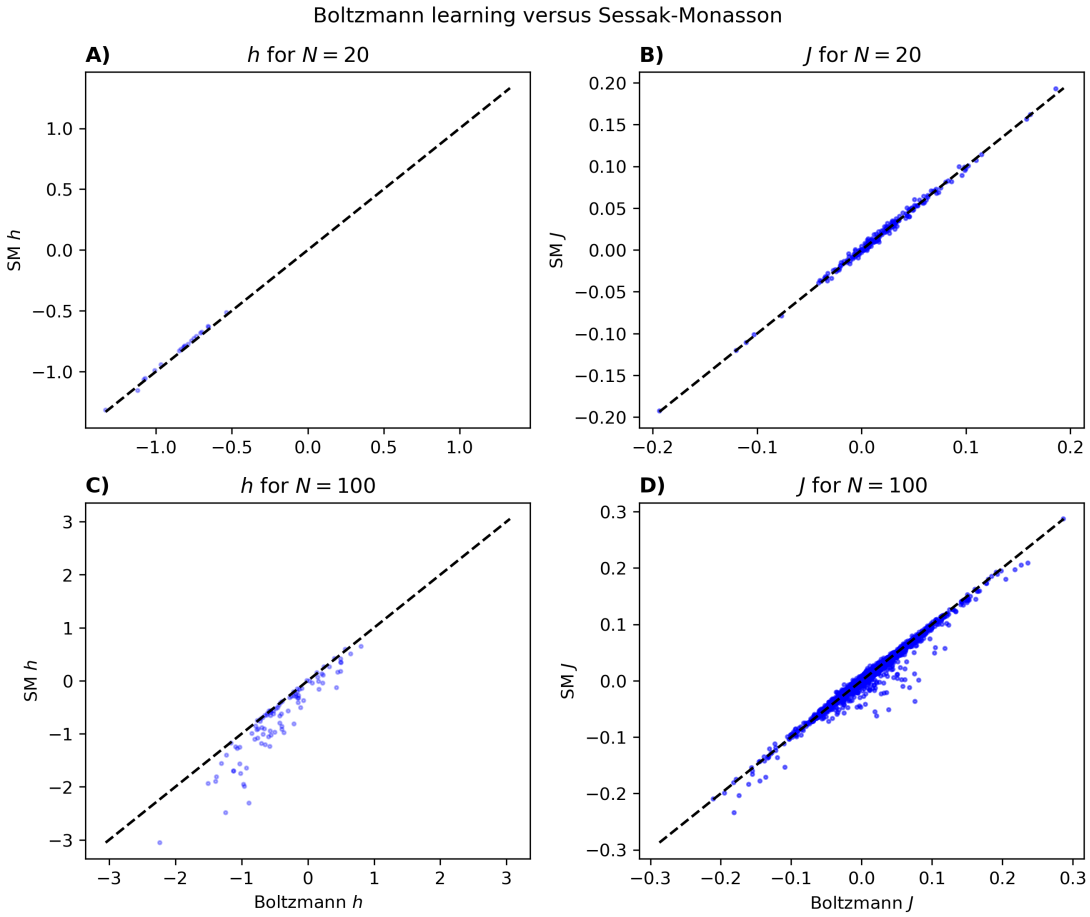


Figure 2.4: Boltzmann learning against SM for neural data. (A-B) A random subpopulation of 20 out of the 495 neurons were chosen. The SM parameters h_i^{SM} and J_{ij}^{SM} were plotted against the Boltzmann learning parameters h_i^{boltz} and J_{ij}^{boltz}, which used a learning rate of $\eta = 0.01$, 40000 iterations, and 10000 samples per iteration. (C-D) A random subpopulation of 100 out of the 495 neurons were chosen. The SM parameters h_i^{SM} and J_{ij}^{SM} were plotted against the Boltzmann learning parameters h_i^{boltz} and J_{ij}^{boltz}, which used a learning rate of $\eta = 0.01$, 80000 iterations, and 50000 samples per iteration. This figure shows that the SM approximation gives smaller parameters than Boltzmann learning, which is more apparent for larger N.

2.3.6 Pseudolikelihood maximization

An approximation that compromise between the reliability of Boltzmann learning and the speed of the approximate closed-form solutions would be ideal. Pseudolikelihood maximisation (PL; Besag, 1975) may fill this role. This approach decomposes the problem of finding the biases and couplings into N independent subproblems by considering the conditional distribution of each neuron s_i given all the others neurons $s_{/i}$:

$$p(s_i|s_{/i}) = \frac{\exp\left(s_i[h_i + \sum_{j\neq i} J_{ij}s_j]\right)}{2\cosh\left(h_i + \sum_{j\neq i} J_{ij}s_j\right)} = \frac{1}{1+\exp\left(-2s_i[h_i + \sum_{j\neq i} J_{ij}s_i]\right)} \tag{18}$$

The sum of these conditional distributions would replace the likelihood function and be maximised over h_i and J_{ij}, where the final parameters are denoted by h_i^{PL} and J_{ij}^{PL}. Equivalently, the condi-

tional distributions define N independent logistic regression problems, each resulting in an h_i (the zeroth coefficient) and a row i in the coupling matrix J (the coefficients in front of s_{ji}). This also means that the pseudolikelihood approach uses all the data, rather than just its means and correlations like in, for example, Boltzmann learning. This is typically advantageous, but leads to poor approximations if we have very few samples. Approximating the biases and couplings using pseudolikelihood is significantly faster than Boltzmann learning, and converge to the maximum likelihood values of h_i and J_{ij} in the limit of infinite samples. Additionally, pseudolikelihood is more reliable than the closed-form approximations where we might have data that violates assumptions (e.g., large J_{ij}s). As expected, we see that the Boltzmann learning and pseudolikelihood approximations of h and J are very similar (Figure 2.5). When not stated otherwise, pseudolikelihood has been used to approximate h and J from data.

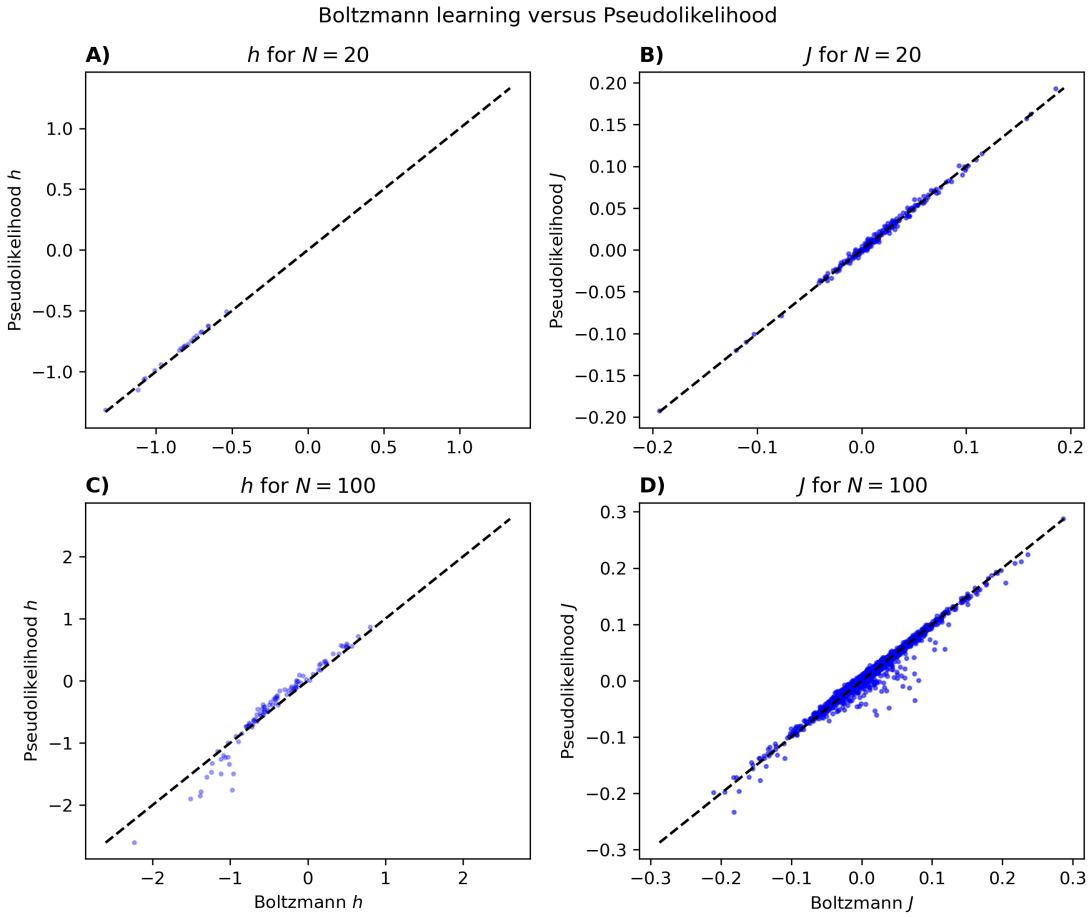


Figure 2.5: Boltzmann learning against pseudolikelihood maximization for neural data. (A-B) A random subpopulation of 20 out of the 495 neurons were chosen. The pseudolikelihood parameters h_i^{PL} and J_{ij}^{PL} were plotted against the Boltzmann learning parameters h_i^{boltz} and J_{ij}^{boltz}, which used a learning rate of $\eta = 0.01$, 40000 iterations, and 10000 samples per iteration. (C-D) A random subpopulation of 100 out of the 495 neurons were chosen. The pseudolikelihood parameters h_i^{PL} and J_{ij}^{PL} were plotted against the Boltzmann learning parameters h_i^{boltz} and J_{ij}^{boltz}, which used a learning rate of $\eta = 0.01$, 80000 iterations, and 50000 samples per iteration. This figure shows that Boltzmann learning and pseudolikelihood find almost the same parameters for a population of 20 and 100 neurons.

2.4 Assessing Performance

After fitting the pairwise model, we want to know how well it captures the data. That is, we want to assess how similar the pairwise distribution p_{pair} with the approximated parameters h and J is to the true distribution p_{true} that would emerge if we had infinite data. Of course we don't have infinite data, so p_{true} is often taken to be the frequency of each state in the data, denoted by p_{data}. The similarity is measured as the Kullback-Leibler (KL) divergence between the pairwise and true distribution:

$$d_{\text{pair}} \equiv D_{\text{KL}}(p_{\text{true}} \parallel p_{\text{pair}}) = \sum_{\mathbf{s}} p_{\text{true}}(\mathbf{s}) \ln \frac{p_{\text{true}}(\mathbf{s})}{p_{\text{pair}}(\mathbf{s})} = S_{\text{pair}} - S_{\text{true}}. \tag{19}$$

However, this quantity is difficult to interpret in isolation. We therefore compare it to the KL divergence between the independent and true distribution, given by

$$d_{\text{ind}} \equiv D_{\text{KL}}(p_{\text{true}} \parallel p_{\text{ind}}) = \sum_{\mathbf{s}} p_{\text{true}}(\mathbf{s}) \ln \frac{p_{\text{true}}(\mathbf{s})}{p_{\text{ind}}(\mathbf{s})} = S_{\text{ind}} - S_{\text{true}}. \tag{20}$$

This allows us to define a performance measure G,

$$G \equiv 1 - \frac{d_{\text{pair}}}{d_{\text{ind}}} = \frac{S_{\text{ind}} - S_{\text{pair}}}{S_{\text{ind}} - S_{\text{true}}}, \tag{21}$$

which increases from 0 to 1 as the true distribution is described better by the pairwise distribution than the independent distribution. Note that d_{pair}, d_{ind}, and G also can be expressed using entropies S as in Eq. (19), (20), and (21), given that the pairwise model fit is exact (i.e., $\langle s_i \rangle_{\text{pair}} = \langle s_i \rangle_{\text{data}}$ and $\langle s_i s_j \rangle_{\text{pair}} = \langle s_i s_j \rangle_{\text{data}}$; Roudi, Nirenberg et al., 2009). Notice also that G depends on the independent distribution p_{ind}, which can be defined either using the biases h_i^{ind} defined in Eq. (5) or the biases h_i inferred for the pairwise model (thus merely setting $J_{ij} = 0$). The former, h_i^{ind}, is traditionally used because it defines a maximum entropy distribution. However, it may be interesting to consider what happens to G when using h_i. That is, how the performance of the pairwise model changes when removing its couplings J_{ij}. Therefore, we consider both the traditional performance measure G, using h_i^{ind}, and another performance measure G^{RC} (removed couplings), using h_i.

2.5 Let's assess the performance then!

Several previous studies have used G (or $1-G$) to evaluate the pairwise model fitted to small (~ 10) neuronal populations (Chelaru et al., 2021; Ganmor et al., 2011a; Schneidman et al., 2006; Shlens et al., 2006; Tang et al., 2008; Yu et al., 2008; Zanoci et al., 2019), and find good performance ($G \approx 0.90$). This has been found in a variety of species and brain areas, including salamander retina, guinea pig retina, primate retina, cat visual cortex, rat and primate cortical cultures, and primate sensory and executive cortices. Similarly, large G have been found for small N (regions of interest, in this case) when applying the pairwise model to human fMRI data (Ezaki et al., 2017; Watanabe et al., 2013).

Given these promising findings, one might wonder whether it's possible to simplify the pairwise model by only considering some of the potential couplings J_{ij} while retaining the good performance, and thus describing the data with even fewer parameters. Such simplifications include only allowing

adjacent neurons to be connected (Shlens et al., 2009; Shlens et al., 2006), only considering large couplings (Ganmor et al., 2011a), and distributing all pairs of neurons into clusters sharing one coupling parameter (Ganmor et al., 2011a). These reduced pairwise models were all applied in the retina, so while they retained good performance, it is not clear whether this is a general phenomena. It is also not obvious whether the pairwise model will continue to reproduce the data distribution well as we consider more neurons. For example, in systems where one would expect higher-order correlations to be prominent, like natural images, the pairwise model might be sufficient for low dimensionality (few pixels), but not for large (Bethge and Berens, 2007). Unfortunately, G is difficult to calculate for larger N because the number of states grows exponentially. Thus, other methods have been used to evaluate the pairwise model for large N, such as comparing third-order correlations, the number of simultaneously active neurons, or the probability of highly sampled states (Ganmor et al., 2009, 2011a, 2011b; Meshulam et al., 2017, 2021; Shlens et al., 2009; Tkacik et al., 2006; Tkačik et al., 2014; Tkačik et al., 2009; Zanoci et al., 2019). Here, the results have been more mixed. The performance generally seems to be good, apart from when N becomes very large ($N \approx 120$; Tkačik et al., 2014) or the input is highly structured (Ganmor et al., 2011b). Worse performance have also been found when the neurons are anatomically close together compared to far apart (Ohiorhenuan et al., 2010), but others have found the opposite (Meshulam et al., 2017, 2021). Finally, larger G have been found in sensory cortical compared to executive cortical areas (Chelaru et al., 2021). Some have therefore searched for other constraints to maximum entropy models that improve performance in these conditions (in particular, for large N). These constraints include the probability of simultaneous silence (Shimazaki et al., 2015), the probability of highly sampled states (Ganmor et al., 2011b), and the probability of k simultaneously active neurons (Tkačik et al., 2014; Tkačik et al., 2013). While these extensions generally show good performance, they have also not been evaluated with G but with third-order correlations and/or the number of simultaneously active neurons.

The problem with these alternative performance measures, relative to G, is that we don't know what they are missing. That is, if they show poor performance we know that the model fits the data poorly, but if they show good performance we can't know that the model fits the data well because there could be large deviations in other higher-order correlations. Therefore, it would be preferable if we knew how G scales with N. Then we could make stronger claims about the adequacy of the pairwise maximum entropy model, and maximum entropy models generally, to account for neuronal data. In light of this, Roudi, Nirenberg et al. (2009) performed a perturbative expansion in $N\bar{v}\delta t$, where $\bar{v}$ is the mean firing rate and δt is the binsize. They showed that the quality of the pairwise model should be linear in $N\bar{v}\delta t$ when $N\bar{v}\delta t \ll 1$, regardless of what the true distribution is. This is in agreement with the performance of the pairwise model obtained in other studies (e.g., Schneidman et al., 2006; Tang et al., 2008; Yu et al., 2008). However, how G behaves outside of this perturbative regime, when $N\bar{v}\delta t > 1$, is less clear. This prevents us from making general statements about the performance of the pairwise model for neuronal data.

To rectify this, we here consider how well the pairwise maximum entropy model describes neuronal data and how this depends on the number of neurons, the firing rates, the binsize, and the cortical area. The first three questions are directly related to the predictions of Roudi, Nirenberg et al. (2009), while the last one is more interesting neuroscientifically. We will also look at how using suboptimal parameters from the nMF, TAP, IP, and SM approximation affects the scaling of G, and at how some additional performance measures compares to G. Finally, we will consider what the performance of the pairwise model might tell us about how a neural network functions.

We do this both using G, representing how the pairwise model compares to the best independent model, and G^{RC}, representing how the pairwise model compares to itself without connections. For brevity, we use G to refer to both G^{RC} and G until we consider them separately in Section 4.2 and 4.3, respectively.

3 Approximating G

There are two problematic steps in calculating G from Eq. (21). First, the partition function Z of the pairwise model is a sum over all states. This forces us to consider the case where N is small enough to evaluate this sum, and the case where N is not, separately. Second, one often takes $p_{\mathrm{true}} = p_{\mathrm{data}}$, which is of course not strictly true. The perhaps clearest example is that one can't say that unsampled states are impossible to observe after recording for only a couple of hours. This affects quantities calculated using p_{data} instead of p_{true}, such as entropies, KL divergences, and G (e.g., Panzeri et al., 2007). To rectify this, a finite sampling correction could be applied.

3.1 Finite sampling correction

To correct some sampling-dependent quantity K for finite sampling, different proportions $r \in \{1/2, 1/5, 1/10, 1\}$ of the data samples were used, resulting in T samples. When $r < 1$, mutually exclusive proportions of the data were used to calculate K, and the mean $\bar{K}$ of these was used further. To be clear, this means that if K depends on parameters inferred from the data, such as h and J, these were inferred using only some proportion of the data. Now, K and $\bar{K}$ were fit to a second-order polynomial in $1/T$. Taking the limit $T \to \infty$, giving the intercept, results in the corrected value of K (Strong et al., 1998).

This correction is generally small, but in the expected direction. That is, S becomes larger, d_{pair} and d_{pair} becomes smaller, and G becomes larger. Because the correction is small and predictable, most of our results are uncorrected, but see Section 4.2.8 for an elaboration on the finite sampling bias. This allows for a more thorough analysis, as the correction is computationally expensive.

3.2 Performance for small N – summing over all states

For small N one can calculate the partition function Z of the pairwise distribution by summing over all 2^N states. This can then be used to calculate d_{pair}, and thus G, by summing over all sampled states $\hat{\mathbf{s}}$ (the unsampled states $\mathbf{s}$ have $p_{\mathrm{data}}(\mathbf{s}) = 0$). Note that we can sum over all sampled states even for large N. Here, we calculate Z exactly for up to $N = 20$.

This method can only be used for small N, but it can be used for arbitrary binsize δt and mean firing rate $\bar{v}$. Therefore, we will consider the effect of changing δt and $\bar{v}$ in subpopulations of $N = 20$ neurons. It is straight-forward to bin the spike times with whatever binsize we want. However, getting the mean firing rate we want is more difficult because we have to pick subpopulations of 20 out of the 495 neurons that give a mean firing rate close to the desired one. When choosing a subpopulation, we first pick one neuron i, and then pick remaining neurons j with a similar firing rate to the first one. Because we have substantially more neurons with small firing rates than

with large ones, the first neuron is picked with a probability proportional to its firing rate. The remaining neurons are picked with a probability inversely proportional to the difference between their and the first neuron's firing rate. That is, the remaining neurons j were picked (without replacement) with probability $p(j) = \frac{1}{|v_i - v_j|^3} / \sum_k^N \frac{1}{|v_i - v_k|^3}$, where the exponent controls the spread of firing rates within each subpopulation. This procedure was found to produce subpopulations that spread out nicely along the range of possible mean firing rates. Then, G was calculated for these subpopulations, with some binsize, by summing over all states.

3.3 Performance for large N – approximating Z

As N becomes large the number of states becomes so enormous that it is intractable to sum over all of them. One solution would be to simply measure performance in some other way, such as comparing the third-order correlations or the number of simultaneously active neurons (Tkačik et al., 2014; Tkačik et al., 2009). Another approach would be to approximate Z, S_{pair}, or d_{pair}, either of which would allow us to approximate G. Doing this, does not explicitly account for effect of finite sampling. That is, there may still be a bias in S_{pair} that we ignore by considering d_{pair} with $p_{\text{true}} = p_{\text{data}}$. However, this does not seem to have a significant effect on the approximation of G, as discussed in Section 4.2.8 and 4.3.7. Here, we consider a new way to approximate the partition function Z of the pairwise model. Note that we can calculate p_{ind}, and thus d_{ind}, without approximating Z_{ind} because the neurons are independent.

The idea is to find the $\hat{Z}$ that gets the unnormalized probabilities of the pairwise distribution as close to the probabilities/frequencies in the data as possible, only considering sampled states $\hat{\mathbf{s}}$. More precisely, we minimise

$$L \equiv \sum_{\hat{\mathbf{s}}} (p_{\text{data}}(\hat{\mathbf{s}}) - \frac{1}{\hat{Z}} \exp\,[\sum_i h_i s_i + \sum_{i<j} J_{ij} s_i s_j])^2 \tag{22}$$

over $\hat{Z}$. Note that the resulting pairwise distribution don't sum to 1 exactly. The approximation of Z allows us to approximate $p_{\text{pair}}(\mathbf{s})$ which can be used to approximate d_{pair}, and thus G, by summing over all sampled states. This approximation can be used for both G and G^{RC}. Finding $\hat{Z}$ is a convex one-variable minimization problem with simple function evaluations that can be solved numerically with, for example, Brent's algorithm (Brent, 1971). However, simply taking the derivative of Eq. (22) with respect to $\hat{Z}$, setting it equal to 0, and solving for $\hat{Z}$, yields an analytical expression for $\hat{Z}$:

$$\hat{Z} \equiv \frac{\sum_{\hat{\mathbf{s}}} \exp\,[2\sum_i h_i s_i + 2\sum_{i<j} J_{ij} s_i s_j]}{\sum_{\hat{\mathbf{s}}} p_{\text{data}}(\hat{\mathbf{s}}) \exp\,[\sum_i h_i s_i + \sum_{i<j} J_{ij} s_i s_j]}. \tag{23}$$

This allows for rapid evaluations of $\hat{Z}$, and thus $\hat{G}$. In a sense, this procedure evaluates the performance of the pairwise model as charitably as possible, as it maximises the similarity between the pairwise and data distribution. One can think about this as comparing the shape, rather than the exact probabilities, of the pairwise and data distribution. However, for $\hat{Z}$ to be close to Z, the real pairwise model should not get closer to the data it was inferred from by dividing by some number (by "changing" $\hat{Z}$). This suggests that the approximation of the parameters h and J have

to be close to optimal (i.e., producing a maximum entropy distribution fulfilling Eq. (6) to get a good approximation of Z, and in turn, of G. To substantiate this suspicion, we look at the difference between $\hat{G}$ and G (and $\hat{G}^{\mathrm{RC}}$ and G^{RC}) as a function of Boltzmann learning iterations in Figure 3.1.

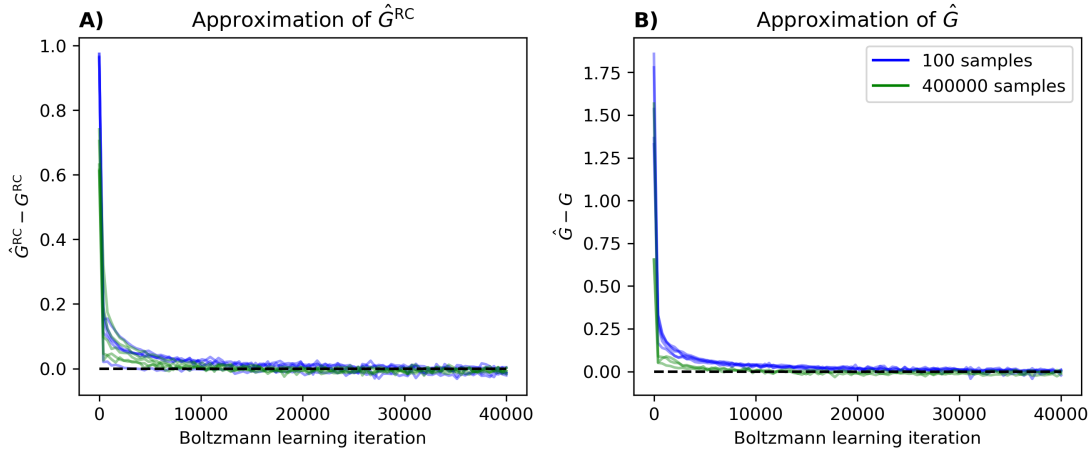


Figure 3.1: $\hat{G} - G$ as a function of Boltzmann learning iteration for synthetic data. Five sets of 400000 and 100 samples were taken from a pairwise model of 20 spins with Gaussian parameters ($M = 0$, $SD = 1/\sqrt{N-1}$). Then, new pairwise models were inferred from these datasets using Boltzmann learning with a learning rate of $\eta = 0.001$, 40000 iterations, and 5000 samples per iteration. $\hat{G} - G$ (A) and $\hat{G}^{\mathrm{RC}} - G^{\mathrm{RC}}$ (B) was only calculated for 100 sets of parameters between 0 and 40000 iterations to save time. The magnitude of the fluctuations around 0 after convergence is largely controlled by the learning rate η. In (B), only three sets of 400000 were used. This figure shows that $\hat{G}$ becomes very similar to G as the parameters h and J are better approximated.

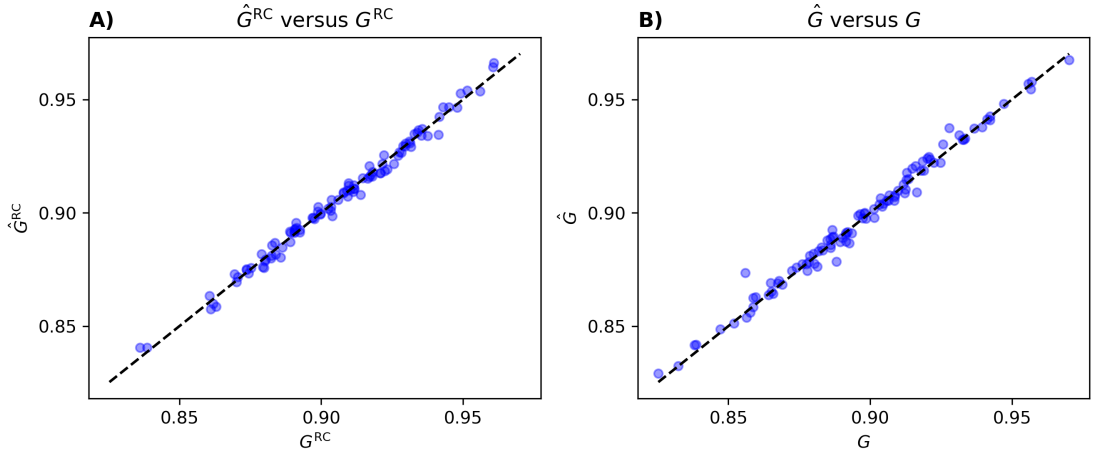


Figure 3.2: Scatterplot of $\hat{G}$ and $\hat{G}^{\mathrm{RC}}$ versus G using pseudolikelihood on synthetic data. Each of the 100 dots corresponds to the performance of a pairwise model of 20 spins inferred from 400000 samples of an arbitrary pairwise model with Gaussian parameters ($M = 0$, $SD = 1/\sqrt{N-1}$). Pseudolikelihood was used to approximate h and J. G is compared to $\hat{G}$ (A) and G^{RC} is compared to $\hat{G}^{\mathrm{RC}}$. This figure shows that $\hat{G}$ is very similar to G for $N = 20$.

Importantly, the number of data samples used to approximate the pairwise model seems to be irrelevant for the fast convergence of $\hat{G} - G$ (and $\hat{G}^{\mathrm{RC}} - G^{\mathrm{RC}}$). Thus, even severe undersampling should not be a problem as long as we have good parameter approximations. It is also worth noting

that using suboptimal parameters typically leads to an overestimation of G. In our testing this is often, but not necessarily, the case. This may be because the probabilities of highly sampled states are more underestimated by the pairwise model than the probabilities of rarely sampled states are overestimated, resulting in an underestimation of Z due to the squared error in Eq. (22), and finally an overestimation of G. As a sanity check, Figure 3.2 compares $\hat{G}$ and G directly.

It is worth commenting that calculating $\hat{G}$ from $\hat{Z}$ for very small N ($<\sim 5$) occasionally results in a $\hat{G}$ larger than 1. To avoid this, one could let $\hat{Z}_{\text{min}} = \sum_{\hat{\mathbf{s}}} \exp\left[\sum_i h_i s_i + \sum_{i<j} J_{ij} s_i s_j\right] \leq Z$ be a lower bound on $\hat{Z}$. That is, if $\hat{Z}$ from Eq. (23) turns out to be smaller than $\hat{Z}_{\text{min}}$, we set $\hat{Z} = \hat{Z}_{\text{min}}$. A second way of avoiding the occasional $\hat{G}$ larger than 1 would be to simply sum over all states when N is small enough, for example when $N \leq 15$. This is what we do in Section 4.

Another way to approximate Z relies on the assumption that the silent state ($s_i = -1$ for all i) is approximated well by the pairwise model because it typically is highly sampled in the data. This allows us to simply take $\hat{Z} = \exp\left[\sum_i h_i s_i + \sum_{i<j} J_{ij} s_i s_j\right] / p_{\text{data}}(\mathbf{s})$ where $\mathbf{s}$ is the silent state (Ashourvan et al., 2021; Ganmor et al., 2011a; Tkačik et al., 2014). This is analogous to the above procedure in that it finds the $\hat{Z}$ that get the probability of the silent state from the pairwise distribution as close to its frequency in the data as possible (by equating them). That is, instead of considering all sampled states, we only consider the silent state. This is obviously a good approximation when the probability of the silent state is a constraint in the maximum entropy model (Tkačik et al., 2014). But in the context of the pairwise model (Ashourvan et al., 2021; Ganmor et al., 2011a), this approximation becomes unreliable because the probability of the silent state frequently is underestimated.

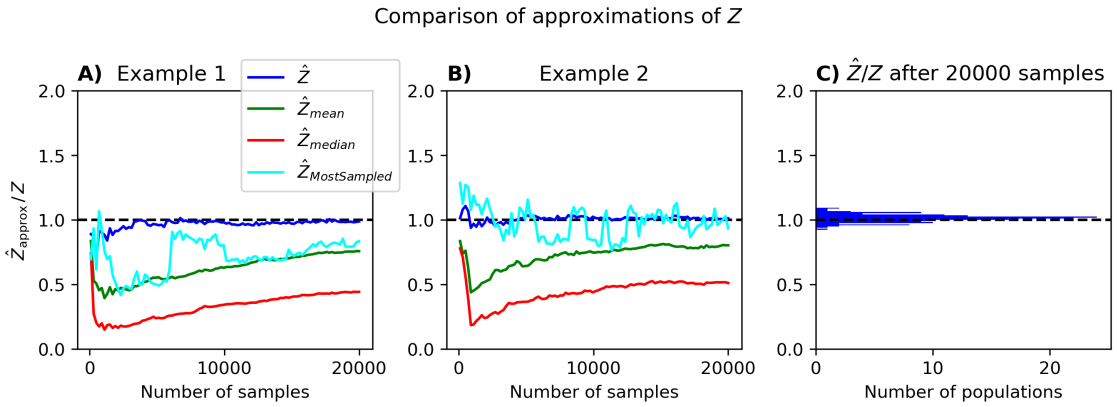


Figure 3.3: Comparison of $\hat{Z}$ (Eq. 23) and approximations based on $\hat{Z} = \exp\left[\sum_i h_i s_i + \sum_{i<j} J_{ij} s_i s_j\right] / p_{\text{data}}(\mathbf{s})$. A total of 20000 samples were taken from pairwise models with 15 spins and Gaussian parameters ($M = 0$, $SD = 1/\sqrt{N-1}$). The approximations of Z were calculated using only some of the samples, where the number of included samples range from 100 to 20000 and only 100 evenly spaced numbers in this range were chosen. The h and J that go into approximating Z were estimated using Boltzmann learning with a learning rate of $\eta = 0.001$, 30000 iterations, and 5000 samples per iteration. This procedure was performed for a total of 150 sets of Gaussian parameters, two examples of which are shown in (A) and (B). The ratio of the approximated Z to the actual Z is displayed for $\hat{Z}$ (blue), $\hat{Z}_{\text{mean}}$ (green), $\hat{Z}_{\text{median}}$ (red), and $\hat{Z}_{\text{MostSampled}}$ (turquoise). (C) The ratio $\hat{Z}/Z$ after 20000 samples for all 150 sets of Gaussian parameters. This figure shows that our approximation of Z outperforms a previously used approximation, and extensions thereof.

We can generalize the approximation based on the silent state somewhat. First, instead of

deciding ahead of time that we approximate Z from the silent state we can use the most sampled state (which is the silent state in most cases), giving $\hat{Z}_{\text{MostSampled}}$. Second, we can consider the approximation $\hat{Z} = \exp\left[\sum_i h_i s_i + \sum_{i<j} J_{ij} s_i s_j\right]/p_{\text{data}}(\mathbf{s})$ where $\mathbf{s}$ based on every sampled state $\mathbf{s}$ and use their mean or median as our approximation $\hat{Z}_{\text{mean}}$ or $\hat{Z}_{\text{median}}$. A comparison of $\hat{Z}$ and these approximations is displayed in Figure 3.3.

Eq. (23) is constructed so that the probabilities of the sampled states are as similar as possible in the data and the pairwise model. The obvious way the approximation could fail is that the unsampled states are very dissimilar in the data and the pairwise model. While the unsampled states have a probability of zero in the data distribution, they have a non-zero probability in the pairwise model. Therefore, $\hat{Z}$ can fail if the unsampled states carry a substantial amount of the probability in the pairwise model. Then, the approximated pairwise distribution $\hat{p}_{\text{pair}}(\mathbf{s})$ will have a sum larger than one. So, one might expect that $\hat{Z}$ is a good approximation when the pairwise distribution does not have too much entropy (i.e., is too flat). Then a minority of the possible states will carry the majority of the probability. Thus, $\hat{Z}$ should be progressively better for more constrained maximum entropy models.

Eq. (22) suggests another performance measure that may be informative. We can compare the minimum value of L for the pairwise and independent distribution:

$$G_L \equiv 1 - \frac{L_{\text{pair}}}{L_{\text{ind}}}. \tag{24}$$

Remember that L_{pair} and L_{ind} is the squared error between the data distribution, and the approximated pairwise and independent distribution, respectively. Like for G, G_L increases from 0 to 1 as the pairwise distribution describe the data distribution better than the independent distribution. For G_L to stay between 0 and 1, L_{ind} must be larger than or equal to L_{pair}, which should always be the case because p_{pair} is a more constrained maximum entropy distribution than p_{ind}. Looking at G_L may be a sensible thing to do because it, like G, measure how close p_{pair} is to p_{data}, relative to p_{ind}. Therefore, we would expect that $\hat{G}$ and G_L show a similar scaling with the perturbative parameter. However, G_L offer some advantages over $\hat{G}$, such as being marginally faster to compute and not caring about whether Z is over- or underestimated. Remember that we, like for G, define $\hat{G}^{\text{RC}}$ and G_L^{RC} the same way as $\hat{G}$ and G_L, except that h is used in place of h^{ind}.

3.4 What about other performance measures?

Due to the difficulty of evaluating G for large N, it has been suggested that performance could instead be evaluated by comparing the third-order interactions in the data with those in the inferred pairwise model (Ganmor et al., 2011b; Tkacik et al., 2006; Tkačik et al., 2014; Tkačik et al., 2009). The third-order correlation coefficient C_{ijk} for each of the $N(N-1)(N-2)/3!$ distinct triplets of neurons is defined as

$$C_{ijk} \equiv \langle s_i s_j s_k \rangle, \tag{25}$$

analogous to the second-order (i.e., pairwise) correlation coefficient. While some use this directly (Tkacik et al., 2006; Tkačik et al., 2009), others correct for the influence of lower-order correlations by considering the connected third-order correlation coefficient

$$\widetilde{C}_{ijk} \equiv \langle (s_i - \langle s_i \rangle)(s_j - \langle s_j \rangle)(s_k - \langle s_k \rangle) \rangle \tag{26}$$

instead (Tkačik et al., 2014). It is worth noting that when comparing C_{ijk} or $\widetilde{C}_{ijk}$ in the data and the pairwise model, they distribute differently around the diagonal, an example of which is displayed in Figure 3.4. This difference holds both for small (~ 5) and large (~ 100) N. Notice in particular that $\widetilde{C}_{ijk}$ centres around zero. This is expected because the pairwise model, by definition, does not include any third-order correlations, beyond those induced by first- and second-order correlations. To the extent that these are corrected for by $\widetilde{C}_{ijk}$, we would expect $\widetilde{C}^{\text{pair}}_{ijk}$ to stay around zero, ignoring sampling noise. This argument hints that a comparison of $\widetilde{C}_{ijk}$s may not have much utility as a performance measure.

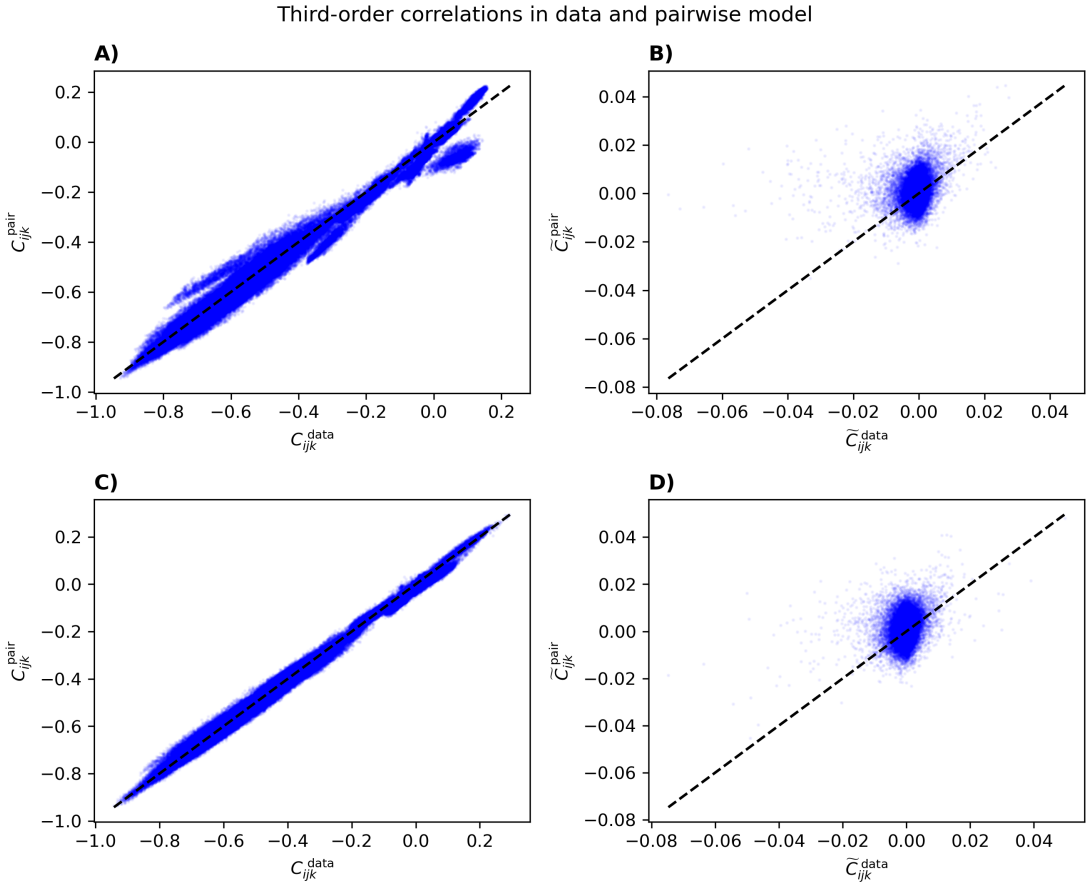


Figure 3.4: Example comparisons of third-order correlations and connected third-order correlations from the data and the inferred pairwise model. Two random subpopulations of 100 neurons were chosen. The pairwise model was inferred using pseudolikelihood maximization before it was sampled using the Metropolis-Hastings algorithm (as many samples as in the data). The third-order correlations were then calculated from these samples using Eq. (25) (A and C), and the connected third-order correlations using Eq. (26) (B and D). (A) and (B) displays the first example subpopulation, while (C) and (D) displays the second. This figure illustrates the difference between how C_{ijk} and $\widetilde{C}_{ijk}$ assess the performance of the pairwise model

The more fundamental problem with using third-order correlations to measure performance, as mentioned previously, is that even if the third-order correlations match well, we cannot say that the pairwise and data distributions match well. Here, we will attempt to determine whether this is the case and thus whether similarity of third-order correlations could be a suitable proxy for G. To that end, we define a performance measure based on the root mean squared error of third-order

correlations as

$$G_C \equiv 1 - \sqrt{\frac{\sum_{i<j<k}(C_{ijk}^{\text{data}} - C_{ijk}^{\text{pair}})^2}{\sum_{i<j<k}(C_{ijk}^{\text{data}} - C_{ijk}^{\text{ind}})^2}} \tag{27a}$$

$$G_{\widetilde{C}} \equiv 1 - \sqrt{\frac{\sum_{i<j<k}(\widetilde{C}_{ijk}^{\text{data}} - \widetilde{C}_{ijk}^{\text{pair}})^2}{\sum_{i<j<k}(\widetilde{C}_{ijk}^{\text{data}} - \widetilde{C}_{ijk}^{\text{ind}})^2}}, \tag{27b}$$

where C_{ijk}^{data} and $\widetilde{C}_{ijk}^{\text{data}}$ are calculated from the data, C_{ijk}^{pair} and $\widetilde{C}_{ijk}^{\text{pair}}$ are calculated from samples of the pairwise model, and C_{ijk}^{ind} and $\widetilde{C}_{ijk}^{\text{ind}}$ are calculated from samples of the independent model. This performance measure is analogous and directly comparable to G.

Performance of the pairwise model has also been evaluated by comparing the number of simultaneously active neurons in the data and the pairwise model (Ganmor et al., 2009, 2011a, 2011b; Tkacik et al., 2006; Tkačik et al., 2014; Tkačik et al., 2013; Tkačik et al., 2009). That is, if the probability of m arbitrary neurons firing, denoted by H_m, is similar in the data and the pairwise model, one might suspect that the pairwise model accounts well for higher-order correlations. However, like for third-order correlations, we cannot know this. Therefore, we define a performance measure based on the root mean squared error of H_m that, again, is analogous to G:

$$G_H \equiv 1 - \sqrt{\frac{\sum_m (H_m^{\text{data}} - H_m^{\text{pair}})^2}{\sum_m (H_m^{\text{data}} - H_m^{\text{ind}})^2}}. \tag{28}$$

Here, H_m^{data} is calculated from the data, H_m^{pair} is calculated from samples of the pairwise model, and H_m^{ind} is calculated from samples of the independent model. For intuition, an example of how H_m^{pair} compares to H_m^{data} is displayed in Figure 3.5.

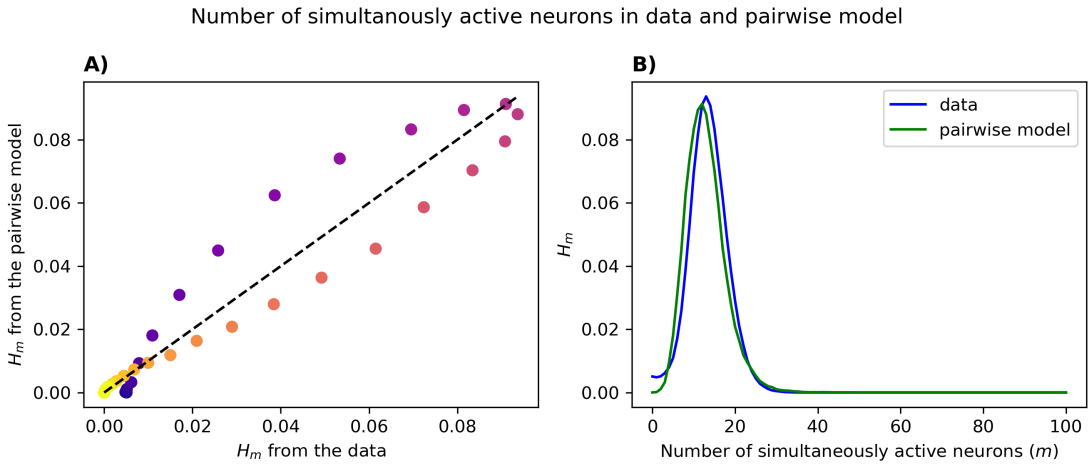


Figure 3.5: Example comparison of the number of simultaneously active neurons in the data and the pairwise model. A random subpopulation of 100 neurons were chosen. The pairwise model was inferred using pseudolikelihood maximization before it was sampled using the Metropolis-Hastings algorithm (as many samples as in the data). The H_ms were then calculated from these samples and the data. (A) Scatterplot of H_m from the data and pairwise model, where small m is represented by dark blue and larger m becomes light yellow (where $m > 30$ is all yellow). (B) H_m as a function of m, showing how the pairwise model systematically over- and underestimates H_m. This figure shows an example comparison of the number of simultaneously active neurons in the data and the pairwise model.

Again, G_C^{RC}, $G_{\widetilde{C}}^{\text{RC}}$, and G_H^{RC} is defined the same way as G_C, $G_{\widetilde{C}}$, and G_H, except that h is

used in place of h^{ind}.

4 Performance of the Pairwise Model on Neuronal Date

4.1 Preprocessing of data

We use a Neuropixel dataset recorded from the visual, auditory, somatosensory, and motor cortices of freely moving rats (Mimica et al., 2022). Each ~ 20 minute session consisted of the rat foraging in an octagonal ($2 \times 2 \times 0.8$ m) arena in dim light, in darkness, with a small weight attached to the implant, or with random-interval white noise playing. Here, we mainly consider the neurons shared across six such sessions recorded from the same probe in the same animal on the same day. This results in about 2 hours of data from $N = 495$ neurons, 130 of which are from auditory cortices, and the remaining 365 neurons are from visual cortices. These sessions were concatenated and binned with binsize δt. When not stated otherwise, a binsize of $\delta t = 0.02$ seconds is used, giving about 450000 bins or samples. In Section 4.2.7 the performance of the pairwise model in visual, auditory, somatosensory, and motor cortices will be compared. For the comparison, we used data from four sessions recorded from the same probe in the same animal on the same day. These sessions, constituting about 1 hour and 20 minutes, were concatenated and binned like above. $N = 539$ neurons were recorded from visual cortices, $N = 376$ from auditory cortices, $N = 287$ from somatosensory cortices, and $N = 1115$ from motor cortex.

4.2 $G^{\mathbf{RC}}$ – the Pairwise Model Compared to the Independent Model using h_i

Probably the simplest way of investigating how the performance of the pairwise model scales with N, $\bar{\nu}$, and δt is to simply plot the performance measure G^{RC} or $\hat{G}^{\text{RC}}$ from many subpopulations as a function of the perturbative parameter $N\bar{\nu}\delta t$. Thus, that will be the format of most of the following results.

4.2.1 Changing N — small N

For small enough populations, $N <\sim 20$, we can sum over all possible states and calculate the pairwise distribution exactly. As a first test, we select 100 populations of $N = 2, 3, ..., 20$ neurons randomly from the 495 neurons we have the most data on. Their parameters h and J were then approximated using pseudolikelihood according to Eq. (18). Finally, the performance of the resulting pairwise model was calculated from Eq. (21) and plotted against the perturbative parameter $N\bar{\nu}\delta t$ in Figure 4.1.

We see an approximately linear scaling of G^{RC} for $N\bar{\nu}\delta t \ll 1$. Interestingly, the linear scaling seems to continue up to $N\bar{\nu}\delta t \approx 2$. Also, the slope is fairly small, suggesting that the pairwise model is almost completely reliant on its couplings in achieving its good performance.

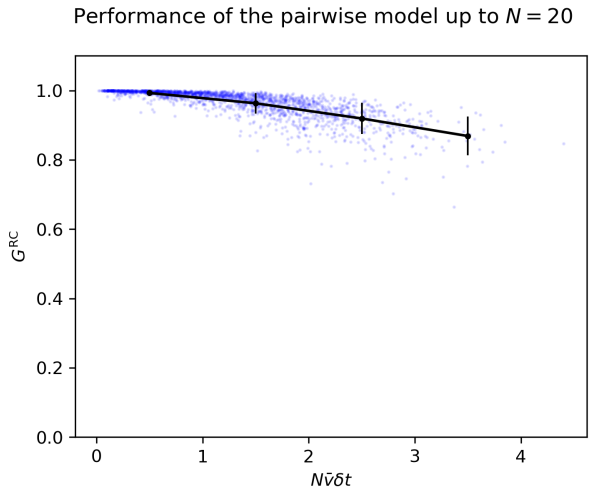


Figure 4.1: Performance of the pairwise model inferred with pseudolikelihood from neural data, for small N. 100 populations of size N were chosen randomly from the 495 neurons that were shared over 6 sessions, where N varied from 2 to 20. A constant binsize of $\delta t = 0.02$ was used. The mean firing rate $\bar{\nu}$ had a mean of $M = 6.04$ and standard deviation of $SD = 2.24$ over all populations. Pseudolikelihood was used to approximate h and J. G^{RC} was calculated by summing over all states. The black lines between the black dots represent the means while the black vertical lines represent standard deviations. This figure shows that G^{RC} initially scales linearly with $N\bar{\nu}\delta t$ when changing N.

4.2.2 Changing N — large N

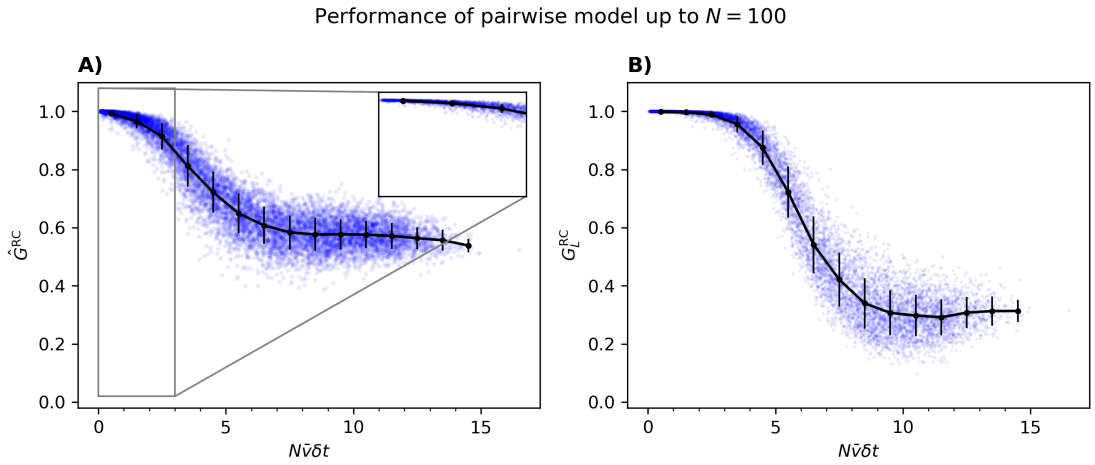


Figure 4.2: Performance of the pairwise model inferred with pseudolikelihood from neural data, for large N. 100 populations of size N were chosen randomly from the 495 neurons that were shared over 6 sessions, where N varied from 2 to 100. A constant binsize of $\delta t = 0.02$ was used. The mean firing rate $\bar{\nu}$ had a mean of $M = 6.07$ and standard deviation of $SD = 1.24$ over all populations. Pseudolikelihood was used to approximate h and J. The black lines between the black dots represent the means while the black vertical lines represent standard deviations. (A) For subpopulations consisting of 15 or fewer neurons, G^{RC} was calculated by summer over all states. For subpopulations with more than 15 neurons, $\hat{G}^{\mathrm{RC}}$ was calculated using $\hat{Z}$ from Eq. (22). (B) G_L^{RC} was calculated from Eq. (24) for all subpopulations. This figure shows that both $\hat{G}^{\mathrm{RC}}$ and G_L^{RC} has an initial linear scaling with $N\bar{\nu}\delta t$ followed by a sharp fall and a levelling off, although $\hat{G}^{\mathrm{RC}}$ level off before G_L^{RC}.

As discussed previously (Section 3.3), we need to approximate G^{RC} when N becomes large. Like for Figure 4.1, we plot performance as a function of the perturbative parameter in Figure 4.2A, except that $\hat{G}^{\mathrm{RC}}$ is used in place of G^{RC} and we pick 100 populations up to $N = 100$ rather than $N = 20$. The initial approximately linear scaling is still present, before $\hat{G}^{\mathrm{RC}}$ decreases sharply and levels off around $\hat{G}^{\mathrm{RC}} = 0.6$. This suggests that while the pairwise model might rely on mostly on its couplings when $N\bar{\nu}\delta t$ is small, its biases become more influential as $N\bar{\nu}\delta t$ increases. However, $\hat{G}^{\mathrm{RC}}$ does not fall to 0, indicating that removing the couplings from a pairwise model always makes it noticeably worse. The alternative performance measure G_L^{RC} (Eq. 24) was calculated for the same random populations with the same parameters as in Figure 4.2A, and is displayed in Figure 4.2B as a function of $N\bar{\nu}\delta t$. We see a qualitatively similar scaling of G_L^{RC} as of G^{RC}. That is, an initial approximately linear decrease, followed by a sharp fall before levelling off. The agreement between these two measures further suggest that the biases of the pairwise model really are more responsible for its performance as $N\bar{\nu}\delta t$ increases.

4.2.3 Changing $\bar{\nu}$ – semi-random 20-neuron populations

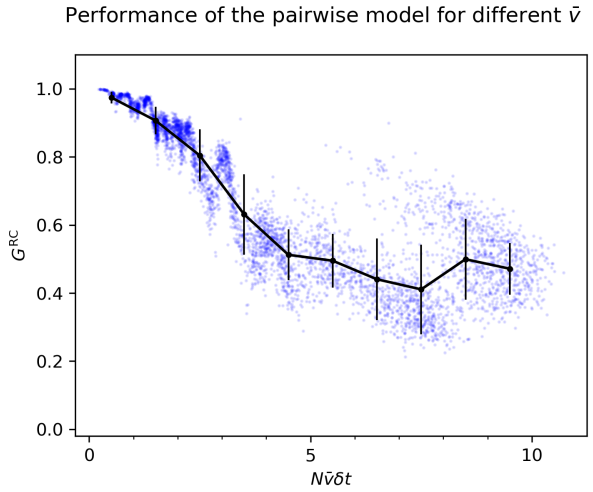


Figure 4.3: Performance of the pairwise model inferred with pseudolikelihood from neural data, for different $\bar{\nu}$. G^{RC} was calculated for 5000 subpopulations of 20 neurons, where neurons in the same subpopulation were chosen so that they have similar firing rates (see Section 3.2), by summing over all states. A constant binsize of $\delta t = 0.02$ was used. The mean firing rate $\bar{\nu}$ had a mean of $M = 11.22$ and standard deviation of $SD = 7.15$ over all populations. Pseudolikelihood was used to approximate the parameters h and J. The black lines between the black dots represent the means while the black vertical lines represent standard deviations. This figure shows that G^{RC} has a similar scaling with $N\bar{\nu}\delta t$ when changing $\bar{\nu}$ as seen in Section 4.2.2.

To investigating how performance scales with the mean firing rate $\bar{\nu}$ we choose non-random subpopulations out of the 495 neurons, like described in Section 3.2. In Figure 4.3 we pick 5000 such subpopulations of 20 neurons with similar firing rates, approximate their parameters using pseudolikelihood, and calculate G^{RC} by summing over all states.

We observe that increasing $\bar{\nu}$ gives approximately the same scaling of G^{RC} as increasing N, even though the latter was based on an approximate G^{RC}. One notable difference, however, is that the levelling off happens for a smaller value of G^{RC} and with a larger standard deviation than in Figure 4.2. Also, the somewhat spotty coverage likely is a result of us not being able to choose $\bar{\nu}$

directly.

4.2.4 Changing δt – random 20-neuron populations

Exploring how G^{RC} changes with binsize δt follows much the same procedure as above. Figure 4.4 displays 5000 random subpopulations of 20 neurons chosen from the 495 neurons. They were binned with a uniformly random binsize between 0.005 and 0.05 seconds, their parameters were approximated using pseudolikelihood, and G^{RC} was calculated by summing over all states. Also here, we see a similar pattern in the scaling of G^{RC}: an initial linear scaling followed by a sharp decrease and a levelling off. Interestingly, we see a similar levelling off for a smaller value of G^{RC} that we saw in Figure 4.3.

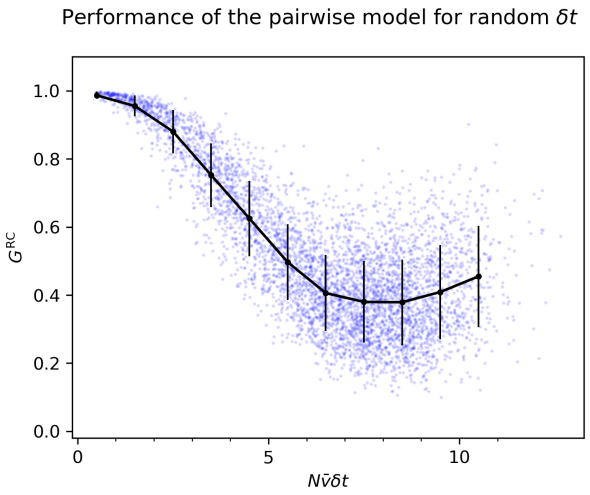


Figure 4.4: Performance of the pairwise model inferred with pseudolikelihood from neural data, for different δt. G^{RC} was calculated for 5000 subpopulations of 20 neurons, where the binsize δt was chosen uniformly between 0.005 and 0.2 seconds, by summing over all states. The mean firing rate $\bar{v}$ had a mean of $M = 3.79$ and standard deviation of $SD = 1.42$ over all populations. Pseudolikelihood was used to approximate the parameters h and J. The black lines between the black dots represent the means while the black vertical lines represent standard deviations. This figure shows that G^{RC} has a similar scaling with $N\bar{v}\delta t$ when changing δt as seen in Section 4.2.2 and 4.2.3.

4.2.5 Changing N, $\bar{v}$, and δt together

As a final test of the scaling of G^{RC} with the perturbative parameter, the number of neurons N, the mean firing rate $\bar{v}$, and the binsize δt were changed independently. This necessitated using $\hat{G}^{\mathrm{RC}}$ in place of G^{RC} again. Like in the previous two sections, Figure 4.5 shows the approximated $\hat{G}^{\mathrm{RC}}$ of 5000 subpopulations with parameters approximated using pseudolikelihood. Also here, we see a scaling reminiscent of the above results. We also see that $\hat{G}^{\mathrm{RC}}$ does not fall considerably as $N\bar{v}\delta t$ grows beyond $N\bar{v}\delta t \approx 15$. The somewhat sharp boundaries in Figure 4.5 are also notable, but they are likely a result of how N, $\bar{v}$, and δt were chosen.

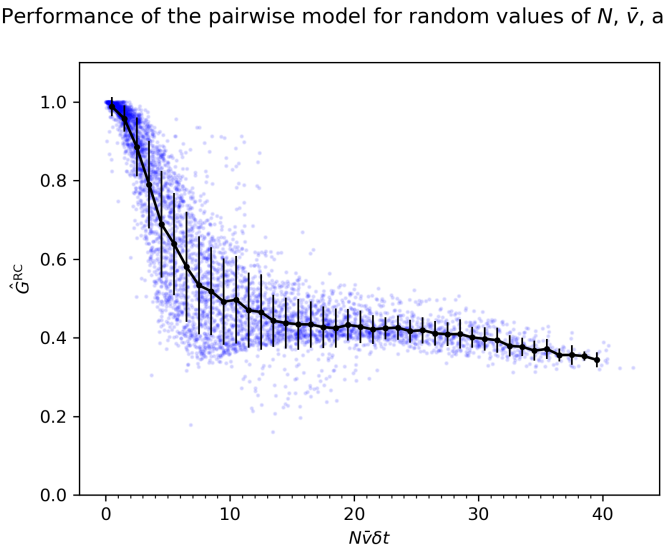


Figure 4.5: Performance of the pairwise model inferred with pseudolikelihood from neural data, for different N, $\bar{v}$, and δt. 5000 pairwise models were constructed and evaluated using uniformly random $N \in [2, 100]$, random $\delta t \in [0.005, 0.05]$, and semi-random mean firing rate $\bar{v}$ (see Section 3.2). The mean firing rate $\bar{v}$ had a mean of $M = 8.97$ and standard deviation of $SD = 5.54$ over all populations. $\hat{G}^{\mathrm{RC}}$ was calculated using $\hat{Z}$ from Eq. (23). The black lines between the black dots represent the means while the black vertical lines represent standard deviations. This figure shows that G^{RC} has a similar scaling with $N\bar{v}\delta t$ when changing N, $\bar{v}$, and δt independently as seen in Section 4.2.2, 4.2.3, and 4.2.4.

Another way to change the number of neurons N, the mean firing rate $\bar{v}$, and the binsize δt together, without relying on $\hat{G}^{\mathrm{RC}}$, is to choose non-random subpopulations with different $\bar{v}$ and of different sizes and then vary their binsize. In Figure 4.6 we choose subpopulations with similar firing rates, and let $N \in \{20, 10\}$ and $\delta t \in \{0.02, 0.02, 0.005\}$. 1500 subpopulations were picked for the six combinations of N and δt. Their parameters were approximated using pseudolikelihood and G^{RC} was calculated by summing over all states.

We again see the initial linear scaling followed by a drop-off. However, the levelling off is not apparent here, likely because $N\bar{v}\delta t$ does not become large enough. Still, we consistently find that larger values of N, $\bar{v}$, and δt results in a smaller G^{RC}. Although, here smaller binsizes do induce a somewhat larger G^{RC} than expected, as the smaller binsize dots exists in the top region of the larger binsize dots in Figure 4.6 (e.g., green on top of red).

The similarity between the scaling of G^{RC} in Figure 4.6, 4.3, and 4.4 and the scaling of $\hat{G}^{\mathrm{RC}}$ in Figure 4.2 and 4.5, suggests that our findings are robust and not an artifact of the approximation of Z. We find that the scaling with $N\bar{v}\delta t$ largely holds when changing N (Figure 4.1 and 4.2), $\bar{v}$ (Figure 4.3), δt (Figure 4.4), and all of them together (Figure 4.5 and 4.6).

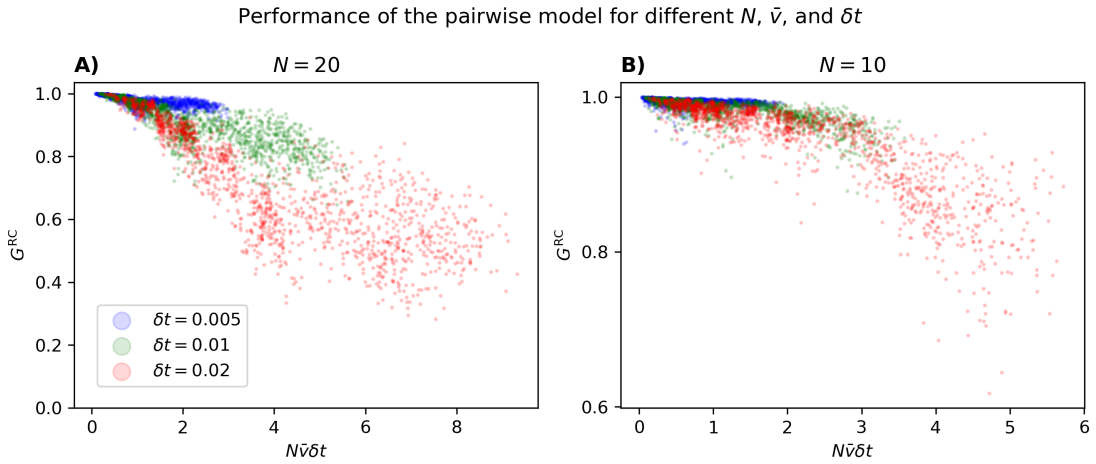


Figure 4.6: Performance of the pairwise model inferred with pseudolikelihood from neural data, for different N, $\bar{\nu}$, and δt. Subpopulations with semi-random mean firing rate $\bar{\nu}$ (see Section 3.2), $N \in \{20, 10\}$, and $\delta t \in \{0.02, 0.02, 0.005\}$ were chosen. 1500 subpopulations were picked for the six combinations of N and δt. G^{RC} was calculated by summing over all states and correcting for finite sampling. Pseudolikelihood was used to approximate the parameters h and J. (A) Subpopulations consists of $N = 20$ neurons. (B) Subpopulations consists of $N = 10$ neurons. This figure shows that G^{RC} has a similar scaling with $N\bar{\nu}\delta t$ as seen in Section 4.2.1-4.2.4 when changing N, $\bar{\nu}$, and δt together.

4.2.6 Performance using nMF, TAP, IP, or SM parameters

We have already seen that pseudolikelihood maximization is a good compromise between speed and accuracy when approximating h and J (Section 2.3.1-2.3.6; Nguyen et al., 2017). However, it might still be interesting to look at how using parameters derived using nMF, TAP, IP, or SM affects the scaling of $\hat{G}^{\mathrm{RC}}$. This is because inaccurate parameters can have different effects on the pairwise model they define. Typically, inaccuracies in parameters of small magnitude affects the probability of different states the more than inaccuracies in large-magnitude parameters. To investigate this, we here present results analogous to Figure 4.2, using nMF, TAP, IP, and SM rather than pseudolikelihood. Note that the same random subpopulations have been chosen for each approximation method. From previous investigations (Roudi, Aurell et al., 2009; Roudi, Tyrcha et al., 2009) we expect that TAP and SM are better than nMF and IP, and thus that using them should result in a scaling of $\hat{G}^{\mathrm{RC}}$ more similar to the pseudolikelihood case (Figure 4.2). Further, from Figure 3.1 we expect that inaccurate parameters results in a larger $\hat{G}^{\mathrm{RC}}$.

As a general trend, we see that using these closed-form approximations leads to an overestimation of $\hat{G}^{\mathrm{RC}}$, reflected in an earlier levelling-off, and a more uncertain $\hat{G}^{\mathrm{RC}}$, reflected in larger standard deviations. Notably, using SM parameters seems to be an exception to this trend. However, in this case an overflow error occurred during the calculation of $\hat{G}^{\mathrm{RC}}$ in 385 out of the 9800 subpopulations due to a vastly overestimated magnitude of some h_is, indicating that the small-J assumption of the SM approximation has been violated. Therefore, it is unclear how informative the later levelling-off is. We also notice that the IP approximation seems to be fairly stable in that $\hat{G}^{\mathrm{RC}}$ don't fall below 0. However, this approximation leads to G_L^{RC} becoming significantly more unstable as $N\bar{\nu}\delta t$ increases. These results show that using inaccurate parameters does have an effect on $\hat{G}^{\mathrm{RC}}$. Still, the stereotypical scaling found in Figure 4.1-4.6 persists.

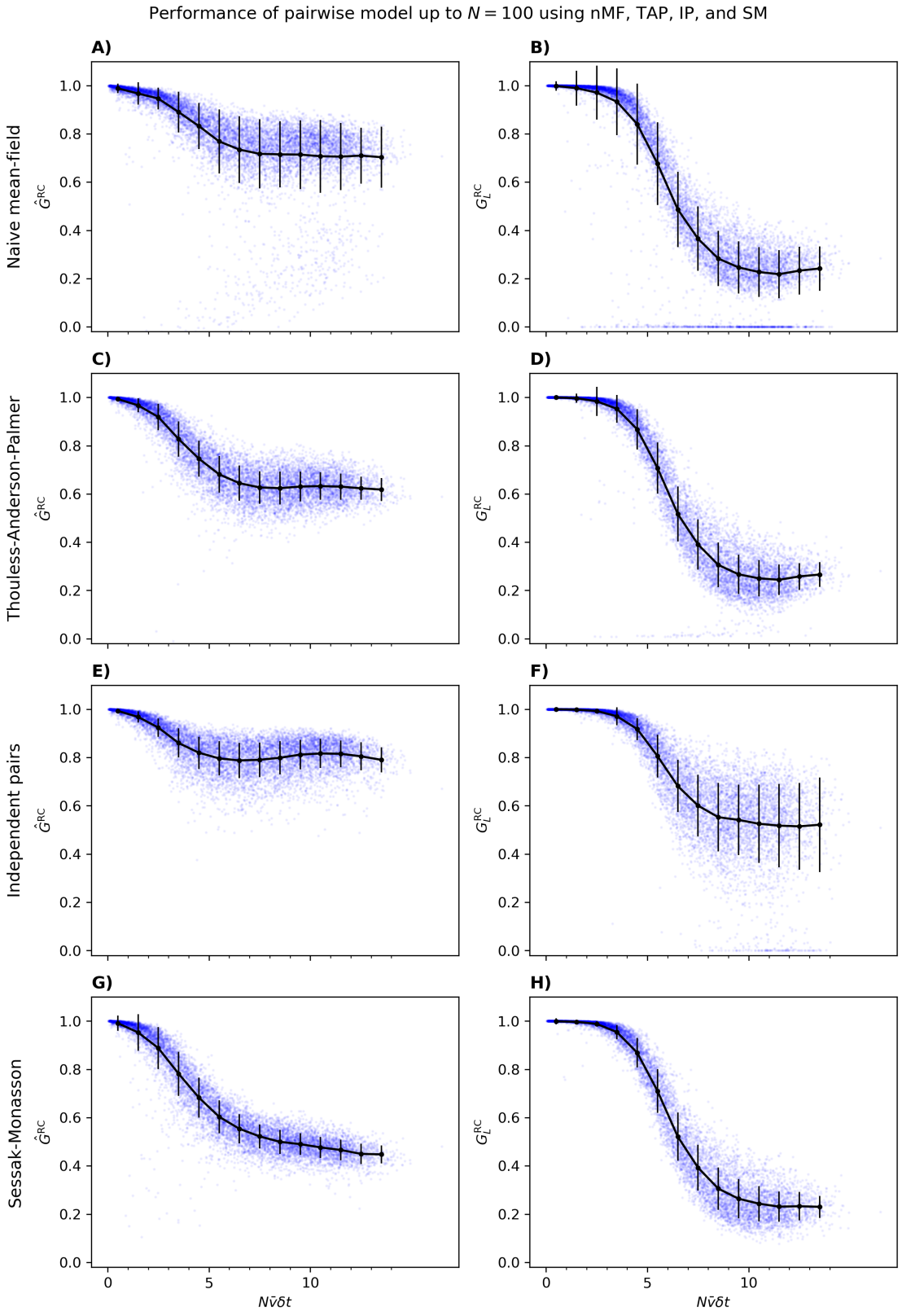


Figure 4.7: Performance of the pairwise model inferred with nMF, TAP, IP, and SM from neural data, for large N. This is identical to Figure 4.2 except that h and J have been approximated with nMF (A-B), TAP (C-D), IP (E-F), or SM (G-H) rather than pseudolikelihood. Using nMF, TAP, IP, and SM resulted in 35, 2, 0, and 5 outliers with $\hat{G}^{\mathrm{RC}} < 0$, respectively, which were omitted from the plots. Additionally, for the SM approximation (G-H), 385 (out of 9800) $\hat{G}^{\mathrm{RC}}$s were completely omitted due to overflow errors. This figure shows that inaccurate parameters does have an effect on $\hat{G}^{\mathrm{RC}}$, but the characteristic scaling persists.

4.2.7 Performance for different brain areas

Here, we follow the same format as for the different approximation methods and show how $\hat{G}^{\mathrm{RC}}$ scales with $N\bar{v}\delta t$ in recordings from rat visual, auditory, motor, and somatosensory cortices. Remember that the data we have looked at thus far also comes from visual and auditory cortices, so we expect other recordings from these areas to look familiar.

The scaling of $\hat{G}^{\mathrm{RC}}$ in visual and auditory cortex does indeed look similar to Figure 4.2. This similarity is more apparent for the visual cortex, which makes sense given that the majority (365/495) of the available neurons in Figure 4.2 were located in the visual cortex. However, in the auditory cortex the convergence was not clear with the usual 100 subpopulations with N varying between 2 and 100. This is due to a smaller firing rate $\bar{v}$ than in visual cortex, which makes $N\bar{v}\delta t$ smaller. Therefore, 50 additional subpopulations were added per N between 101 and 200. These extra subpopulations were also added for the motor (Figure 4.9A-B) and somatosensory (Figure 4.10A-B) cortices, for the same reason. Despite the larger subpopulations, the trend in $\hat{G}^{\mathrm{RC}}$ is still not clear in the motor and somatosensory cortex. In Figure 4.9A we see a drop off followed by convergence for the motor cortex, but the drop off is small as $\hat{G}^{\mathrm{RC}}$ converges to a larger value than usual (~ 0.85). For the somatosensory cortices in Figure 4.10A, the firing rate is so small that $N\bar{v}\delta t$ does not increase much above 8 for N up to 200, and the linear scaling of $\hat{G}^{\mathrm{RC}}$ persists. Additionally, because some neurons very rarely fire, we occasionally get very negative h_is, which leads to overflow errors when calculating $\hat{Z}$. To avoid this and get a better, and less computationally expensive, picture of how $\hat{G}^{\mathrm{RC}}$ scales with $N\bar{v}\delta t$ in the motor and somatosensory cortex, we increase the binsize from $\delta t = 0.02$ to $\delta t = 0.06$ in Figure 4.9C-D and to $\delta t = 0.14$ in Figure 4.10C-D.

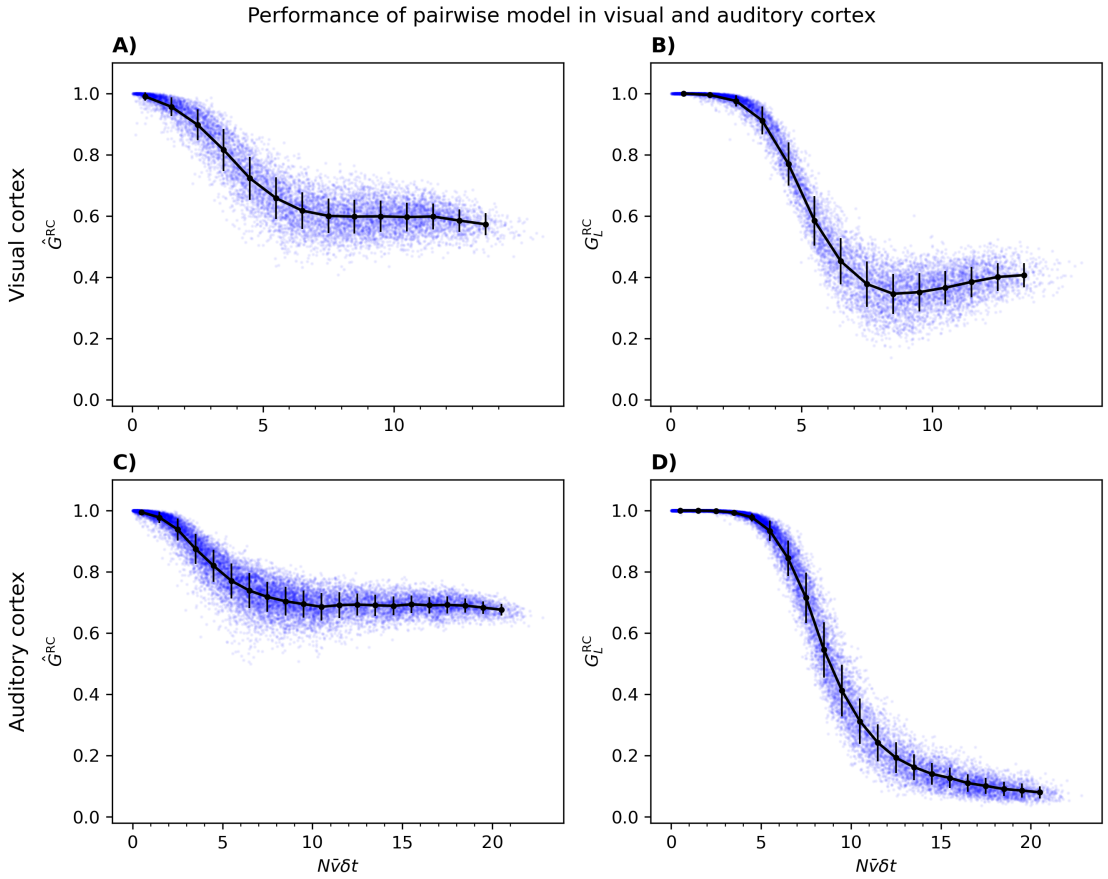


Figure 4.8: Performance of the pairwise model inferred with pseudolikelihood from neural data, for large N in visual and auditory cortex. For both cortical areas, 100 populations of size N were chosen randomly, where N varied from 2 to 100. A constant binsize of $\delta t = 0.02$ was used. Pseudolikelihood was used to approximate h and J. The black lines between the black dots represent the means while the black vertical lines represent standard deviations. For subpopulations consisting of 15 or fewer neurons, ${}^{\mathrm{RC}}G$ was calculated by summer over all states. For subpopulations with more than 15 neurons, $\hat{G}^{\mathrm{RC}}$ was calculated using $\hat{Z}$ from Eq. (22). G_L^{RC} was calculated from Eq. (24) for all subpopulations. (A-B) The mean firing rate $\bar{\nu}$ in the visual cortex had a mean of $M = 6.27$ and standard deviation of $SD = 1.24$ over all populations. (C-D) In addition to the populations up to $N = 100$, 50 populations per N between $N = 101$ and $N = 200$ were are also included. The mean firing rate $\bar{\nu}$ in the auditory cortex had a mean of $M = 4.99$ and standard deviation of $SD = 0.93$ over all populations. This figure again shows the characteristic scaling of $\hat{G}^{\mathrm{RC}}$ in the visual and auditory cortex, in different recordings.

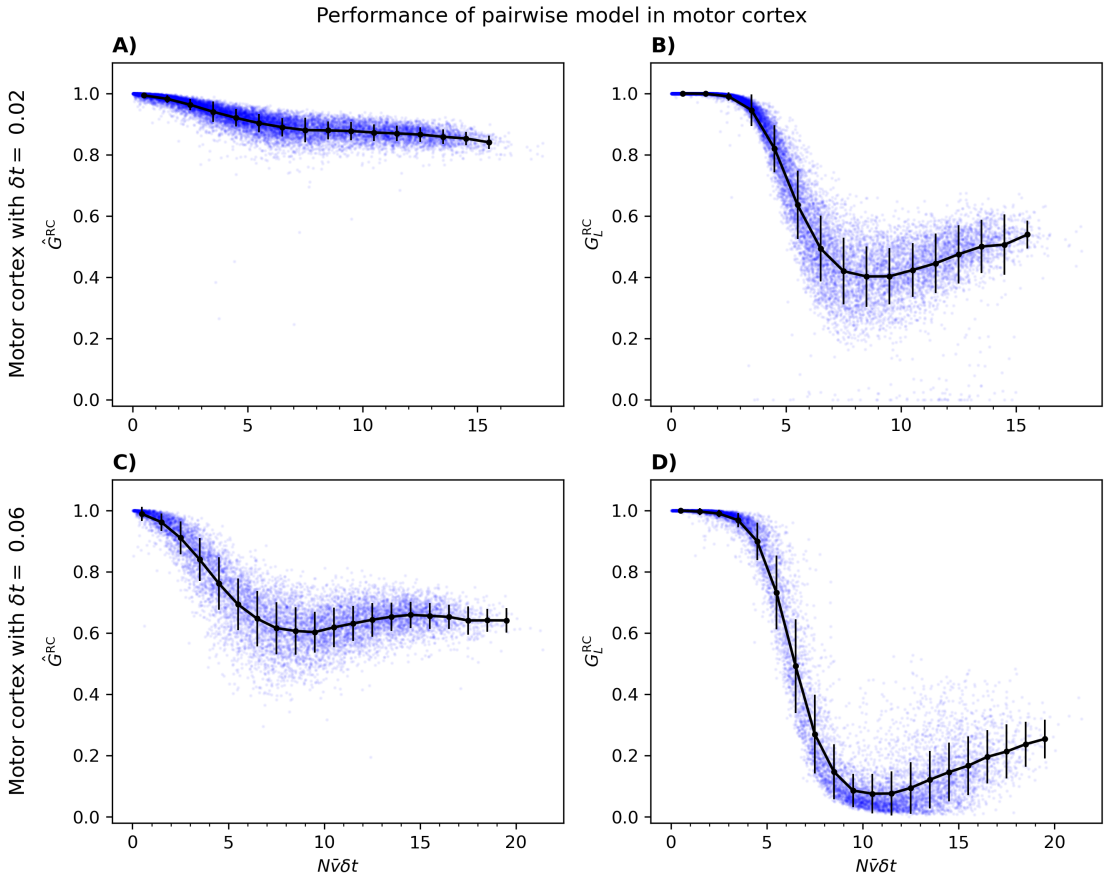


Figure 4.9: Performance of the pairwise model inferred with pseudolikelihood from neural data, for large N in motor cortex. 100 populations of size N were chosen randomly, where N varied from 2 to 100. Pseudolikelihood was used to approximate h and J. The black lines between the black dots represent the means while the black vertical lines represent standard deviations. For subpopulations consisting of 15 or fewer neurons, G was calculated by summer over all states. For subpopulations with more than 15 neurons, $\hat{G}^{\mathrm{RC}}$ was calculated using $\hat{Z}^{\mathrm{RC}}$ from Eq. (22). G_L^{RC} was calculated from Eq. (24) for all subpopulations. (A-B) In addition to the populations up to $N = 100$, 50 populations per N between $N = 101$ and $N = 200$ were are also included. A binsize of $\delta t = 0.02$ was used. The mean firing rate $\bar{\nu}$ had a mean of $M = 3.51$ and standard deviation of $SD = 0.86$ over all populations. 4 populations were omitted due to their parameters being too large and leading to overflow errors. (C-D) A binsize of $\delta t = 0.06$ was used. The mean firing rate $\bar{\nu}$ had a mean of $M = 2.67$ and standard deviation of $SD = 0.64$ over all populations. This figure shows that increasing the binsize recovers the characteristic scaling of $\hat{G}^{\mathrm{RC}}$ in the motor cortex.

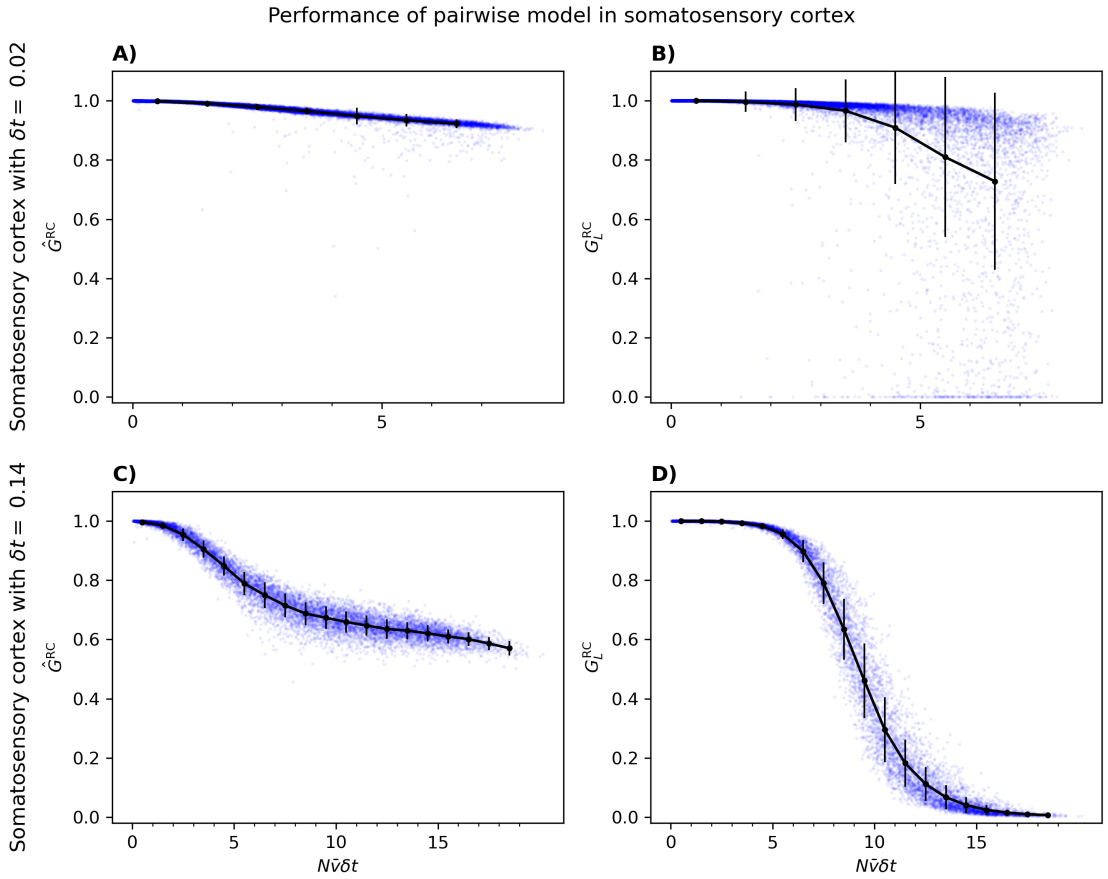


Figure 4.10: Performance of the pairwise model inferred with pseudolikelihood from neural data, for large N in visual, auditory, motor, and somatosensory cortex. For each cortical area, 100 populations of size N were chosen randomly, where N varied from 2 to 100. Pseudolikelihood was used to approximate h and J. The black lines between the black dots represent the means while the black vertical lines represent standard deviations. For subpopulations consisting of 15 or fewer neurons, G^{RC} was calculated by summer over all states. For subpopulations with more than 15 neurons, $\hat{G}^{\mathrm{RC}}$ was calculated using $\hat{Z}$ from Eq. (22). G_L^{RC} was calculated from Eq. (24) for all subpopulations. (A-B) In addition to the populations up to $N = 100$, 50 populations per N between $N = 101$ and $N = 200$ were are also included. The mean firing rate $\bar{v}$ in the motor cortex had a mean of $M = 1.86$ and standard deviation of $SD = 0.40$ over all populations. 87 populations were omitted due to their parameters being too large and leading to overflow errors. (C-D) A binsize of $\delta t = 0.14$ was used. The mean firing rate $\bar{v}$ had a mean of $M = 2.73$ and standard deviation of $SD = 0.55$ over all populations. This figure shows that increasing the binsize recovers the characteristic scaling of $\hat{G}^{\mathrm{RC}}$ in the motor cortex.

First, it is noteworthy that we see a larger $\hat{G}^{\mathrm{RC}}$ for smaller $\bar{v}$ and δt, as expected. Second, the drop off and convergence is much clearer after increasing the binsize. Seeing this characteristic scaling of $\hat{G}^{\mathrm{RC}}$ with $N\bar{v}\delta t$ across different cortical areas suggests that it is a general phenomena.

4.2.8 Effect of finite sampling

To confirm that the above findings would not change significantly if we had more or less data, we here reproduce Figure 4.1 and 4.2 using the finite sampling correction (Strong et al., 1998) and using half of the data.

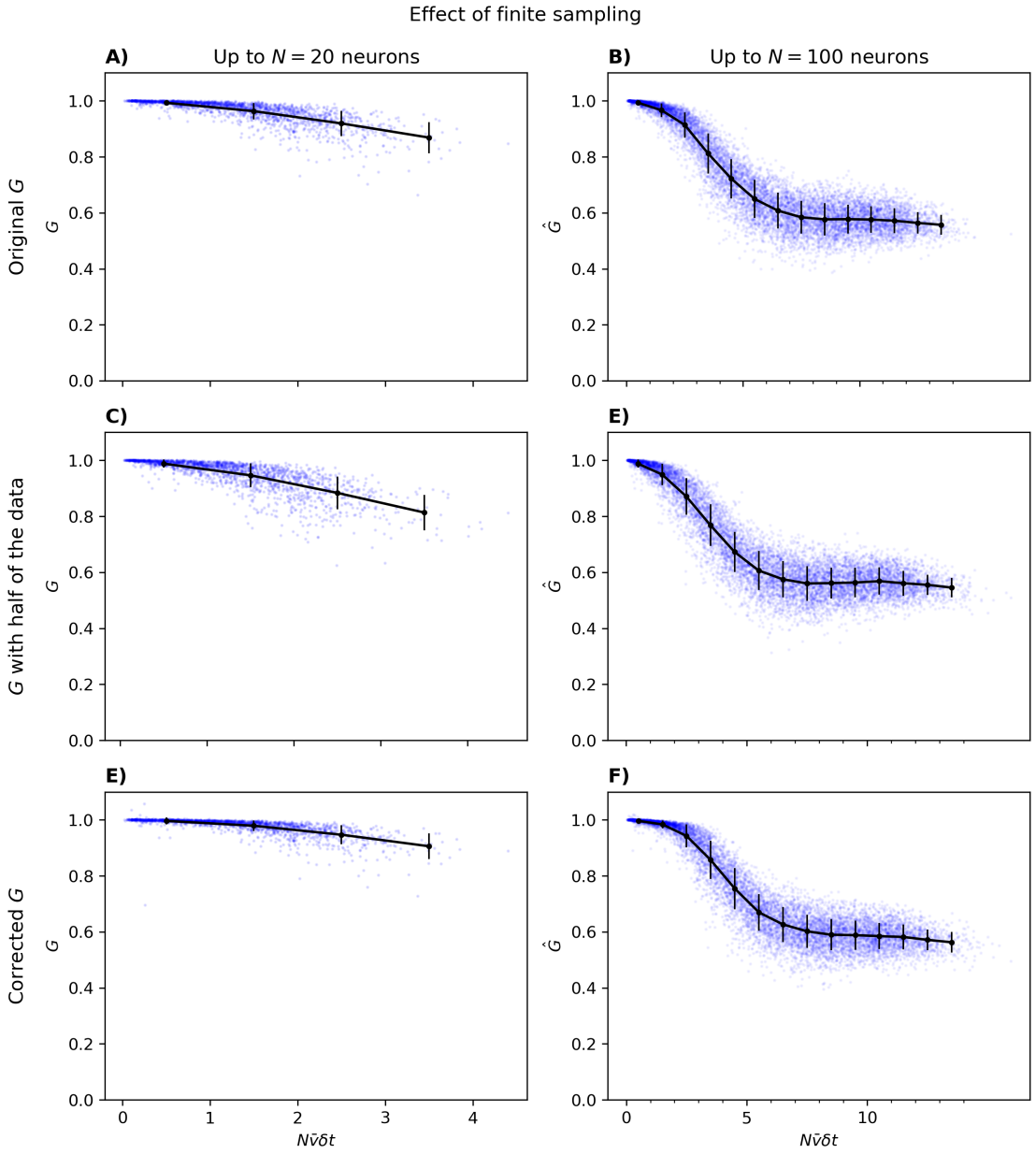


Figure 4.11: (A, C, and E) Same as in Figure 4.1, but with only half the data, and corrected for finite sampling. (B, D, and F) Same as in Figure 4.2, but with only half the data, and corrected for finite sampling. In C and E, a random collection of half of the samples were used for every population. This figure shows that the scaling of G and $\hat{G}$ is not extremely sensitive to the amount of available data.

Both for $N = 20$ and for $N = 100$ we see that having only half of the data increases the standard deviations somewhat and that correcting for finite sampling gives a very slightly larger G^{RC} and $\hat{G}^{\mathrm{RC}}$. However, neither having half of the data nor correcting for finite sampling changes the scaling of G^{RC} or $\hat{G}^{\mathrm{RC}}$ substantially, indicating that our conclusions will not be overly dependent on the amount of data.

4.2.9 Comparison with other performance measures

We now evaluate what third-order correlation can tell us about the performance of the pairwise model. In Figure 4.12A, the root mean squared error of the pairwise correlations from the data and pairwise model are displayed for different subpopulations, while Figure 4.12B shows the same for the connected third-order correlations. We see that neither G_C^{RC} nor $G_{\widetilde{C}}^{\mathrm{RC}}$ show a scaling with $N\bar{v}\delta t$ similar to that of G^{RC} or $\hat{G}^{\mathrm{RC}}$.

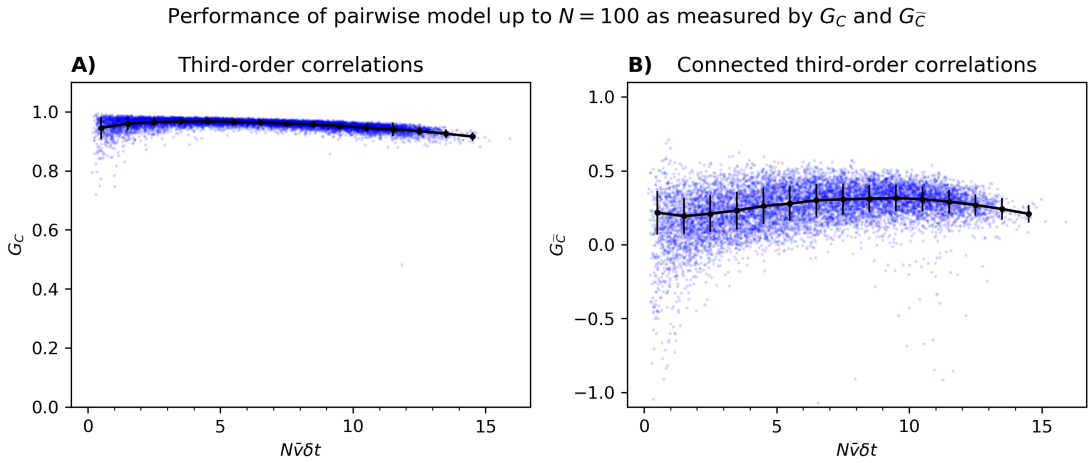


Figure 4.12: Comparison of third-order correlations and connected third-order correlations from neural data and the inferred pairwise model for 5 to 100 neurons. Like in Figure 4.2, 100 populations of size N were chosen randomly from the 495 neurons that were shared over 6 sessions, where N varied from 5 to 100. A constant binsize of $\delta t = 0.02$ was used. Pseudolikelihood was used to approximate h and J. The black lines between the black dots represent the means while the black vertical lines represent standard deviations. The resulting pairwise models were then sampled using the Metropolis-Hastings algorithm (as many samples as in the data) before G_C^{RC} and $G_{\widetilde{C}}^{\mathrm{RC}}$ were calculated from these samples. (A) Performance of the pairwise model as measured by third-order correlations, Eq. (27a), of different populations of neurons. (B) Performance of the pairwise model as measured by connected third-order correlations, Eq. (27b), of different populations of neurons. 17 outliers with $G_{\widetilde{C}}^{\mathrm{RC}} < -1.1$ were omitted.

Finally, In Figure 4.13 we investigate whether comparing the number of simultaneously active neurons in the data and the pairwise model might be a good substitute for G^{RC}. Also here, we see a fairly linear scaling dissimilar from that of G^{RC} and $\hat{G}^{\mathrm{RC}}$.

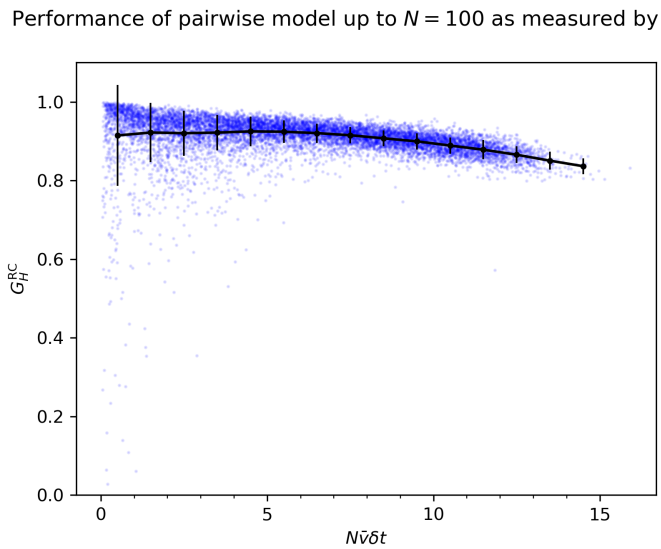


Figure 4.13: Comparison of number of simultaneously active neurons H_m from neural data and the inferred pairwise model for 2 to 100 neurons. Like in Figure 4.2, 100 populations of size N were chosen randomly from the 495 neurons that were shared over 6 sessions, where N varied from 2 to 100. A constant binsize of $\delta t = 0.02$ was used. Pseudolikelihood was used to approximate h and J. The black lines between the black dots represent the means while the black vertical lines represent standard deviations. The resulting pairwise models were then sampled using the Metropolis-Hastings algorithm (as many samples as in the data) before G_H^{RC} were calculated from these samples. The Figure displays the performance of the pairwise model, as measured by G_H^{RC}, Eq. (28), of different populations of neurons. 15 outliers with $G_H^{\mathrm{RC}} < 0$ were omitted, all of which from subpopulations with $N < 5$.

4.3 G – the Pairwise Model Compared to the Independent Model using $h_i^{\mathbf{ind}}$

Now that we have looked at how the pairwise model performs compared to itself without couplings, we will consider how it compares to the maximum entropy independent model. That is, we now investigate the scaling of G and $\hat{G}$, rather than G^{RC} and $\hat{G}^{\mathrm{RC}}$, with $N\bar{v}\delta t$. To do that, we follow the same format as in Section 4.2 and plot the performance against $N\bar{v}\delta t$ under a variety of circumstances. The only difference is that the independent distribution is constructed from h^{ind}, defined in Eq. (5), rather than h, inferred for the pairwise model via pseudolikelihood. To avoid repetition, in this section we only briefly comment on the significance of each figure. Details on the setup can be found either from the corresponding figure in Section 4.2 or in the figure caption.

4.3.1 Changing N — small N

In Figure 4.14, G follows a similar trend as G^{RC} in Figure 4.1 for $N \leq 20$. In this case, however, G starts dropping significantly faster. Still, the large G for small $N\bar{v}\delta t$ fits our expectation (Roudi, Nirenberg et al., 2009). For ease of comparison, both with previous results for small N and with Figure 4.16 and 4.17, $N \leq 10$ is coloured blue while $N \geq 11$ is coloured green.

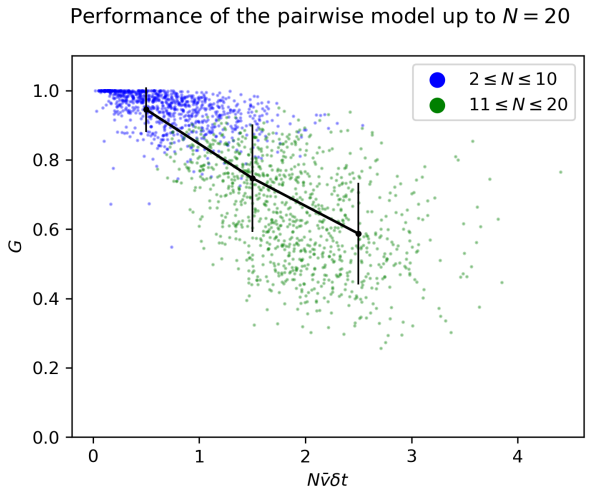


Figure 4.14: Performance of the pairwise model inferred with pseudolikelihood from neural data, for small N. 100 populations of size N were chosen randomly from the 495 neurons that were shared over 6 sessions, where N varied from 2 to 20. The blue dots represent populations of 10 or less neurons, while the green dots represent populations of 11 or more neurons. A constant binsize of $\delta t = 0.02$ was used. The mean firing rate $\bar{\nu}$ had a mean of $M = 6.04$ and standard deviation of $SD = 2.24$ over all populations. Pseudolikelihood was used to approximate h and J. G was calculated by summing over all states. The black lines between the black dots represent the means while the black vertical lines represent standard deviations. This figure shows that G falls as N increases.

4.3.2 Changing N — large N

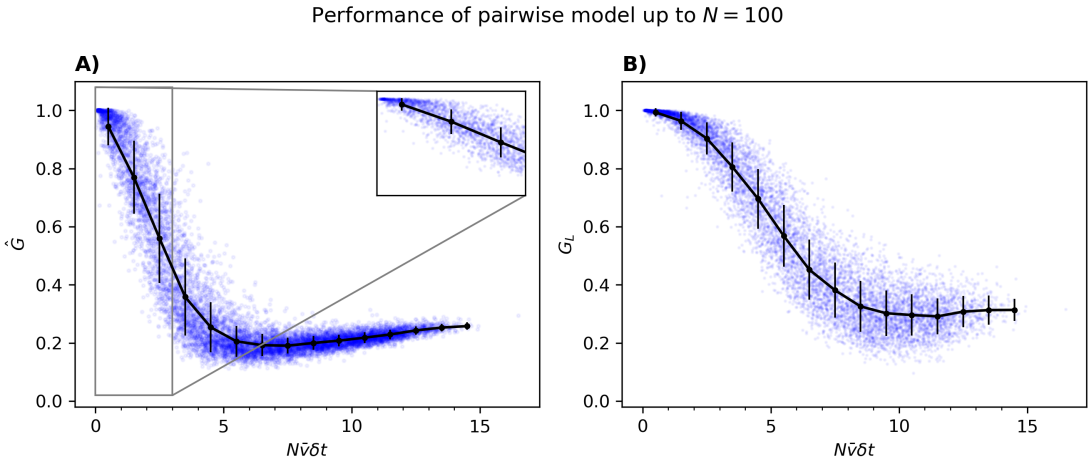


Figure 4.15: Performance of the pairwise model inferred with pseudolikelihood from neural data, for large N. 100 populations of size N were chosen randomly from the 495 neurons that were shared over 6 sessions, where N varied from 2 to 100. A constant binsize of $\delta t = 0.02$ was used. The mean firing rate $\bar{\nu}$ had a mean of $M = 6.07$ and standard deviation of $SD = 1.24$ over all populations. Pseudolikelihood was used to approximate h and J. The black lines between the black dots represent the means while the black vertical lines represent standard deviations. (A) For subpopulations consisting of 15 or fewer neurons, G was calculated by summer over all states. For subpopulations with more than 15 neurons, $\hat{G}$ was calculated using $\hat{Z}$ from Eq. (22). (B) G_L was calculated from Eq. (24) for all subpopulations. This figure shows that both $\hat{G}$ and G_L falls as the number of neurons increases, before levelling off, although $\hat{G}$ level off before G_L.

Like for small N, $\hat{G}$ and G_L follows a similar trend as $\hat{G}^{\mathrm{RC}}$ and G_L^{RC}, except that they start decreasing sharply for a smaller $N\bar{v}\delta t$. So, while the pairwise model might fit the data well when $N\bar{v}\delta t$ is small, its performance quickly falls as $N\bar{v}\delta t$ increases. Further, we see a qualitatively similar scaling of G_L and $\hat{G}$, suggesting that the performance of the pairwise model really does drop as $N\bar{v}\delta t$ increases. Unlike $\hat{G}^{\mathrm{RC}}$, $\hat{G}$ seems to increase slightly as $N\bar{v}\delta t$ becomes large, which is curious.

4.3.3 Changing $\bar{v}$ – semi-random 20-neuron populations

When changing $\bar{v}$, however, G and G^{RC} reacts differently. G^{RC} clearly decreases as $\bar{v}$ increases, while G does not increase or decrease consistently with $\bar{v}$. This is somewhat contrary to our expectations about G going towards one as $N\bar{v}\delta t$ becomes small. Therefore, we calculate G for an additional set of populations, but this time with $N = 10$ (blue). As expected, this results in G increasing, but it still does not seem to converge to one as $N\bar{v}\delta t$ becomes small.

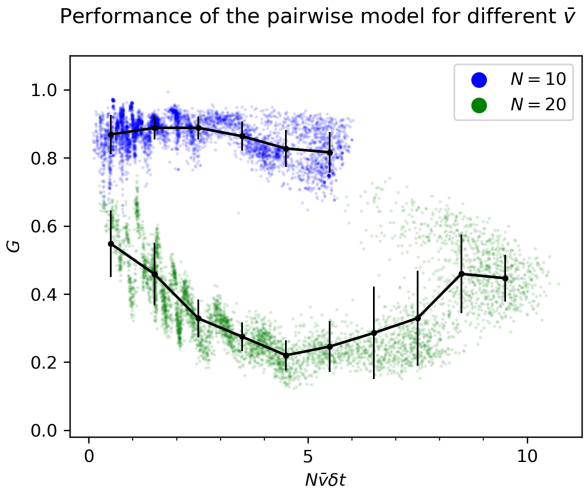


Figure 4.16: Performance of the pairwise model inferred with pseudolikelihood from neural data, for different $\bar{v}$. G was calculated for 5000 subpopulations of 20 neurons (green) and 5000 subpopulations of 10 neurons (blue), by summing over all states. The subpopulations were chosen so that the neurons have similar firing rates (see Section 3.2). A constant binsize of $\delta t = 0.02$ was used. The mean firing rate $\bar{v}$ had a mean of $M = 11.73$ and standard deviation of $SD = 8.29$ over all populations. Pseudolikelihood was used to approximate the parameters h and J. The black lines between the black dots represent the means while the black vertical lines represent standard deviations. This figure shows that there is no clear monotonic relationship between G and $\bar{v}$.

4.3.4 Changing δt – random 20-neuron populations

When changing $\bar{v}$, we observe that G falls as $N\bar{v}\delta t$ increases, like for G^{RC} in Figure 4.3, though not as much. Moreover, the binsize might need to become smaller then 5 ms for G to converge to one, both when $N = 20$ and $N = 10$.

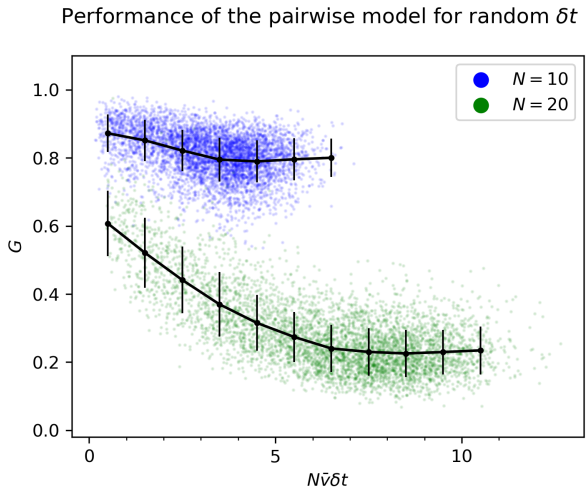


Figure 4.17: Performance of the pairwise model inferred with pseudolikelihood from neural data, for different δt. G was calculated for 5000 subpopulations of 20 neurons (green) and 5000 subpopulations of 10 neurons (blue), by summing over all states. The binsize δt was chosen uniformly between 0.005 and 0.2 seconds. The mean firing rate $\bar{\nu}$ had a mean of $M = 3.79$ and standard deviation of $SD = 1.53$ over all populations. Pseudolikelihood was used to approximate the parameters h and J. The black lines between the black dots represent the means while the black vertical lines represent standard deviations. This figure shows that G falls as δt increases, though not as much as when increasing N.

4.3.5 Performance using nMF, TAP, IP, or SM parameters

Here, we again see that increasing N leads to a considerable drop in $\hat{G}$, even when using other parameter approximation methods. The perhaps biggest difference between $\hat{G}$ in Figure 4.18 and $\hat{G}^{\mathrm{RC}}$ in Figure 4.7 is that there are substantially more outliers with $\hat{G} < 0$ in the former case. This suggests that good parameter approximations are even more important for $\hat{G}$ than for $\hat{G}^{\mathrm{RC}}$.

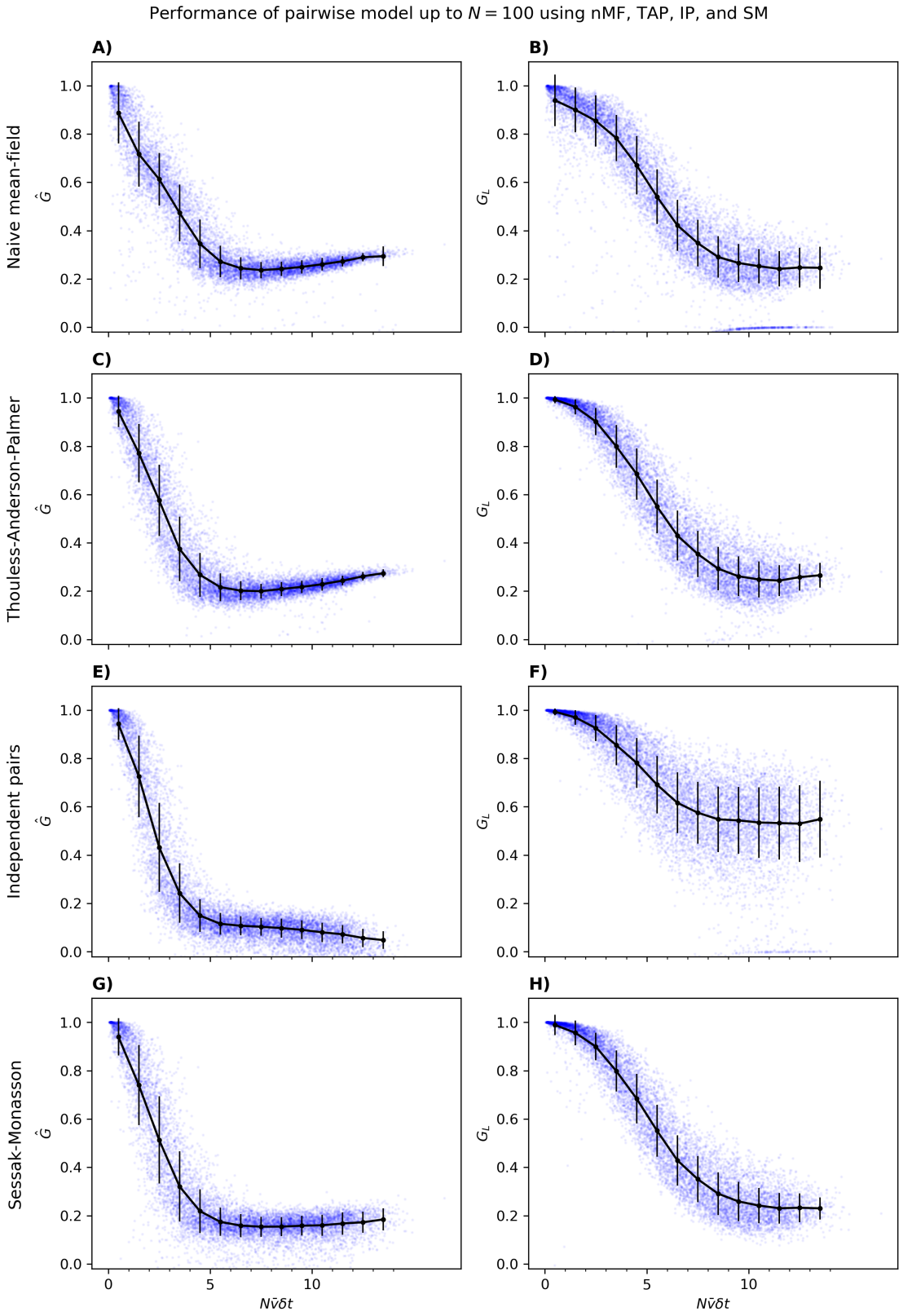


Figure 4.18: Performance of the pairwise model inferred with nMF, TAP, IP, and SM from neural data, for large N. This is identical to Figure 4.15 except that h and J have been approximated with nMF (A-B), TAP (C-D), IP (E-F), or SM (G-H) rather than pseudolikelihood. Using nMF, TAP, IP, and SM resulted in 672, 60, 655, and 288 outliers with $\hat{G} < 0$, respectively, which were omitted. Additionally, for the SM approximation (G-H), 385 (out of 9800) $\hat{G}$s were completely omitted due to overflow errors. This figure shows that inaccurate parameters does have an effect on $\hat{G}$, but the characteristic scaling persists.

4.3.6 Performance for different brain areas

We see from Figure 4.19 that $\hat{G}$ again falls sharply before levelling off as $N\bar{v}\delta t$ increases, even for other recordings from the visual and auditory cortices. This is also the case for recordings from the motor and somatosensory cortices, where we, like in Figure 4.9 and 4.10, use a binsize of $\delta t = 0.06$ in the motor cortex and $\delta t = 0.14$ in the somatosensory cortex to compensate for low firing rates. Even though we see a familiar relationship between $\hat{G}$ and $N\bar{v}\delta t$ in different cortical areas, the scaling of G_L varies more.

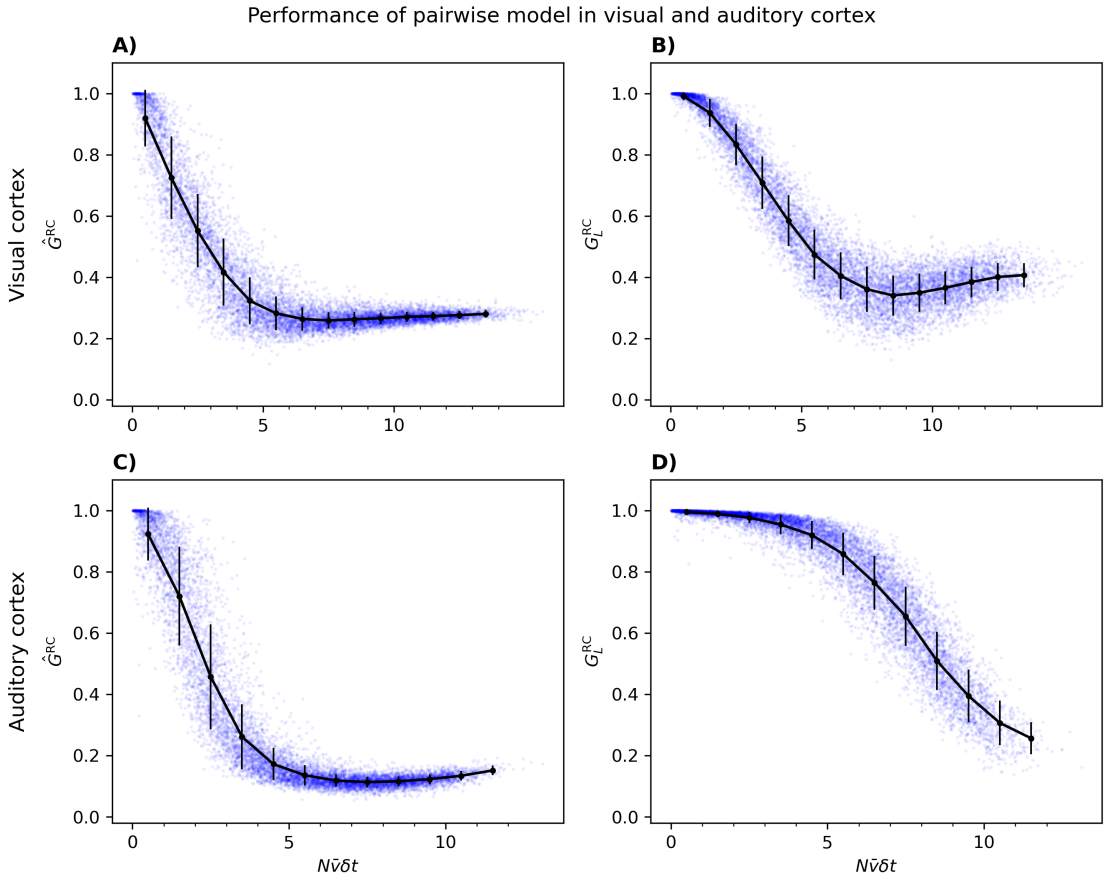


Figure 4.19: Performance of the pairwise model inferred with pseudolikelihood from neural data, for large N in visual and auditory cortex. For both cortical areas, 100 populations of size N were chosen randomly, where N varied from 2 to 100. A constant binsize of $\delta t = 0.02$ was used. Pseudolikelihood was used to approximate h and J. The black lines between the black dots represent the means while the black vertical lines represent standard deviations. For subpopulations consisting of 15 or fewer neurons, G was calculated by summer over all states. For subpopulations with more than 15 neurons, $\hat{G}$ was calculated using $\hat{Z}$ from Eq. (22). G_L was calculated from Eq. (24) for all subpopulations. (A-B) The mean firing rate $\bar{v}$ in the visual cortex had a mean of $M = 6.27$ and standard deviation of $SD = 1.24$ over all populations. (C-D) The mean firing rate $\bar{v}$ in the auditory cortex had a mean of $M = 4.99$ and standard deviation of $SD = 0.93$ over all populations. This figure again shows that different recordings from the visual and auditory cortices exhibit a similar scaling of $\hat{G}$ as that seen in Figure 4.15.

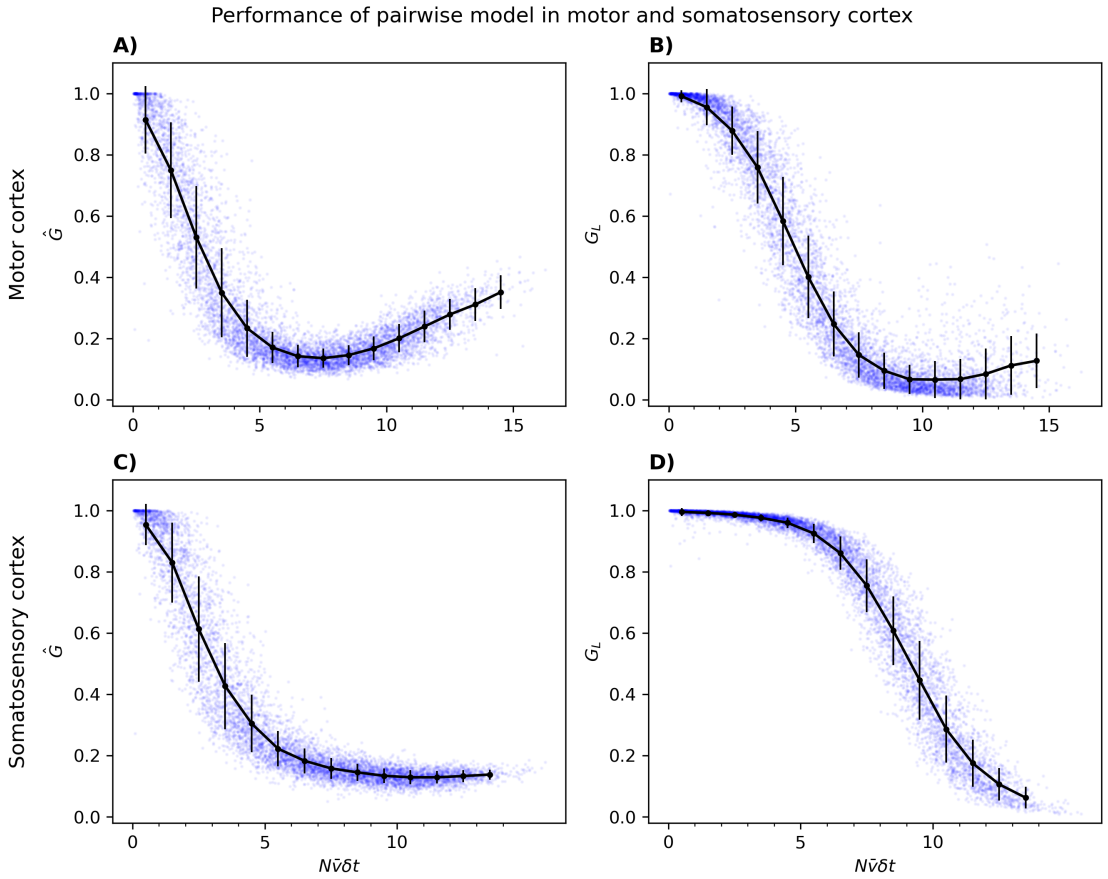


Figure 4.20: Performance of the pairwise model inferred with pseudolikelihood from neural data, for large N in motor and somatosensory cortices. For both cortical areas, 100 populations of size N were chosen randomly, where N varied from 2 to 100. Pseudolikelihood was used to approximate h and J. The black lines between the black dots represent the means while the black vertical lines represent standard deviations. For subpopulations consisting of 15 or fewer neurons, G was calculated by summer over all states. For subpopulations with more than 15 neurons, $\hat{G}$ was calculated using $\hat{Z}$ from Eq. (22). G_L was calculated from Eq. (24) for all subpopulations. (A-B) A constant binsize of $\delta t = 0.06$ was used. The mean firing rate $\bar{v}$ in the visual cortex had a mean of $M = 6.27$ and standard deviation of $SD = 1.24$ over all populations. (C-D) A constant binsize of $\delta t = 0.14$ was used. The mean firing rate $\bar{v}$ in the auditory cortex had a mean of $M = 4.99$ and standard deviation of $SD = 0.93$ over all populations. This figure shows a scaling of $\hat{G}$ in the motor and somatosensory cortices reminiscent of that in Figure 4.15.

4.3.7 Effect of finite sampling

Like for G^{RC} and $\hat{G}^{\mathrm{RC}}$ in Figure 4.2.8, we see that having half of the data increases the standard deviations somewhat and that correcting for finite sampling makes G and $\hat{G}$ very slightly larger. But again, the scaling does not change substantially, suggesting that our conclusions are not skewed too much by a finite sampling bias.

Note also that while it might seem intuitive that G would scale with the average number of samples per state (Ezaki et al., 2017), we see little evidence of that here, even when N range between 2 and 75.

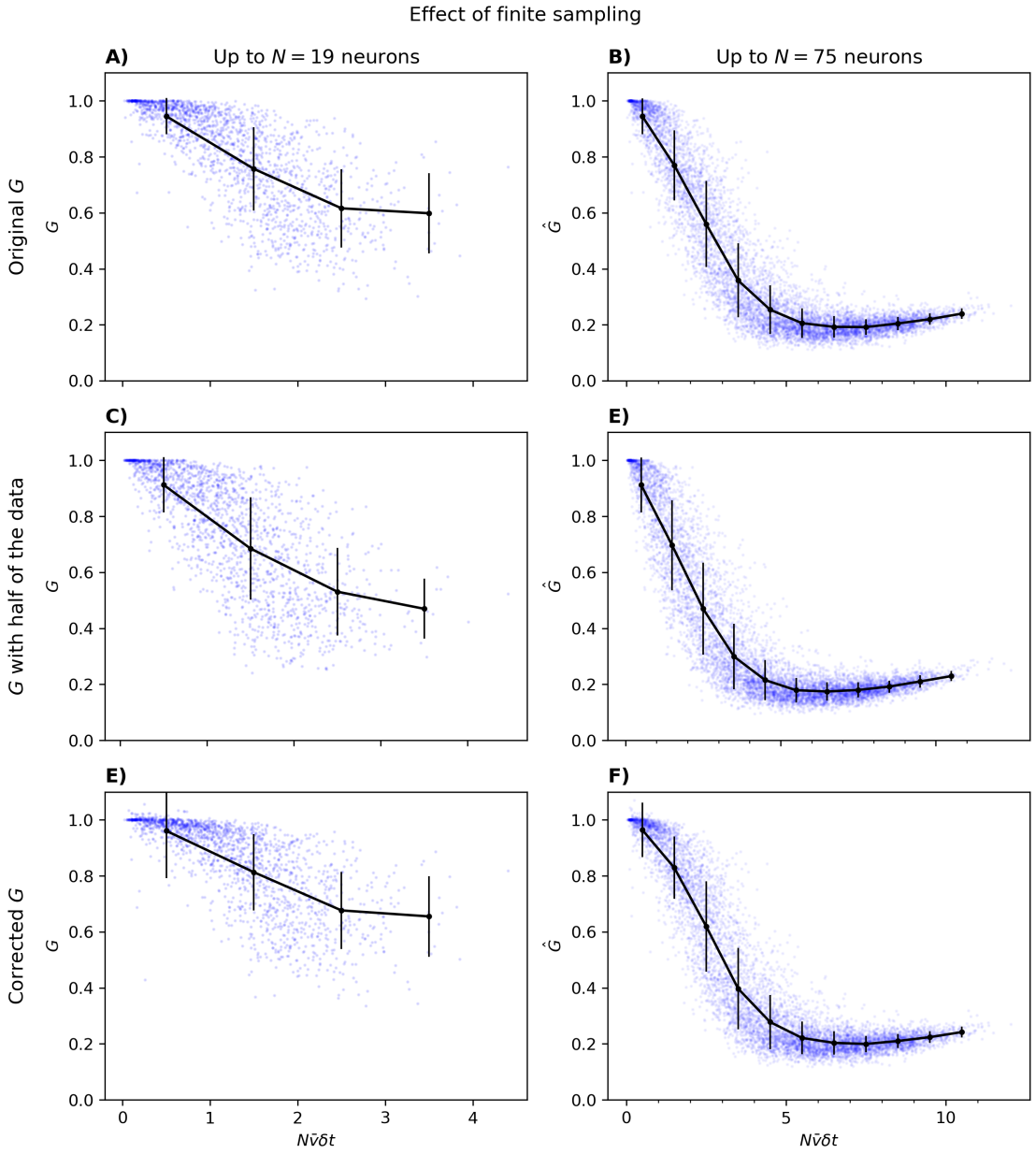


Figure 4.21: (A, C, and E) Same as in Figure 4.14, but with only half the data, and corrected for finite sampling (and up to $N = 19$ rather than $N = 20$). (B, D, and F) Same as in Figure 4.15, but with only half the data, and corrected for finite sampling (and up to $N = 75$ rather than $N = 100$). In C and E, a random collection of half of the samples were used for every population. This figure shows that the scaling of G and $\hat{G}$ is not extremely sensitive to the amount of available data.

4.3.8 Comparison with other performance measures

Even when comparing third-order correlations in the pairwise model and the maximum entropy independent model, which makes G_C and $G_{\widetilde{C}}$ decrease, we still don't see a scaling similar to that of G or $\hat{G}$.

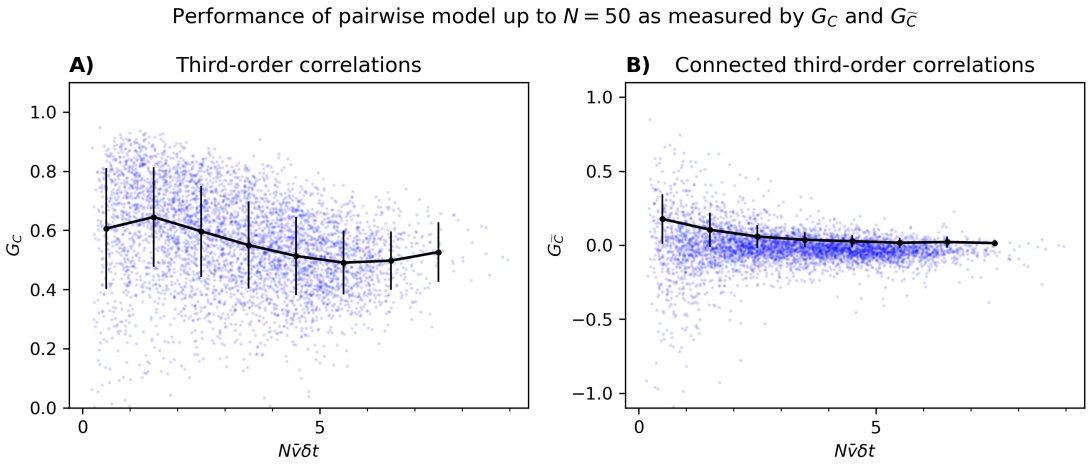


Figure 4.22: Comparison of third-order correlations and connected third-order correlations from neural data and the inferred pairwise model for 5 to 50 neurons. Like in Figure 4.15, 100 populations of size N were chosen randomly from the 495 neurons that were shared over 6 sessions, but here N varied from 5 to 50. A constant binsize of $\delta t = 0.02$ was used. Pseudolikelihood was used to approximate h and J. The black lines between the black dots represent the means while the black vertical lines represent standard deviations. The resulting pairwise models were then sampled using the Metropolis-Hastings algorithm (as many samples as in the data) before G_C and $G_{\widetilde{C}}$ were calculated from these samples. (A) Performance of the pairwise model as measured by third-order correlations, Eq. (27a), of different populations of neurons. 13 outliers with $G_C < 0$ were omitted. (B) Performance of the pairwise model as measured by connected third-order correlations, Eq. (27b), of different populations of neurons. 2 outliers with $G_{\widetilde{C}} < -1.1$ were omitted.

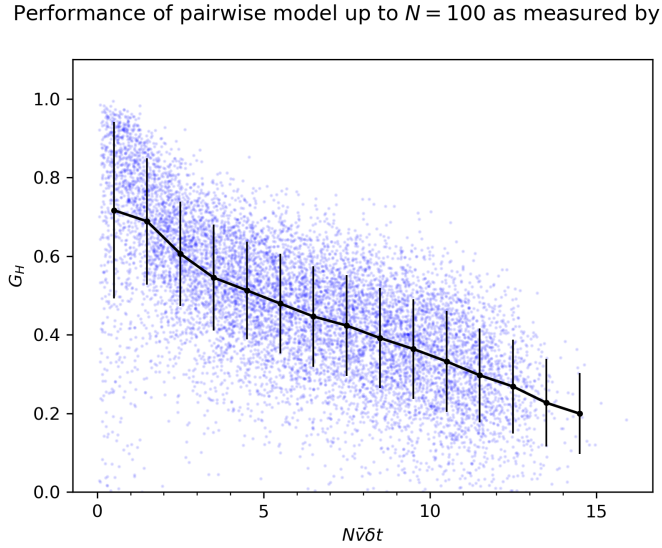


Figure 4.23: Comparison of number of simultaneously active neurons H_m from neural data and the inferred pairwise model for 2 to 100 neurons. Like in Figure 4.2, 100 populations of size N were chosen randomly from the 495 neurons that were shared over 6 sessions, where N varied from 2 to 100. A constant binsize of $\delta t = 0.02$ was used. Pseudolikelihood was used to approximate h and J. The black lines between the black dots represent the means while the black vertical lines represent standard deviations. The resulting pairwise models were then sampled using the Metropolis-Hastings algorithm (as many samples as in the data) before G_H were calculated from these samples. The Figure displays the performance of the pairwise model, as measured by G_H, Eq. (28), of different populations of neurons. 148 outliers with $G_H < 0$ were omitted.

Comparing the number of simultaneously active neurons, however, does reveal a performance more similar to that G or $\hat{G}$. Even though G_H does not display the sharp fall that $\hat{G}$ does, it is substantially closer than G_C and $G_{\widetilde{C}}$, and even G_H^{RC}. Despite this, G_H occasionally (in 148/9800 populations) becomes negative, meaning that the independent model occasionally performs better than the pairwise model.

5 Is the model any good?

Given our systematic investigation of the relationship between G and N, $\bar{v}$, and δt, we are in a much better position than previous studies to make claims about how well the pairwise model accounts for neural data generally.

5.1 The results are consistent with previous findings

Like many previous studies (Chelaru et al., 2021; Ganmor et al., 2011a; Schneidman et al., 2006; Shlens et al., 2006; Tang et al., 2008; Yu et al., 2008), we also find that the pairwise model performs well for small N (Figure 4.14). This finding holds across all tested approximation methods (Figure 4.18) and brain areas (Figure 4.19 and 4.20). This also holds regardless of which biases, h or h^{ind}, are used for the independent model. Further, our findings are consistent with the predicted (Roudi, Nirenberg et al., 2009) linear scaling of G for $N\bar{v}\delta t \ll 1$. Comparing Figure 4.15, 4.16, and 4.17, we see that this relationship is much clearer when changing N than when changing $\bar{v}$ or δt. This might suggest that we did not consider a small enough $\bar{v}$ and/or δt. If we instead consider G^{RC}, the initial linear scaling holds when changing N, $\bar{v}$, or δt independently (Figure 4.1-4.4) and together (Figure 4.5-4.6). Further, some have hinted that the performance of the pairwise model might decrease as N becomes large (Ashourvan et al., 2021; Barreiro et al., 2014; Ezaki et al., 2017; Roudi, Nirenberg et al., 2009; Tkačik et al., 2014), which we here show systematically.

5.2 G decreases with the number of neurons

After the initial linear scaling of $\hat{G}$ and G with $N\bar{v}\delta t$ we see a sharp drop before it levels off. This stereotypical scaling holds when changing N in different cortical areas (Figure 4.19 and 4.20) and using different parameter approximation methods (4.18), but it does not unambiguously hold when only changing $\bar{v}$ or δt (4.16, and 4.17). However, for G^{RC} and $\hat{G}^{\mathrm{RC}}$ the stereotypical scaling persists when changing N, $\bar{v}$, and δt independently (Figure 4.1-4.4) and together (Figure 4.5-4.6), in addition to for all tested approximation methods (Figure 4.7) and brain areas (Figure 4.8-4.10). Although, changing N, $\bar{v}$, and δt independently did result in G and $\hat{G}$ converging to slightly different values. Finally, the alternative performance measure G_L and G_L^{RC} scales similar to G and G^{RC} with $N\bar{v}\delta t$, further increasing our confidence that this scaling is a general phenomena of pairwise models fitted to (our) neural data.

These findings directly addresses the question posed by Roudi, Nirenberg et al. (2009) (their Figure 2), in that we now have a strong case for how the performance of the pairwise model scales with N (and $\bar{v}$ and δt). Importantly, we clearly see that the pairwise model does not universally capture the probability of neural states. Given that we would expect good performance for small

$N\bar{v}\delta t$ regardless of what the true distribution is (Roudi, Nirenberg et al., 2009), neuronal data might not be structured in a way that makes the pairwise model unusually good. This is of course not to say that the pairwise model is useless. It still seems to account well for neuronal activity in the perturbative regime (small $N\bar{v}\delta t$). Moreover, sometimes we might not care that some higher-order correlations are neglected, for example when using the pairwise model as a decoder (Posani et al., 2017). However, if we are looking for the parameters that account maximally for neuronal activity, pairwise correlations might not be it.

Since we find that G increases as δt decreases, one might be able to achieve arbitrarily good performance simply by choosing a sufficiently small binsize δt. However, if one continues to make the binsize smaller, one will violate the independent timebins assumption made when taking the true distribution to be the frequency distribution of time binned states (i.e., when taking $p_{\text{true}} = p_{\text{data}}$). Thus, if we use an excessively small timebin (smaller than the correlation time of the spike trains), $p_{\text{true}} = p_{\text{data}}$ becomes a worse assumption and we would instead want to infer a temporally correlated distribution (Roudi, Nirenberg et al., 2009), such as a kinetic Ising model.

5.3 Good parameter approximations matter

First, we note that using other parameter approximation methods does change the scaling of $\hat{G}$ and $\hat{G}^{\text{RC}}$ somewhat. For $\hat{G}$, using suboptimal parameters led to both under- and overestimations, while it for $\hat{G}^{\text{RC}}$ primarily resulted in overestimations. Because TAP and SM typically approximates h and J better than nMF and IP (Roudi, Aurell et al., 2009; Roudi, Tyrcha et al., 2009), the deviations are more apparent for nMF and IP than for TAP and SM (Figure 4.7 and 4.18). Still, we see substantially more populations with a $\hat{G}$ or $\hat{G}^{\text{RC}}$ smaller than zero when using nMF, TAP, IP, or SM compared to the more reliable pseudolikelihood. This likely results from violating the assumptions in the closed-form approximations. The biases h in the SM approximation are particularly sensitive to violations of the small-J assumption, which resulted in overflow errors in 385 (out of 9800) $\hat{G}$s. This makes the usefulness of the closed-form approximations dataset-dependent. In the dataset tested here, using the TAP approximation resulted in $\hat{G}$s and $\hat{G}^{\text{RC}}$s closest to the ones from pseudolikelihood. Curiously, G_L and G_L^{RC} seems less sensitive to suboptimal parameters (except for the IP approximation), suggesting nMF, TAP, and/or SM could be useful for evaluating pairwise models with G_L or G_L^{RC}.

5.4 G_C and $G_{\widetilde{C}}$ are not good proxies for G, but G_H might be

Neither the third-order correlations nor the connected third-order correlations show a scaling with $N\bar{v}\delta t$ expected from G, $\hat{G}$, G_L, G^{RC}, $\hat{G}^{\text{RC}}$, or G_L^{RC}. Figure 4.22A and 4.12A shows that the third-order correlations C_{ijk} continue to be largely accounted for by the pairwise model. Thus, the synchronized activity of triplets of neurons occurs close to a rate expected from their firing rates and pairwise correlations. Figure 4.22B and 4.12B shows that the third-order correlations $\widetilde{C}_{ijk}$ that go beyond this expectation are almost equally well accounted for by the pairwise and independent model. The fact that G_C, $G_{\widetilde{C}}$, G_C^{RC}, and $G_{\widetilde{C}}^{\text{RC}}$ do not change much with $N\bar{v}\delta t$, while $\hat{G}$ and $\hat{G}^{\text{RC}}$ does, likely reflects two facts: (1) that $\hat{G}$ and $\hat{G}^{\text{RC}}$ include contributions from the pairwise model failing to predict higher-order correlations and (2) that even small errors in predicting third-order correlations (and for that matters higher-order ones) may lead to a large decrease in $\hat{G}$ and $\hat{G}^{\text{RC}}$, because the number of kth order correlations initially increase exponentially with k.

In contrast, measuring performance based on the number of simultaneously active neurons (Figure 4.23 and 4.13) promote a somewhat different conclusion depending on which independent model is used. We see that G_H does decrease substantially with $N\bar{v}\delta t$, as opposed to G_H^{RC}. This suggests that the higher-order correlations implied by the number of simultaneously active neurons are partially responsible for the drop $\hat{G}$ but not $\hat{G}^{\mathrm{RC}}$. From this, one might suspect that G could be approximated by comparing a collection of correlations of different orders.

Despite the somewhat promising results for G_H, the third-order correlations and the number of simultaneously active neurons does not fully capture the performance of the pairwise model. This suggests that a large portion of the possible higher-order correlations has a small but non-negligible contribution. That is, the amount of entropy accounted for by progressively more higher-order correlations may not level off.

5.5 The couplings are responsible for the good performance for small $N\bar{v}\delta t$, but less so for large $N\bar{v}\delta t$

G^{RC} and $\hat{G}^{\mathrm{RC}}$ tells us how much a pairwise model relies on its couplings J, as opposed to its biases h, for matching the data distribution. As G^{RC} and $\hat{G}^{\mathrm{RC}}$ never go to zero, we see that removing the couplings always makes the pairwise model worse. However, the removal of couplings is a lot more destructive in populations with small $N\bar{v}\delta t$. Conversely, removing couplings is less influential for large $N\bar{v}\delta t$, which makes sense considering that G and $\hat{G}$ also fall as $N\bar{v}\delta t$ increases. One would be removing couplings that did not give a good performance in the first place. Yet, this explanation is unlikely to fully explain the scaling of $\hat{G}^{\mathrm{RC}}$ with $N\bar{v}\delta t$ because it holds regardless of whether N, $\bar{v}$, or δt is changed (Figure 4.1-4.6), while the scaling of $\hat{G}$ does not. A perhaps supplementary explanation comes from noticing that decreasing N, $\bar{v}$, and δt all lower the average number of spikes in each sample. Thus, one might suspect that how much pairwise models rely on their couplings, as opposed to their biases, is determined by the average number of simultaneously active neurons. This would mean that it is difficult for couplings to capture the probability of many neurons being active at a time. This fits nicely with the finding that the couplings are almost entirely responsible for the ability of the pairwise model to account for third-order correlations (Figure 4.12) and the number of simultaneously active neurons (Figure 4.13).

5.6 $\hat{Z}$ could facilitate the evaluation of other maximum entropy models

Interest in maximum entropy models extends well beyond the pairwise model (Yeh et al., 2010). For example, some have removed constraints from the pairwise model (Ganmor et al., 2011a; Shlens et al., 2009; Shlens et al., 2006), while others have added constraints (Ganmor et al., 2011b; Shimazaki et al., 2015; Tkačik et al., 2014; Tkačik et al., 2013). Yet others have constructed maximum entropy models by adding temporal correlations (Tang et al., 2008; Vasquez et al., 2012) or stimulus-dependence (Granot-Atedgi et al., 2013). In all cases, this have been with the explicit aim of finding models with sets of constraints that describe neural data well. However, in evaluating the performance of these models for large N one again encounters the exponentially increasing number of states. A good approximation of Z could alleviate this in the same way as for the pairwise model: by enabling an approximation of G via a sum over all sampled states rather than all possible states. Our approximation of Z could therefore aid the evaluation of other maximum entropy models. Although, it should of course be tested for that particular model first.

6 What does this tell us about the brain?

So, we have a way of evaluating the performance of the pairwise maximum entropy model. How does this help us understand the brain? This discussion is primarily concerned with G and $\hat{G}$ because they have a more intuitive meaning than G^{RC} and $\hat{G}^{\mathrm{RC}}$.

6.1 G measures the importance of higher-order correlations in a set of spike trains

G is large when d_{ind} is large relative to d_{pair} and small when d_{pair} is large relative to d_{ind}. This means that G is large when the pairwise model fits the data distribution substantially better than the independent distribution. However, G can be small both when the pairwise and independent model are equally bad at accounting for higher-order correlations and when they are equally good at capturing a truly independent distribution. This latter case complicates the interpretation of G but, fortunately, we rarely record completely independent neurons. With this caveat, one could say that G measures how much of the total (i.e., second to Nth-order) correlations in the data that is accounted for by pairwise correlations. G^{RC}, however, is more difficult to interpret. This is because the means $\langle s_i \rangle_{\mathrm{ind}}$ and $\langle s_i \rangle_{\mathrm{data}}$ don't match when using h, which makes d_{ind}, and thus G^{RC}, larger than when using h^{ind}. This is what we see in, for example, Figure 4.1 and 4.14. The means $\langle s_i \rangle_{\mathrm{ind}}$ and $\langle s_i \rangle_{\mathrm{data}}$ not matching means that some of the magnitude of G^{RC} comes from the means $\langle s_i \rangle_{\mathrm{pair}}$ and $\langle s_i \rangle_{\mathrm{data}}$ now matching. This makes it difficult to interpret G^{RC}, as it does not simply reflect the ability of pairwise correlations to account for all correlations in the data.

In any case, when interpreting G or G^{RC}, it is important to remember that our data is a matrix consisting of N binned spike trains. The correlation structure in this matrix depends on N, $\bar{v}$, and δt. So does, as we have seen, G. If we want to say something general about the statistics of some neuronal population, we probably want that statement to be independent of the values of N, $\bar{v}$, and δt. This means that G will be most informative when considered in the context of $N\bar{v}\delta t$.

6.2 The scaling of G with $N\bar{v}\delta t$ measures the importance of higher-order correlations in a local circuit

Interpreting G and $N\bar{v}\delta t$ together is a first step in moving from statements about spike trains to statements about neuronal activity generally. When $N\bar{v}\delta t$ is small, we expect G to scale linearly with $N\bar{v}\delta t$ regardless of what the true distribution is (Roudi, Nirenberg et al., 2009). However, the scaling of G with $N\bar{v}\delta t$ when moving beyond the perturbative regime, when the scaling becomes non-linear, can inform us about the importance of higher-order correlations in the recorded neural activity. While N seems to affect the scaling of G the most, followed by δt and $\bar{v}$, it is still clear that higher-order correlations become more important as $N\bar{v}\delta t$ increases. This might, speculatively, be explained by (1) there being more opportunities for strong higher-order correlations in larger populations, (2) a larger firing rate making it more likely for groups of neurons to spike together, and (3) neurons being more synchronized over larger timescales. However, the scaling of G with $N\bar{v}\delta t$ is different depending on how $N\bar{v}\delta t$ was changed (e.g., Figure 4.14-4.17). This makes it difficult to draw firm conclusions about a neuronal population from the scaling of G with $N\bar{v}\delta t$. This might suggest that considering G as a function of N, instead of $N\bar{v}\delta t$, could be more interpretable. Still,

the average G or $\hat{G}$ for different values of $N\bar{\nu}\delta t$ reflects the degree of importance of higher-order correlations in the data.

In our analyses we had recordings of hundreds or thousands of neurons but only considered a subpopulation of those at a time. However, we see fairly good agreement between the Gs of the different subpopulations, hinting that higher-order correlations are about equally important in the entire population of recorded neurons. Notably, from the perspective of one subpopulation of neurons, there is no difference between a recorded but not included neuron and a not recorded neuron. This suggests that the conclusions based on many subpopulations of recorded neurons can be extended somewhat to non-recorded neurons in a, vaguely defined, local circuit.

6.3 Higher-order correlations seems to be more important in visual and auditory cortices than in somatosensory and motor cortices

Comparing how G or $\hat{G}$ scales with $N\bar{\nu}\delta t$ in different populations might teach us more about the significance of the distribution of Gs and $\hat{G}$s than looking at a single population. That is, we can say something about the relative importance of higher-order correlations. While we here compare different cortical areas, such comparative analyses can of course be extended far beyond that, for example to comparing the effect of different experimental conditions. We find that, in all cortical areas, $\hat{G}$ start decreasing almost immediately, when $N\bar{\nu}\delta t < 1$. However, there are slight differences in the value $\hat{G}$ converges to, suggesting that higher-order correlations are more important in some areas than others. This might be interpreted as varying degrees of collaborative (i.e. population) coding, which might reflect that different areas carry different amounts of information about some stimuli (Cayco-Gajic et al., 2015). Additionally, in visual and motor cortices we see that $\hat{G}$ start increasing slightly when $N\bar{\nu}\delta t$ becomes larger. It is unclear what is causing this, but from Figure 4.16 one might suspect that a large firing rate may be involved. It is also noteworthy that we see such a similar scaling in the four cortical areas despite increasing the binsize for motor and somatosensory cortices. From Figure 4.17, one might expect that $\hat{G}$ should converge to a smaller value for a larger binsize, suggesting that higher-order correlations are more important in visual and auditory than in motor and somatosensory cortices.

6.4 Potential mechanisms behind higher-order correlations

The higher-order correlations responsible for the drop in G and $\hat{G}$ must come from somewhere. In early sensory areas, they could be driven by some external stimulus that synchronize neuronal activity. Barreiro et al. (2014) suggest that higher-order correlations could be induced in small $(3-16)$ networks of retinal ganglion cells (RGCs) by a bimodal stimulus distribution where the RGCs are active above some stimulus threshold. However, we don't see any signs of this in our data as G is large for small N (Figure 4.1). Our findings are unlikely to be fully explained by stimulus-induced correlations both because of this, and because our data is concatenated over different sessions recorded during different experimental conditions (see Section 4.1). Thus, the higher-order correlations should be generated by some biological mechanism (Shlens et al., 2009). For example, higher-order correlations could be driven by fast recurrent excitation (Barreiro et al., 2014). This might be mimicked by considering a timescale (i.e., binsize) so large that activity have time to propagate through a network, which is consistent with G decreasing as δt increases. Other potential sources of higher-order correlations could be unobserved neurons (Meshulam et al.,

2021), neuromodulation, oscillations, or tripartite synapses. Of course, it is difficult to evaluate which mechanism(s) are most influential from the current dataset. But, this work could help guide experiments to areas in which such mechanisms are likely to contribute substantially to network dynamics.

6.5 Limitations and future work

Despite these insights, there are many things the pairwise model can't tell us. Most notably, time is completely ignored as each sample/state is assumed to be independent and identically distributed when estimating p_{true} with relative frequencies. This makes it impossible to detect plasticity or account for time-varying stimuli. The uncertainty in the model parameters h and J were not considered here, beyond a general statement about the approximation method that generated them (Figure 2.1-2.5). Estimating the uncertainty of each parameter could help us determine whether G is significantly affected by the model fit (Zanoci et al., 2019), thus allowing stronger conclusions about the performance of the pairwise model. As alluded to previously, comparing the scaling of G under different experimental conditions is also an interesting direction for future investigations. Both because one might suspect that the pairwise model performs differently under less varied stimuli than considered here, and because it might tell us something about the relative degree of population coding under different conditions.

Another topic worthy of further study is the accuracy of $\hat{Z}$. If $\hat{G}$ is to be applied more broadly, we need a better understanding of when and how $\hat{Z}$ is a poor approximation. It was suggested here that $\hat{Z}$ might be a good approximation when the entropy is not too large, but this was not quantified. We also need a better understanding of how $\hat{Z}$ affects $\hat{G}$ and interacts with other sources of bias. Some initial testing suggests that $\hat{G}$ sometimes overestimates G when applied to neural data, but this should be investigated further. One can decompose the error in $\hat{G}$ into two sources in Eq. 21. The first due to the approximation of the model entropy S_{pair}, which $\hat{Z}$ does affect, and the second due to the approximation of the data entropy S_{data}, which $\hat{Z}$ does not affect. Here, these two sources have been lumped together by calculating the KL divergence d_{pair} rather than the entropies S_{pair} and S_{data}. However, considering them separately may be beneficial when searching for a better understanding of when $\hat{G}$ fails. To disentangle the influence of different approximations one could calculate the entropies, KL divergences, and Gs in different ways. The true data entropy S_{data} can be calculated either by the plugin method where the frequency distribution p_{data} is used directly, resulting in $S_{\text{data}}^{\text{plugin}} = \sum_{\mathbf{s}} p_{\text{data}}(\mathbf{s}) \ln p_{\text{data}}(\mathbf{s})$, or by some entropy approximation algorithm (e.g., Archer et al., 2013; Strong et al., 1998), whose entropy we denote as $S_{\text{data}}^{\text{approx}}$. The entropy of the pairwise model S_{pair} can be calculated either with a sum over all states when N is small, giving $S_{\text{pair}}^{\text{exact}} = \sum_{\mathbf{s}} p_{\text{pair}}(\mathbf{s}) \ln p_{\text{pair}}(\mathbf{s})$, or by sampling the pairwise model and using one of the above entropy approximation algorithms when N is large, denoted by $S_{\text{pair}}^{\text{sampling}}$. The KL divergence can be expressed as $\sum_{\mathbf{s}} p_{\text{data}}(\mathbf{s}) \ln \frac{p_{\text{data}}(\mathbf{s})}{p_{\text{pair}}(\mathbf{s})} = \sum_{\mathbf{s}} p_{\text{data}}(\mathbf{s}) \ln p_{\text{data}}(\mathbf{s}) - \sum_{\mathbf{s}} p_{\text{data}}(\mathbf{s}) \ln p_{\text{pair}}(\mathbf{s})$, where the plugin entropy $S_{\text{data}}^{\text{plugin}}$ can be replaced by an approximation that corrects for finite sampling $S_{\text{data}}^{\text{approx}}$, and the cross entropy $S_{\text{pair}}^{\text{cross}} = \sum_{\mathbf{s}} p_{\text{data}}(\mathbf{s}) \ln p_{\text{pair}}(\mathbf{s})$ can be replaced with $S_{\text{pair}}^{\text{exact}}$ or $S_{\text{pair}}^{\text{sampling}}$. Finally, either of these possible KL divergences d_{pair}, with its corresponding d_{ind}, can be used to calculate G. So, there are many possible comparisons to make. We suspect that a good starting point for investigating the relative effect of underestimating the data entropy due to finite sampling and replacing the cross entropy with a true entropy is to compare four approximations of the KL divergence: (1) $d_{\text{pair}} = S_{\text{data}}^{\text{plugin}} - S_{\text{pair}}^{\text{cross}}$, (2) $d_{\text{pair}} = S_{\text{data}}^{\text{plugin}} - S_{\text{pair}}^{\text{exact}}$, (3) $d_{\text{pair}} = S_{\text{data}}^{\text{approx}} - S_{\text{pair}}^{\text{cross}}$, and

(4) $d_{\text{pair}} = S_{\text{data}}^{\text{approx}} - S_{\text{pair}}^{\text{exact}}$. Note that we used the first of these approximations in our results. Subsequently, the effect of inaccuracies in $\hat{Z}$ could be assessed simply be replacing Z in p_{pair} with $\hat{Z}$. Using $\hat{Z}$ would also allow us to compare the first and third of the four d_{pair} approximations for large N, while using $S_{\text{pair}}^{\text{sampling}}$ in place of $S_{\text{pair}}^{\text{exact}}$ enables the calculation of approximation two and four as well.

Finally, we still have to choose a binsize to construct the states we are approximating a probability distribution over. If we care about spike statistics and relationships between neurons, this is probably fine provided that we recognize how the chosen binsize affects the results. If we care about coding (e.g., Tkačik et al., 2014), however, it may be misleading to consider every state $\mathbf{s}$ to be unique like when calculating KL divergences. That is, G tells us about how well the pairwise model preserves the correlation structure in the data, it does not tell us how the pairwise model preserves some meaning or representation in the data. For example, given some value of G, it is not obvious whether the torus representation found in grid cells (Gardner et al., 2022) is preserved in the model. For this, it might be more interesting to consider states that are similar, in terms of conveying the same information (i.e., location on the torus), together. One might suspect that similar states according to Hamming distance (syntactic similarity) convey similar information (semantic similarity), but this does not seem be the case (Ganmor et al., 2015). Instead, (semantic) similarity between states could be learned from the information they contain about some stimulus. Note that which states that are similar to each other might change under different conditions. These similarities could be used to define clusters of states containing approximately the same information (Ganmor et al., 2015). Looking at clusters of states may allow for an interpretation that is more related to coding and representations, though it should be done with caution (Brette, 2019; de-Wit et al., 2016). Further, this might simplify our analyses and interpretations in that we could consider less than 2^N states, as exemplified by Wolf et al. (2023) primarily caring about the mean activity in two brain areas. Even though $\hat{G}$ may have mitigated some of the trouble of looking at all 2^N states individually, it will likely not scale to thousands or millions of neurons.

7 Conclusion

We have, for the first time, systematically evaluated the performance of the pairwise maximum entropy model for large N. Our results are consistent with previous work investigating the performance of pairwise models, but expand the work in important ways by showing that the good performance observed for small N does not continue indefinitely. Additionally, we see that good approximations of h and J are important for obtaining reliable $\hat{G}$s and that third-order correlations are not responsible for the drop in $\hat{G}$, while the number of simultaneously active neurons might be responsible for some of it. We also speculate that our approximation of Z (Eq. 23) could be useful more broadly, such as when evaluating other maximum entropy models, but this should be investigated in future work. Finally, we discuss what G might tell us about neuronal dynamics, and conclude that higher-order correlations plays an important role in populations with many neurons. That is, pairwise correlations are not sufficient to account for neural activity in the general case.

Bibliography

Ackley, D. H., Hinton, G. E., & Sejnowski, T. J. (1985). A learning algorithm for boltzmann machines. *Cognitive science*, *9*(1), 147–169.

Archer, E. W., Park, I. M., & Pillow, J. W. (2013). Bayesian entropy estimation for binary spike train data using parametric prior knowledge. *Advances in neural information processing systems*, *26*.

Ashourvan, A., Shah, P., Pines, A., Gu, S., Lynn, C. W., Bassett, D. S., Davis, K. A., & Litt, B. (2021). Pairwise maximum entropy model explains the role of white matter structure in shaping emergent co-activation states. *Communications Biology*, *4*(1), 210.

Barreiro, A. K., Gjorgjieva, J., Rieke, F., & Shea-Brown, E. (2014). When do microcircuits produce beyond-pairwise correlations? *Frontiers in computational neuroscience*, *8*, 10.

Besag, J. (1975). Statistical analysis of non-lattice data. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *24*(3), 179–195.

Bethge, M., & Berens, P. (2007). Near-maximum entropy models for binary neural representations of natural images. *Advances in neural information processing systems*, *20*.

Betzel, R. F., & Bassett, D. S. (2017). Generative models for network neuroscience: Prospects and promise. *Journal of The Royal Society Interface*, *14*(136), 20170623.

Brent, R. P. (1971). An algorithm with guaranteed convergence for finding a zero of a function. *The computer journal*, *14*(4), 422–425.

Brette, R. (2019). Is coding a relevant metaphor for the brain? *Behavioral and Brain Sciences*, *42*, e215.

Cayco-Gajic, N. A., Zylberberg, J., & Shea-Brown, E. (2015). Triplet correlations among similarly tuned cells impact population coding. *Frontiers in computational neuroscience*, *9*, 57.

Chelaru, M. I., Eagleman, S., Andrei, A. R., Milton, R., Kharas, N., & Dragoi, V. (2021). High-order interactions explain the collective behavior of cortical populations in executive but not sensory areas. *Neuron*, *109*(24), 3954–3961.

Cofré, R., Herzog, R., Corcoran, D., & Rosas, F. E. (2019). A comparison of the maximum entropy principle across biological spatial scales. *Entropy*, *21*(10), 1009.

Cunningham, J. P., & Yu, B. M. (2014). Dimensionality reduction for large-scale neural recordings. *Nature neuroscience*, *17*(11), 1500–1509.

de-Wit, L., Alexander, D., Ekroll, V., & Wagemans, J. (2016). Is neuroimaging measuring information in the brain? *Psychonomic bulletin & review*, *23*, 1415–1428.

Ezaki, T., Watanabe, T., Ohzeki, M., & Masuda, N. (2017). Energy landscape analysis of neuroimaging data. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *375*(2096), 20160287.

Ganmor, E., Segev, R., & Schneidman, E. (2009). How fast can we learn maximum entropy models of neural populations? *Journal of Physics: Conference Series*, *197*(1), 012020.

Ganmor, E., Segev, R., & Schneidman, E. (2011a). The architecture of functional interaction networks in the retina. *Journal of Neuroscience*, *31*(8), 3044–3054.

Ganmor, E., Segev, R., & Schneidman, E. (2011b). Sparse low-order interaction network underlies a highly correlated and learnable neural population code. *Proceedings of the National Academy of sciences*, *108*(23), 9679–9684.

Ganmor, E., Segev, R., & Schneidman, E. (2015). A thesaurus for a neural population code. *Elife*, *4*, e06134.

Gao, P., & Ganguli, S. (2015). On simplicity and complexity in the brave new world of large-scale neuroscience. *Current opinion in neurobiology*, *32*, 148–155.

Gardner, R. J., Hermansen, E., Pachitariu, M., Burak, Y., Baas, N. A., Dunn, B. A., Moser, M.-B., & Moser, E. I. (2022). Toroidal topology of population activity in grid cells. *Nature*, *602*(7895), 123–128.

Ghojogh, B., Nekoei, H., Ghojogh, A., Karray, F., & Crowley, M. (2020). Sampling algorithms, from survey sampling to monte carlo methods: Tutorial and literature review. *arXiv preprint arXiv:2011.00901*.

Granot-Atedgi, E., Tkačik, G., Segev, R., & Schneidman, E. (2013). Stimulus-dependent maximum entropy models of neural population codes. *PLoS computational biology*, *9*(3), e1002922.

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications.

Hertz, J., Roudi, Y., & Tyrcha, J. (2011). Ising models for inferring network structure from spike data. *arXiv preprint arXiv:1106.1752*.

Jaynes, E. T. (1957). Information theory and statistical mechanics. *Physical review*, *106*(4), 620.

Kadirvelu, B., Hayashi, Y., & Nasuto, S. J. (2017). Inferring structural connectivity using ising couplings in models of neuronal networks. *Scientific reports*, *7*(1), 1–12.

Meshulam, L., Gauthier, J. L., Brody, C. D., Tank, D. W., & Bialek, W. (2017). Collective behavior of place and non-place neurons in the hippocampal network. *Neuron*, *96*(5), 1178–1191.

Meshulam, L., Gauthier, J. L., Brody, C. D., Tank, D. W., & Bialek, W. (2021). Successes and failures of simplified models for a network of real neurons. *arXiv preprint arXiv:2112.14735*.

Mimica, B., Tombaz, T., Battistin, C., Fuglstad, J. G., Dunn, B. A., & Whitlock, J. R. (2022). Behavioral decomposition reveals rich encoding structure employed across neocortex. *bioRxiv*.

Nguyen, H. C., Zecchina, R., & Berg, J. (2017). Inverse statistical problems: From the inverse ising problem to data science. *Advances in Physics*, *66*(3), 197–261.

Ohiorhenuan, I. E., Mechler, F., Purpura, K. P., Schmid, A. M., Hu, Q., & Victor, J. D. (2010). Sparse coding and high-order correlations in fine-scale cortical networks. *Nature*, *466*(7306), 617–621.

Panzeri, S., Senatore, R., Montemurro, M. A., & Petersen, R. S. (2007). Correcting for the sampling bias problem in spike train information measures. *Journal of neurophysiology*, *98*(3), 1064–1072.

Plefka, T. (1982). Convergence condition of the tap equation for the infinite-ranged ising spin glass model. *Journal of Physics A: Mathematical and general*, *15*(6), 1971.

Posani, L., Cocco, S., Ježek, K., & Monasson, R. (2017). Functional connectivity models for decoding of spatial representations from hippocampal ca1 recordings. *Journal of Computational Neuroscience*, *43*(1), 17–33.

Roudi, Y., Aurell, E., & Hertz, J. A. (2009). Statistical physics of pairwise probability models. *Frontiers in computational neuroscience*, 22.

Roudi, Y., Nirenberg, S., & Latham, P. E. (2009). Pairwise maximum entropy models for studying large biological systems: When they can work and when they can't. *PLoS computational biology*, *5*(5), e1000380.

Roudi, Y., Tyrcha, J., & Hertz, J. (2009). Ising model for neural data: Model quality and approximate methods for extracting functional connectivity. *Physical Review E*, *79*(5), 051915.

Savin, C., & Tkačik, G. (2017). Maximum entropy models as a tool for building precise neural controls. *Current opinion in neurobiology*, *46*, 120–126.

Schneidman, E. (2016). Towards the design principles of neural population codes. *Current opinion in neurobiology*, *37*, 133–140.

Schneidman, E., Berry, M. J., Segev, R., & Bialek, W. (2006). Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature*, *440*(7087), 1007–1012.

Sessak, V., & Monasson, R. (2009). Small-correlation expansions for the inverse ising problem. *Journal of Physics A: Mathematical and Theoretical*, *42*(5), 055001.

Sherrington, D., & Kirkpatrick, S. (1975). Solvable model of a spin-glass. *Physical review letters*, *35*(26), 1792.

Shimazaki, H., Sadeghi, K., Ishikawa, T., Ikegaya, Y., & Toyoizumi, T. (2015). Simultaneous silence organizes structured higher-order interactions in neural populations. *Scientific reports*, *5*(1), 1–13.

Shlens, J., Field, G. D., Gauthier, J. L., Greschner, M., Sher, A., Litke, A. M., & Chichilnisky, E. (2009). The structure of large-scale synchronized firing in primate retina. *Journal of Neuroscience*, *29*(15), 5022–5031.

Shlens, J., Field, G. D., Gauthier, J. L., Grivich, M. I., Petrusca, D., Sher, A., Litke, A. M., & Chichilnisky, E. (2006). The structure of multi-neuron firing patterns in primate retina. *Journal of Neuroscience*, *26*(32), 8254–8266.

Sompolinsky, H. (1988). Statistical mechanics of neural networks. *Physics Today*, *41*(21), 70–80.

Stephens, G. J., Osborne, L. C., & Bialek, W. (2011). Searching for simplicity in the analysis of neurons and behavior. *Proceedings of the National Academy of Sciences*, *108*(supplement_3), 15565–15571.

Stevenson, I. H., & Kording, K. P. (2011). How advances in neural recording affect data analysis. *Nature neuroscience*, *14*(2), 139–142.

Strong, S. P., Koberle, R., Van Steveninck, R. R. D. R., & Bialek, W. (1998). Entropy and information in neural spike trains. *Physical review letters*, *80*(1), 197.

Tang, A., Jackson, D., Hobbs, J., Chen, W., Smith, J. L., Patel, H., Prieto, A., Petrusca, D., Grivich, M. I., Sher, A., et al. (2008). A maximum entropy model applied to spatial and temporal correlations from cortical networks in vitro. *Journal of Neuroscience*, *28*(2), 505–518.

Thouless, D. J., Anderson, P. W., & Palmer, R. G. (1977). Solution of'solvable model of a spin glass'. *Philosophical Magazine*, *35*(3), 593–601.

Timme, N. M., & Lapish, C. (2018). A tutorial for information theory in neuroscience. *eneuro*, *5*(3).

Tkacik, G., Schneidman, E., Berry II, M. J., & Bialek, W. (2006). Ising models for networks of real neurons. *arXiv preprint q-bio/0611072*.

Tkačik, G., Marre, O., Amodei, D., Schneidman, E., Bialek, W., & Berry, M. J. (2014). Searching for collective behavior in a large network of sensory neurons. *PLoS computational biology*, *10*(1), e1003408.

Tkačik, G., Marre, O., Mora, T., Amodei, D., Berry II, M. J., & Bialek, W. (2013). The simplest maximum entropy model for collective behavior in a neural network. *Journal of Statistical Mechanics: Theory and Experiment*, *2013*(03), P03011.

Tkačik, G., Schneidman, E., Berry II, M. J., & Bialek, W. (2009). Spin glass models for a network of real neurons. *arXiv preprint arXiv:0912.5409*.

Vasquez, J. C., Marre, O., Palacios, A. G., Berry II, M. J., & Cessac, B. (2012). Gibbs distribution analysis of temporal correlations structure in retina ganglion cells. *Journal of Physiology-Paris*, *106*(3-4), 120–127.

Watanabe, T., Hirose, S., Wada, H., Imai, Y., Machida, T., Shirouzu, I., Konishi, S., Miyashita, Y., & Masuda, N. (2013). A pairwise maximum entropy model accurately describes resting-state human brain networks. *Nature communications*, *4*(1), 1370.

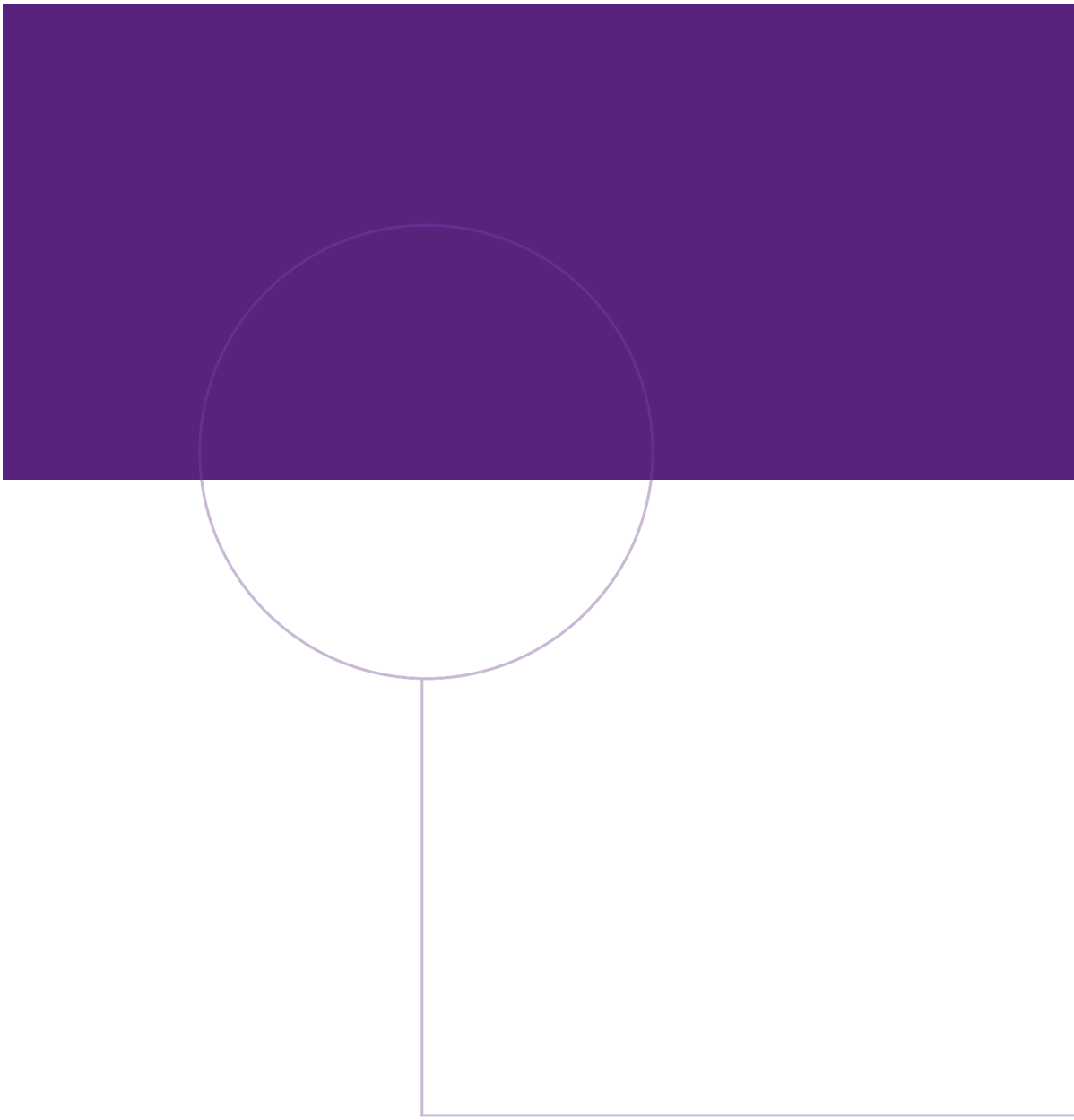
Wolf, S., Le Goc, G., Debrégeas, G., Cocco, S., & Monasson, R. (2023). Emergence of time persistence in a data-driven neural network model. *Elife*, *12*, e79541.

Yeh, F.-C., Tang, A., Hobbs, J. P., Hottowy, P., Dabrowski, W., Sher, A., Litke, A., & Beggs, J. M. (2010). Maximum entropy approaches to living neural networks. *Entropy*, *12*(1), 89–106.

Yu, S., Huang, D., Singer, W., & Nikolić, D. (2008). A small world of neuronal synchrony. *Cerebral cortex*, *18*(12), 2891–2901.

Yuste, R. (2015). From the neuron doctrine to neural networks. *Nature reviews neuroscience*, *16*(8), 487–497.

Zanoci, C., Dehghani, N., & Tegmark, M. (2019). Ensemble inhibition and excitation in the human cortex: An ising-model analysis with uncertainties. *Physical Review E*, *99*(3), 032408.



NTNU

Norwegian University of
Science and Technology