OPEN ACCESS Freely available online

PLoS COMPUTATIONAL BIOLOGY

Pairwise Maximum Entropy Models for Studying Large Biological Systems: When They Can Work and When They Can't

Yasser Roudi[1,2], Sheila Nirenberg[2], Peter E. Latham[1]*

1 Gatsby Computational Neuroscience Unit, University College London, London, United Kingdom, **2** Department of Physiology and Biophysics, Weill Medical College of Cornell University, New York, United States of America

Abstract

One of the most critical problems we face in the study of biological systems is building accurate statistical descriptions of them. This problem has been particularly challenging because biological systems typically contain large numbers of interacting elements, which precludes the use of standard brute force approaches. Recently, though, several groups have reported that there may be an alternate strategy. The reports show that reliable statistical models can be built without knowledge of all the interactions in a system; instead, pairwise interactions can suffice. These findings, however, are based on the analysis of small subsystems. Here, we ask whether the observations will generalize to systems of realistic size, that is, whether pairwise models will provide reliable descriptions of true biological systems. Our results show that, in most cases, they will not. The reason is that there is a crossover in the predictive power of pairwise models: If the size of the subsystem is below the crossover point, then the results have no predictive power for large systems. If the size is above the crossover point, then the results may have predictive power. This work thus provides a general framework for determining the extent to which pairwise models can be used to predict the behavior of large biological systems. Applied to neural data, the size of most systems studied so far is below the crossover point.

Citation: Roudi Y, Nirenberg S, Latham PE (2009) Pairwise Maximum Entropy Models for Studying Large Biological Systems: When They Can Work and When They Can't. PLoS Comput Biol 5(5): e1000380. doi:10.1371/journal.pcbi.1000380

Editor: Olaf Sporns, Indiana University, United States of America

Received November 10, 2008; **Accepted** April 1, 2009; **Published** May 8, 2009

Funding: YR and PEL were supported by the Gatsby Charitable Foundation (http://www.gatsby.org.uk) and by the US National Institute of Mental Health grant R01 MH62447. SN was supported by the US National Eye Institute grant R01 EY12978. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: pel@gatsby.ucl.ac.uk

Introduction

Many fundamental questions in biology are naturally treated in a probabilistic setting. For instance, deciphering the neural code requires knowledge of the probability of observing patterns of activity in response to stimuli [1]; determining which features of a protein are important for correct folding requires knowledge of the probability that a particular sequence of amino acids folds naturally [2,3]; and determining the patterns of foraging of animals and their social and individual behavior requires knowledge of the distribution of food and species over both space and time [4–6].

Constructing these probability distributions is, however, hard. There are several reasons for this: i) biological systems are composed of large numbers of elements, and so can exhibit a huge number of configurations—in fact, an exponentially large number, ii) the elements typically interact with each other, making it impossible to view the system as a collection of independent entities, and iii) because of technological considerations, the descriptions of biological systems have to be built from very little data. For example, with current technology in neuroscience, we can record simultaneously from only about 100 neurons out of approximately 100 billion in the human brain. So, not only are we faced with the problem of estimating probability distributions in high dimensional spaces, we must do this based on a small fraction of the neurons in the network.

Despite these apparent difficulties, recent work has suggested that the situation may be less bleak than it seems, and that an accurate statistical description of systems can be achieved without having to examine all possible configurations [2,3,7–11]. One merely has to measure the probability distribution over pairs of elements and use those to build the full distribution. These "pairwise models" potentially offer a fundamental simplification, as the number of pairs is quadratic in the number of elements, not exponential. However, support for the efficacy of pairwise models has, necessarily, come from relatively small subsystems—small enough that the true probability distribution could be measured experimentally [7–9,11]. While these studies have provided a key first step, a critical question remains: will the results from the analysis of these small subsystems extrapolate to large ones? That is, if a pairwise model predicts the probability distribution for a subset of the elements in a system, will it also predict the probability distribution for the whole system? Here we find that, for a biologically relevant class of systems, this question can be answered quantitatively and, importantly, generically—independent of many of the details of the biological system under consideration. And the answer is, generally, "no." In this paper, we explain, both analytically and with simulations, why this is the case.

Author Summary

Biological systems are exceedingly complicated: They consist of a large number of elements, those elements interact in nonlinear and highly unpredictable ways, and collective interactions typically play a critical role. It would seem surprising, then, that one could build a quantitative description of biological systems based only on knowledge of how pairs of elements interact. Yet, that is what a number of studies have found. Those studies, however, focused on relatively small systems. Here, we ask the question: Do their conclusions extend to large systems? We show that the answer depends on the size of the system relative to a crossover point: Below the crossover point the results on the small system have no predictive power for large systems; above the crossover point the results on the small system may have predictive power. Moreover, the crossover point can be computed analytically. This work thus provides a general framework for determining the extent to which pairwise models can be used to predict the behavior of large biological systems. It also provides a useful heuristic for designing experiments: If one is interested in understanding truly large systems via pairwise interactions, then make sure that the system one studies is above the crossover point.

Results

The extrapolation problem

To gain intuition into the extrapolation problem, let us consider a specific example: neuronal spike trains. Fig. 1A shows a typical spike train for a small population of neurons. Although the raw spike times provide a complete description, they are not a useful representation, as they are too high-dimensional. Therefore, we divide time into bins and re-represent the spike train as 0 s and 1 s: 0 if there is no spike in a bin; 1 otherwise (Fig. 1B) [7–9,11]. For now we assume that the bins are independent (an assumption whose validity we discuss below, and in more detail in the section "Is there anything wrong with using small time bins?"). The problem, then, is to find $p_{true}(\mathbf{r}) \equiv p_{true}(r_1, r_2, \ldots, r_N)$ where r_i is a binary variable indicating no spike ($r_i = 0$) or one or more spikes ($r_i = 1$) on neuron i. Since this, too, is a high dimensional problem (though less so than the original spike time representation), suppose that we instead construct a pairwise approximation to p_{true}, which we denote p_{pair}, for a population of size N. (The pairwise model derives its name from the fact that it has the same mean and pairwise correlations as the true model; see Eq. (15).) Our question, then, is: if p_{pair} is close to p_{true} for small N, what can we say about how close the two distributions are for large N?

To answer this question quantitatively, we need a measure of distance. The measure we use, denoted Δ_N, is defined in Eq. (3) below, but all we need to know about it for now is that if $\Delta_N = 0$ then $p_{pair} = p_{true}$, and if Δ_N is near one then p_{pair} is far from p_{true}. In terms of Δ_N, our main results are as follows. First, for small N, in what we call the perturbative regime, Δ_N is proportional to $N-2$. In other words, as the population size increases, the pairwise model becomes a worse and worse approximation to the true distribution. Second, this behavior is entirely generic: for small N, Δ_N increases linearly, no matter what the true distribution is. This is illustrated schematically in Fig. 2, which shows the generic behavior of Δ_N. The solid red part of the curve is the perturbative regime, where Δ_N is a linearly increasing function of N; the dashed curves show possible behavior beyond the perturbative regime.

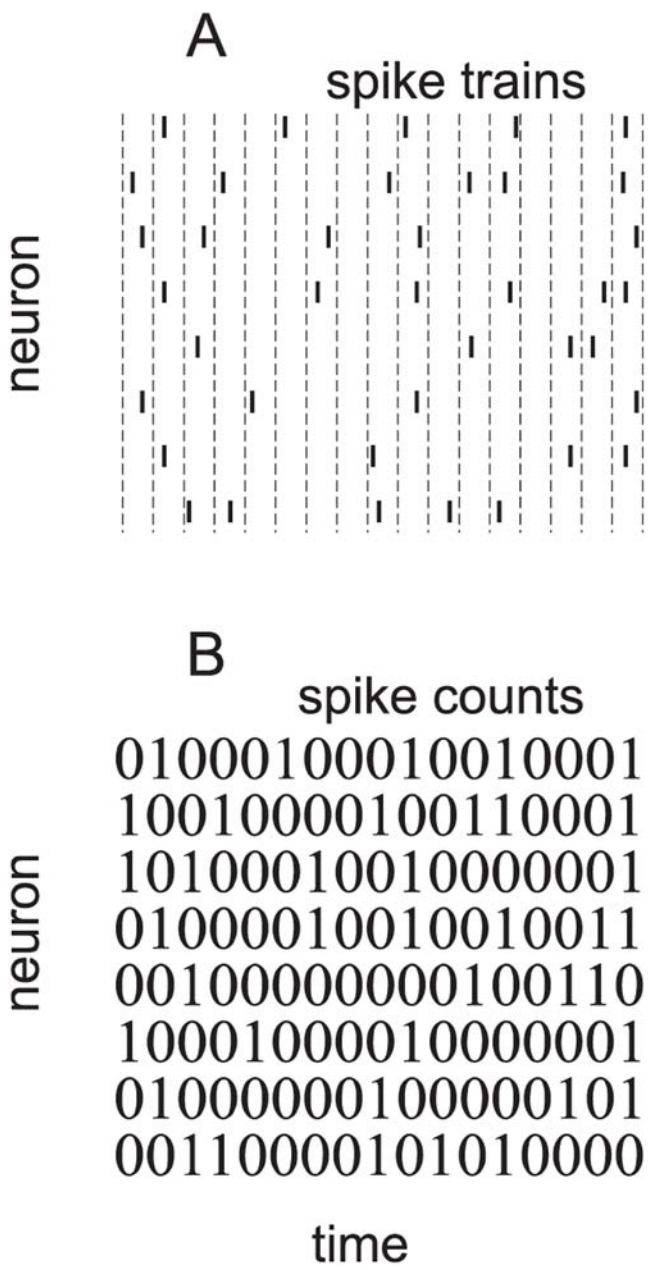

Figure 1. Transforming spike trains to spike count. (A) Spike rasters. Tick marks indicate spike times; different rows correspond to different neurons. The horizontal dashed lines are the bin boundaries. (B) Spike count in each bin. In this example the bins are small enough that there is at most one spike per bin, but this is not necessary—one could use bigger bins and have larger spike counts.
doi:10.1371/journal.pcbi.1000380.g001

These results have an important corollary: if one does an experiment and finds that Δ_N is increasing linearly with N, then one has no information at all about the true distribution. The flip side of this is more encouraging: if one can measure the true distribution for sufficiently large N that Δ_N saturates, as for the dashed blue line in Fig. 2, then there is a chance that extrapolation to large N is meaningful. The implications for the interpretation of experiments is, therefore, that one can gain information about large N behavior only if one can analyze data past the perturbative regime.

Under what conditions is a subsystem in the perturbative regime? The answer turns out to be simple: the size of the system,

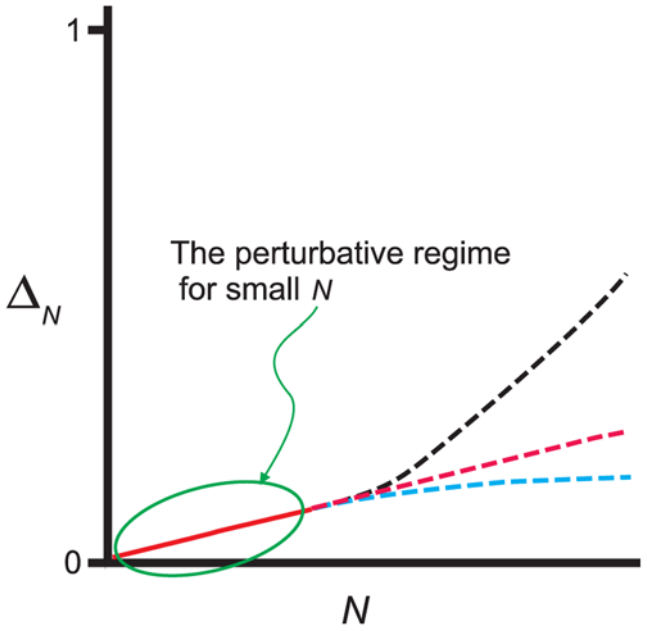

Figure 2. Cartoon illustrating the dependence of Δ_N on N. For small N there is always a perturbative regime in which Δ_N increases linearly with N (solid red line). When N becomes large, Δ_N may continue increasing with N (red and black dashed lines) or it may plateau (cyan dashed line), depending on p_{true}. The observation that Δ_N increases linearly with N does not, therefore, provide much, if any information about the large N behavior.
doi:10.1371/journal.pcbi.1000380.g002

N, times the average probability of observing a spike in a bin, must be small compared to 1. For example, if the average probability is 1/100, then a system will be in the perturbative regime if the number of neurons is small compared to 100. This last observation would seem to be good news: if we divide the spike trains into sufficiently small time bins and ignore temporal correlations, then we can model the data very well with a pairwise distribution. The problem with this, though, is that temporal correlations can be ignored only when time bins are large compared to the autocorrelation time. This leads to a kind of catch-22: pairwise models are guaranteed to work well (in the sense that they describe spike trains in which temporal correlations are ignored) if one uses small time bins, but small time bins is the one regime where ignoring temporal correlations is not a valid approximation.

In the next several sections we quantify the qualitative picture presented above: we write down an explicit expression for Δ_N, explain why it increases linearly with N when N is small, and provide additional tests, besides assessing the linearity of Δ_N, to determine whether or not one is in the perturbative regime.

Quantifying how well the pairwise model explains the data

A natural measure of the distance between p_{pair} and p_{true} is the Kullback-Leibler (KL) divergence [12], denoted $D_{KL}(p_{true}\|p_{pair})$ and defined as

$$D_{KL}(p_{true}\|p_{pair}) = \sum_{\mathbf{r}} p_{true}(\mathbf{r}) \log_2 \frac{p_{true}(\mathbf{r})}{p_{pair}(\mathbf{r})}. \quad (1)$$

The KL divergence is zero if the two distributions are equal; otherwise it is nonzero.

Although the KL divergence is a very natural measure, it is not easy to interpret (except, of course, when it is exactly zero). That is because a nonzero KL divergence tells us that $p_{pair} \neq p_{true}$, but it does not give us any real handle on how good, or bad, the pairwise model really is. To make sense of the KL divergence, we need something to compare it to. A reasonable reference quantity, used by a number of authors [7–9], is the KL divergence between the true distribution and the independent one, the latter denoted p_{ind}. The independent distribution, as its name suggests, is a distribution in which the variables are taken to be independent,

$$p_{ind}(r_1, \ldots, r_N) = \prod_i p_i(r_i), \quad (2)$$

where $p_i(r_i)$ is the distribution of the response of the i^{th} neuron, r_i. With this choice for a comparison, we define a normalized distance measure—a measure of how well the pairwise model explains the data—as

$$\Delta_N = \frac{D_{KL}(p_{true}\|p_{pair})}{D_{KL}(p_{true}\|p_{ind})}. \quad (3)$$

Note that the denominator in this expression, $D_{KL}(p_{true}\|p_{ind})$, is usually referred to as the multi-information [7,13,14].

The quantity Δ_N lies between 0 and 1, and measures how well a pairwise model does relative to an independent model. If it is 0, the pairwise model is equal to the true model ($p_{pair}(\mathbf{r}) = p_{true}(\mathbf{r})$); if it is near 1, the pairwise model offers little improvement over the independent model; and if it is exactly 1, the pairwise model is equal to the independent model ($p_{pair}(\mathbf{r}) = p_{ind}(\mathbf{r})$), and so offers no improvement.

How do we attach intuitive meaning to the two divergences $D_{KL}(p_{true}\|p_{pair})$ and $D_{KL}(p_{true}\|p_{ind})$? For the latter, we use the fact that, as is easy to show,

$$D_{KL}(p_{true}\|p_{ind}) = S_{ind} - S_{true}, \quad (4)$$

where S_{ind} and S_{true} are the entropies [15,16] of p_{ind} and p_{true}, respectively, defined, as usual, to be $S[p] = -\sum_{\mathbf{r}} p(\mathbf{r}) \log_2 p(\mathbf{r})$. For the former, we use the definition of the KL divergence to write

$$\begin{aligned} D_{KL}(p_{true}\|p_{pair}) &= -\sum_{\mathbf{r}} p_{true}(\mathbf{r}) \log_2 (p_{pair}(\mathbf{r})) \\ &\quad - S_{true} \equiv \tilde{S}_{pair} - S_{true}. \end{aligned} \quad (5)$$

The quantity $\tilde{S}_{pair}$ has the flavor of an entropy, although it is a true entropy only when p_{pair} is maximum entropy as well as pairwise (see Eq. (6) below). For other pairwise distributions, all we need to know is that $\tilde{S}_{pair}$ lies between S_{true} and S_{ind}. A plot illustrating the relationship between Δ_N, the two entropies S_{ind} and S_{true}, and the entropy-like quantity $\tilde{S}_{pair}$, is shown in Fig. 3.

Note that for pairwise maximum entropy models (or maximum entropy models for short), Δ_N has a particularly simple interpretation, since in this case $\tilde{S}_{pair}$ really is an entropy. Using S_{maxent} to denote the pairwise entropy of a maximum entropy model, for this case we have

$$\Delta_N = \frac{S_{maxent} - S_{true}}{S_{ind} - S_{true}}, \quad (6)$$

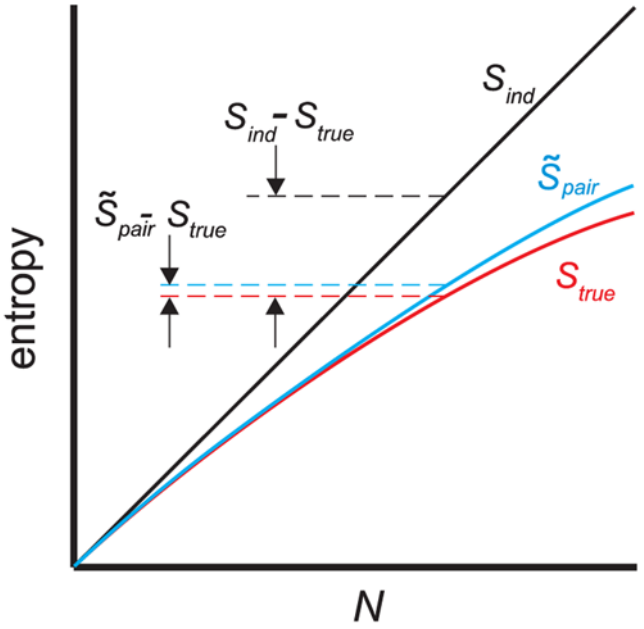

Figure 3. Schematic plot of S_{ind} (black line), $\tilde{S}_{pair}$ (cyan line) and S_{true} (red line). The better the pairwise model, the closer $\tilde{S}_{pair}$ is to S_{true}. This is reflected in the normalized distance measure, Δ_N, which is the distance between the red and cyan lines divided by the distance between the red and black lines.
doi:10.1371/journal.pcbi.1000380.g003

as is easy to see by inserting Eqs. (4) and (5) into (3). This expression has been used previously by a number of authors [7,9].

Δ_N in the perturbative regime

The extrapolation problem discussed above is the problem of determining Δ_N in the large N limit. This is hard to do in general, but there is a perturbative regime in which it is possible. The small parameter that defines this regime is the average number of spikes produced by the whole population of neurons in each time bin. It is given quantitatively by $N\bar{\nu}\delta t$ where δt is the bin size and $\bar{\nu}$ the average firing rate,

$$\bar{\nu} \equiv \frac{1}{N}\sum_i \nu_i, \tag{7}$$

with ν_i the firing rate of neuron i.

The first step in the perturbation expansion is to compute the two quantities that make up Δ_N: $D_{KL}(p_{true}\|p_{ind})$ and $D_{KL}(p_{true}\|p_{pair})$. As we show in the section "Perturbative Expansion" (Methods), these are given by

$$D_{KL}(p_{true}\|p_{ind}) = D^0_{KL}(p_{true}\|p_{ind}) + \mathcal{O}\left((N\bar{\nu}\delta t)^3\right) \tag{8a}$$

$$D_{KL}(p_{true}\|p_{pair}) = D^0_{KL}(p_{true}\|p_{pair}) + \mathcal{O}\left((N\bar{\nu}\delta t)^4\right), \tag{8b}$$

where

$$D^0_{KL}(p_{true}\|p_{ind}) = g_{ind}N(N-1)(\bar{\nu}\delta t)^2 \tag{9a}$$

$$D^0_{KL}(p_{true}\|p_{pair}) = g_{pair}N(N-1)(N-2)(\bar{\nu}\delta t)^3. \tag{9b}$$

Here and in what follows we use $\mathcal{O}((N\bar{\nu}\delta t)^n)$ to denote terms that are proportional to $(N\bar{\nu}\delta t)^n$ in the limit $N\bar{\nu}\delta t \to 0$. The N-dependence in Eq. (9a) has been noted previously [7], although the authors did not compute the prefactor, g_{ind}.

The prefactors g_{ind} and g_{pair}, which are given explicitly in Eqs. (42) and (44), depend on the low order statistics of the spike trains: g_{ind} depends on the second order normalized correlation coefficients, g_{pair} depends on the second and third order normalized correlation coefficients (the normalized correlation coefficients are defined in Eq. (16) below), and both depend on the firing rates of the individual cells. The details of that dependence, however, are not important for now; what is important is that g_{ind} and g_{pair} are independent of N and $\bar{\nu}\delta t$ (at least on average; see next section).

Inserting Eq. (8) into Eq. (3) (into the definition of Δ_N) and using Eq. (9), we arrive at our main result,

$$\Delta_N = \Delta^0_N + \mathcal{O}\left((N\bar{\nu}\delta t)^2\right) \tag{10a}$$

$$\Delta^0_N = \frac{g_{pair}}{g_{ind}}(N-2)\bar{\nu}\delta t \tag{10b}$$

Note that in the regime $N\bar{\nu}\delta t \ll 1$, Δ_N is necessarily small. This explains why, in an analytic study of non-pairwise model in which $N\bar{\nu}\delta t$ was small, Shlens et al. found that Δ_N was rarely greater than 0.1 [8].

We refer to quantities with a superscript zero as "zeroth order." Note that, via Eqs. (4) and (5), we can also define zeroth order entropies,

$$S^0_{true} \equiv S_{ind} - D^0_{KL}(p_{true}\|p_{ind}) \tag{11a}$$

$$\tilde{S}^0_{pair} \equiv S_{ind} - D^0_{KL}(p_{true}\|p_{ind}) + D^0_{KL}(p_{true}\|p_{pair}). \tag{11b}$$

These quantities are important primarily because differences between them and the actual entropies indicate a breakdown of the perturbation expansion (see in particular Fig. 4 below).

Assuming, as discussed in the next section, that g_{ind} and g_{pair} are approximately independent of N, $\bar{\nu}$, and δt, Eq. (10) tells us that Δ_N scales linearly with N in the perturbative regime—the regime in which $N\bar{\nu}\delta t \ll 1$. The key observation about this scaling is that it is independent of the details of the true distribution, p_{true}. This has a very important consequence, one that has major implications for experimental data: if one does an experiment with small $\bar{\nu}\delta t$ and finds that Δ_N is proportional to $N-2$, then the system is, with very high probability, in the perturbative regime, and one does not know whether p_{pair} will remain close to p_{true} as N increases. What this means in practical terms is that if one wants to know whether a particular pairwise model is a good one for large systems, it is necessary to consider values of N that are significantly greater than N_c, where

$$N_c \equiv \frac{1}{\bar{\nu}\delta t}. \tag{12}$$

We interpret N_c as the value at which there is a crossover in the behavior of the pairwise model. Specifically, if $N \ll N_c$, the system is in the perturbative regime and the pairwise model is not informative about the large N behavior, whereas if $N \gg N_c$, the

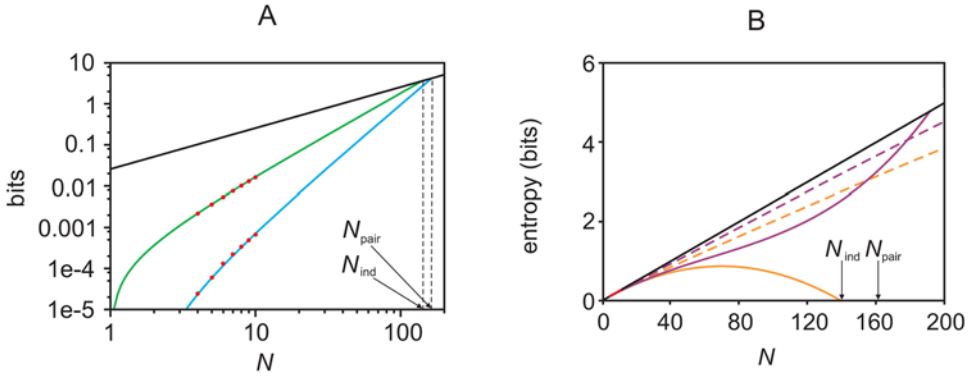

Figure 4. Cartoon showing extrapolations of the zeroth order KL divergences and entropies (see Eqs. (9) and (11)). These extrapolations illustrate why the two natural quantities derived from them, N_{ind} and N_{pair}, occur beyond the point at which the extrapolation is meaningful. (A) Extrapolations on a log-log scale. Black: S_{ind}; green: $D^0_{KL}(p_{true}\|p_{ind})$; cyan: $D^0_{KL}(p_{true}\|p_{pair})$. The red points are the data. The points N_{ind} and N_{pair} label the intersections of the two extrapolations with the independent entropy, S_{ind}. (B) Extrapolation of the entropies rather than the KL divergences, plotted on a linear-linear scale. The data, again shown in red, is barely visible in the lower left hand corner. Black: S_{ind}; solid orange: S^0_{true}; solid maroon: $\tilde{S}^0_{pair}$. The dashed orange and maroon lines are the extrapolations of the true entropy and true pairwise "entropy", respectively.
doi:10.1371/journal.pcbi.1000380.g004

system is in a regime in which it may be possible to make inferences about the behavior of the full system.

The prefactors, g_{ind} and g_{pair}

As we show in Methods (see in particular Eqs. (42) and (44)), the prefactors g_{ind} and g_{pair} depend on which neurons out of the full population are used. Consequently, these quantities fluctuate around their true values (in the sense that different subpopulations produce different values of g_{ind} and g_{pair}), where "true" refers to an average over all possible N-neuron sub-populations. Here we assume that the N neurons are chosen randomly from the full population, so any set of N neurons provides unbiased estimates of g_{ind} and g_{pair}. In our simulations, the fluctuations were small, as indicated by the small error bars on the blue points in Fig. 5. However, in general the size of the fluctuations is determined by the range of firing rates and correlation coefficients, with larger ranges producing larger fluctuations.

Because N does not affect the mean values of g_{ind} and g_{pair}, it is reasonable to think of these quantities—or at least their true values—as being independent of N. They are also independent of $\overline{\nu}$, again modulo fluctuations. Finally, as we show in the section "Bin size and the correlation coefficients" (Methods), g_{ind} and g_{pair} are independent of δt in the limit that δt is small compared to the width of the temporal correlations among neurons. We will assume this limit applies here. In sum, then, to first approximation, g_{ind} and g_{pair} are independent of our three important quantities: N, $\overline{\nu}$, and δt. Thus, we treat them as effectively constant throughout our analysis.

The dangers of extrapolation

Although the behavior of Δ_N in the perturbative regime does not tell us much about its behavior at large N, it is possible that other quantities that can be calculated in the perturbative regime, g_{ind}, g_{pair}, and S_{ind} (the last one exactly), are informative, as others have suggested [7]. Here we show that this is not the case—they also are uninformative.

The easiest way to relate the perturbative regime to the large N regime is to ignore the corrections in Eqs. (8a) and (8b), extrapolate the expressions for the zeroth order terms, and ask what their large N behavior tells us. Generic versions of these extrapolations, plotted on a log-log scale, are shown in Fig. 4A, along with a plot of the independent entropy, S_{ind} (which is necessarily linear in N). The first thing we notice about the extrapolations is that they do not, technically, have a large N behavior: one terminates at the point labeled N_{ind}, which is where $D^0_{KL}(p_{true}\|p_{ind}) = S_{ind}$ (and thus, via Eq. (0a), $S^0_{true} = 0$; continuing the extrapolation implies negative true zeroth order entropy); the other at the point labeled N_{pair}, which is where $D^0_{KL}(p_{true}\|p_{pair}) = S_{ind}$ (and thus, via Eq. (5) and the fact that $\tilde{S}^0_{pair} \le S_{ind}$, $S^0_{true} \le 0$).

Despite the fact that the extrapolations end abruptly, they still might provide information about the large N regime. For example, N_{pair} and/or N_{ind} might be values of N at which something interesting happens. To see if this is the case, in Fig. 4B we plot the naive extrapolations of $\tilde{S}_{pair}$ and S_{true} (that is, the zeroth order quantities given in Eq. (11), $\tilde{S}^0_{pair}$ and S^0_{true}), on a linear-linear plot, along with S_{ind}. This plot contains no new information compared to Fig. 4A, but it does elucidate the meaning of the extrapolations. Perhaps its most striking feature is that the naive extrapolation of S_{true} has a decreasing portion. As is easy to show mathematically, entropy cannot decrease with N (intuitively, that is because observing one additional neuron cannot decrease the entropy of previously observed neurons). Thus, N_{ind}, which occurs well beyond the point where the naive extrapolation of S_{true} is decreasing, has essentially no meaning, something that has been pointed out previously by Bethge and Berens [10]. The other potentially important value of N is N_{pair}. This, though, suffers from a similar problem: when $N = N_{pair}$, S^0_{true} is negative.

How do the naively extrapolated entropies—the solid lines in Fig. 4B—compare to the actual entropies? To answer this, in Fig. 4B we show the true behavior of S_{true} and $\tilde{S}_{pair}$ versus N (dashed lines). Note that S_{true} is asymptotically linear in N, even though the neurons are correlated, a fact that forces $\tilde{S}_{pair}$ to be linear in N, as it is sandwiched between S_{true} and S_{ind}. (The asymptotically linear behavior of S_{true} is typical, even in highly correlated systems. Although this is not always appreciated, it is easy to show; see the section "The behavior of the true entropy in the large N limit," Methods.) Comparing the dashed and solid lines, we see that the naively extrapolated and true entropies, and thus the naively extrapolated and true values of Δ_N, have extremely different behavior. This further suggests that there is very little connection between the perturbative and large N regimes.

In fact, these observations follow directly from the fact that g_{ind} and g_{pair} depend only on correlation coefficients up to third order

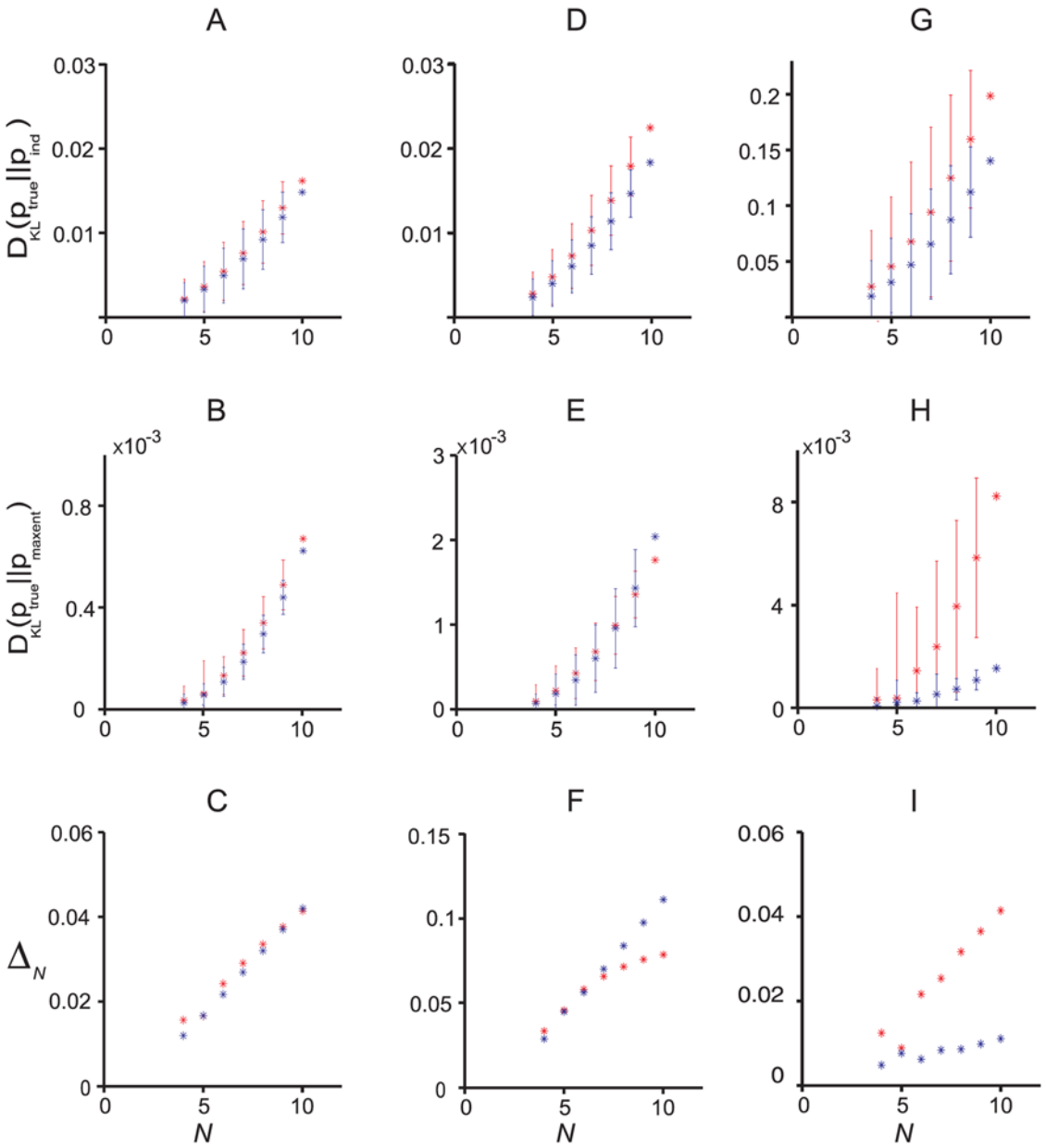

Figure 5. The N dependence of the KL divergences and the normalized distance measure, Δ_N. Data was generated from a third order model, as explained in the section "Generating synthetic data" (Methods), and fit to pairwise maximum entropy models and independent models. All data points correspond to averages over marginalizations of the true distribution (see text for details). The red points were computed directly using Eqs. (1), (3) and (4); the blue points are the zeroth order estimates, $D^0_{KL}(p_{true}\|p_{ind})$, $D^0_{KL}(p_{true}\|p_{pair})$, and Δ^0_N, in rows 1, 2 and 3, respectively. The three columns correspond to $\overline{\nu}\delta t = 0.024$, 0.029, and 0.037, from left to right. (A, B, C) ($\overline{\nu}\delta t = 0.024$). Predictions from the perturbative expansion are in good agreement with the measurements up to $N = 10$, indicating that the data is in the perturbative regime. (D, E, F) ($\overline{\nu}\delta t = 0.029$). Predictions from the perturbative expansion are in good agreement with the measurements up to $N = 7$, indicating that the data is only partially in the perturbative regime. (G, H, I) ($\overline{\nu}\delta t = 0.037$). Predictions from the perturbative expansion are not in good agreement with the measurements, even for small N, indicating that the data is outside the perturbative regime.
doi:10.1371/journal.pcbi.1000380.g005

(see Eqs. (42) and (44)) whereas the large N behavior depends on correlations at all orders. Thus, since g_{ind} and g_{pair} tell us very little, if anything, about higher order correlations, it is not surprising that they tell us very little about the behavior of Δ_N in the large N limit.

Numerical simulations

To check that our perturbation expansions, Eqs. (8–10), are correct, and to investigate the regime in which they are valid, we performed numerical simulations. We generated, from synthetic data, a set of true distributions, computed the true distance measures, $D_{KL}(p_{true}\|p_{ind})$, $D_{KL}(p_{true}\|p_{pair})$, and Δ_N numerically, and compared them to the zeroth order ones, $D^0_{KL}(p_{true}\|p_{ind})$, $D^0_{KL}(p_{true}\|p_{pair})$, and Δ^0_N. If the perturbation expansion is valid, then the true values should be close to the zeroth order values whenever $N\overline{\nu}\delta t$ is small. The results are shown in Fig. 5, and that is, indeed, what we observed. Before discussing that figure, though, we explain our procedure for constructing true distributions.

The set of true distributions we used were generated from a third order model (so named because it includes up to third order interactions). This model has the form

$$p_{true}(r_1,\ldots,r_{N^*}) = \frac{1}{Z_{true}}\exp\left[\sum_i h_i^{true} r_i + \sum_{i<j} J_{ij}^{true} r_i r_j + \sum_{i<j<k} K_{ijk}^{true} r_i r_j r_k\right] \quad (13)$$

where Z_{true} is a normalization constant, chosen to ensure that the probability distribution sums to 1, and the sums over i, j and k run from 1 to N^*. The parameters h_i^{true},J_{ij}^{true} and K_{ijk}^{true} were chosen by sampling from distributions (see the section "Generating synthetic data," Methods), which allowed us to generate many different true distributions. In all of our simulations we calculate the relevant quantities directly from Eq. (13) . Consequently, we do not have to worry about issues of finite data, as would be the case in realistic experiments.

For a particular simulation (corresponding to a column in Fig. 5), we generated a true distribution with $N^*=15$, randomly chose 5 neurons, and marginalized over them. This gave us a 10-neuron true distribution. True distributions with $N<10$ were constructed by marginalizing over additional neurons within our 10-neuron population. To achieve a representative sample, we considered all possible marginalizations (of which there are 10 choose N, or $10!/[N!(10-N)!]$). The results in Fig. 5 are averages over these marginalizations.

For neural data, the most commonly used pairwise model is the maximum entropy model. Therefore, we use that one here. To emphasize the maximum entropy nature of this model, we replace the label "*pair*" that we have been using so far with "*maxent*." The maximum entropy distribution has the form

$$p_{maxent}(\mathbf{r}) = \frac{1}{Z}\exp\left[\sum_i h_i r_i + \sum_{i<j} J_{ij} r_i r_j\right]. \quad (14)$$

Fitting this distribution requires that we choose the h_i and J_{ij} so that the first and second moments match those of the true distribution. Quantitatively, these conditions are

$$\langle r_i\rangle_{maxent} = \langle r_i\rangle_{true} \quad (15a)$$

$$\langle r_i r_j\rangle_{maxent} = \langle r_i r_j\rangle_{true} \quad (15b)$$

where the angle brackets, $\langle\ldots\rangle_{maxent}$ and $\langle\ldots\rangle_{true}$, represent averages with respect to p_{maxent} and p_{true}, respectively. Once we have h_i and J_{ij} that satisfy Eq. (15), we calculate the KL divergences, Eqs. (1) and (4), and use those to compute Δ_N.

The results are shown in Fig. 5. The rows correspond to our three quantities of interest: $D_{KL}(p_{true}\|p_{ind})$, $D_{KL}(p_{true}\|p_{pair})$, and Δ_N (top to bottom). The columns correspond to different values of $\overline{\nu}\delta t$, with the smallest $\overline{\nu}\delta t$ on the left and the largest on the right. Red circles are the true values of these quantities; blue ones are the zeroth order predictions from Eqs. (9) and (10b).

As suggested by our perturbation analysis, the smaller the value of $\overline{\nu}\delta t$, the larger the value of N for which agreement between the true and zeroth order values is good. Our simulations corroborate this: the left column of Fig. 5 has $\overline{\nu}\delta t=0.024$, and agreement is almost perfect out to $N=10$; the middle column has $\overline{\nu}\delta t=0.029$, and agreement is almost perfect out to $N=7$; and the right column has $\overline{\nu}\delta t=0.037$, and agreement is not good for any value of N. Note that the perturbation expansion breaks down for values of N well below N_C (defined in Eq.(12)): in the middle column of Fig. 5 it breaks down when $N/N_c\approx 0.23$, and in the right column it breaks down when $N/N_c\approx 0.15$. This is not, however, especially surprising, as the perturbation expansion is guaranteed to be valid only if $N/N_c\ll 1$.

These results validate the perturbation expansions in Eqs. (8) and (10), and show that those expansions provide sensible predictions—at least for some parameters. They also suggest a natural way to assess the significance of one's data: plot $D_{KL}(p_{true}\|p_{ind})$, $D_{KL}(p_{true}\|p_{pair})$, and Δ_N versus N, and look for agreement with the predictions of the perturbation expansion. If agreement is good, as in the left column of Fig. 5, then one is in the perturbative regime, and one knows very little about the true distribution. If, on the other hand, agreement is bad, as in the right column, then one is out of the perturbative regime, and it may be possible to extract meaningful information about the relationship between the true and pairwise models.

That said, the qualifier "at least for some parameters" is an important one. This is because the perturbation expansion is essentially an expansion that depends on the normalized correlation coefficients, and there is an underlying assumption that they don't exhibit pathological behavior. The k^{th} order normalized correlation coefficient for the distribution $p(\mathbf{r})$, denoted $\rho^p_{i_1 i_2\ldots i_k}$, is written

$$\rho^p_{i_1 i_2\ldots i_k} = \frac{\left\langle\left(r_{i_1}-\langle r_{i_1}\rangle_p\right)\left(r_{i_2}-\langle r_{i_2}\rangle_p\right)\ldots\left(r_{i_k}-\langle r_{i_k}\rangle_p\right)\right\rangle_p}{\langle r_{i_1}\rangle_p\langle r_{i_2}\rangle_p\ldots\langle r_{i_k}\rangle_p}. \quad (16)$$

A potentially problematic feature of the correlation coefficients is that the denominator is a product over mean activities. If the mean activities are small, the denominator can become very small, leading to very large correlation coefficients. Although our perturbation expansion is always valid for sufficiently small time bins (because the correlation coefficients eventually becomes independent of bin size; see the section "Bin size and the correlation coeffcients," Methods), "sufficiently small" can depend in detail on the parameters. For instance, at the maximum population size tested ($N=10$) and for the true distributions that had $\overline{\nu}\delta t<0.03$, the absolute error of the prediction had a median of approximately 16%. However, about 11% of the runs had errors larger than 60%. Thus, the exact size of the small parameter at which the perturbative expansion breaks down can depend on the details of the true distribution.

External fields and pairwise couplings have a simple dependence on firing rates and correlation coefficients in the perturbative regime

Estimation of the KL divergences and Δ_N from real data can be hard, in the sense that it takes a large amount of data for them to converge to their true values. In addition, as discussed above, in the section "The prefactors g_{ind} and g_{pair}", there are fluctuations in Δ_N associated with finite subsampling of the full population of neurons. Those fluctuations tend to keep Δ_N from being purely linear, as can seen, for example, in the blue points in Fig. 5F and 5I. We therefore provide a second set of relationships that can be used to determine whether or not a particular data set is in the perturbative regime. These relationships are between the parameters of the maximum entropy model, the h_i and J_{ij}, and the mean activity and normalized second order correlation coefficient (the latter defined in Eq. (19) below).

Since the quantity $\bar{\nu}\delta t$ plays a central role in our analysis, we replace it with a single parameter, which we denote δ,

$$\delta \equiv \bar{\nu}\delta t. \tag{17}$$

In terms of this parameter, we find (using the same perturbative approach that led us to Eqs. (8–10); see the section "External fields, pairwise couplings and moments," Methods), that

$$h_i = -\log\left[\langle r_i\rangle^{-1} - 1\right] + \mathcal{O}(N\delta) \tag{18a}$$

$$J_{ij} = \log\left[1 + \rho_{ij}\right] + \mathcal{O}(N\delta) \tag{18b}$$

where ρ_{ij}, the normalized second order correlation coefficient, is defined in Eq. (16) with $k=2$; it is given explicitly by

$$\rho_{ij} = \frac{\langle r_i r_j\rangle - \langle r_i\rangle\langle r_j\rangle}{\langle r_i\rangle\langle r_j\rangle}. \tag{19}$$

(We don't need a superscript on ρ or a subscript on the angle brackets because the first and second moments are the same under the true and pairwise distributions.) Equation (18a) can be reconstructed from the low firing rate limit of analysis carried out by Sessak and Monasson [17], as can the first three terms in the expansion of the log in Eq. (18b).

Equation (18) tells us that the N-dependence of the h_i and J_{ij}, the external fields and pairwise couplings, is very weak. In Fig. 6 we confirm this through numerical simulations. Equation (18b) also provides additional information—it gives us a functional relationship between the pairwise couplings and the normalized pairwise correlations function, ρ_{ij}. In Fig. 7A–C we plot the pairwise couplings, J_{ij}, versus the normalized pairwise correlation coefficient, ρ_{ij} (blue dots), along with the prediction from Eq. (18b) (black line). Consistent with our predictions, the data in Fig. 7A–C essentially follows a line—the line given by Eq. (18b).

A relationship between the pairwise couplings and the correlations coefficients has been sought previously, but for the more standard Pearson correlation coefficient [7,9,11]. Our analysis explains why it was not found. The Pearson correlation coefficient, denoted c_{ij}, is given by

$$c_{ij} \equiv \frac{\langle r_i r_j\rangle - \langle r_i\rangle\langle r_j\rangle}{\left[\left(\langle r_i^2\rangle - \langle r_i\rangle^2\right)\left(\langle r_j^2\rangle - \langle r_j\rangle^2\right)\right]^{1/2}}. \tag{20}$$

In the small $\langle r_i\rangle$ limit—the limit of interest—the right hand side of Eq. (20) is approximately equal to $\left(\langle r_i\rangle\langle r_j\rangle\right)^{1/2}\rho_{ij}$. Because $\left(\langle r_i\rangle\langle r_j\rangle\right)^{1/2}$ depends on the external fields, h_i and h_j (see Eq. (18a)) *and* there is a one-to-one relationship between ρ_{ij} and J_{ij} (Eq. (18b)), there can't be a one-to-one relationship between c_{ij} and J_{ij}. We verify the lack of a relationship in Fig. 7D and 7E, where we again plot J_{ij}, but this time versus the standard correlation coefficient, c_{ij}. As predicted, the data in Fig. 7D and 7E is scattered over two dimensions. This suggests that ρ_{ij}, not c_{ij}, is the natural measure of the correlation between two neurons when they have a binary representation, something that has also been suggested by Amari based on information-geometric arguments [18].

Note that the lack of a simple relationship between the pairwise couplings and the standard correlation coefficient has been a major motivation in building maximum entropy models [7,11]. This is for good reason: if there is a simple relationship, knowing the J_{ij}'s adds essentially nothing. Thus, plotting J_{ij} versus ρ_{ij} (but not c_{ij}) is an important test of one's data, and if the two quantities fall on the curve predicted by Eq. (18b), the maximum entropy model is adding very little information, if any.

As an aside, we should point out that the N-dependence is a function of the variables used to represent the firing patterns. Here we use 0 for no spike and 1 for one or more spikes, but another, possibly more common, representation, derived from the Ising model and used in a number of studies [7,9,11], is to use -1 and $+1$ rather than 0 and 1. This amounts to making the change of variables $s_i = 2r_i - 1$. In terms of s_i, the maximum entropy model has the form $p(\mathbf{r}) \sim \exp\left[\sum_i h_i^{ising} s_i + \sum_{i<j} J_{ij}^{ising} s_i s_j\right]$ where h_i^{ising} and J_{ij}^{ising} are given by

$$h_i^{ising} = \frac{h_i}{2} + \sum_{j\neq i}\frac{J_{ij}}{4} \tag{21a}$$

$$J_{ij}^{ising} = \frac{J_{ij}}{4}. \tag{21b}$$

The second term on the right side of Eq. (21a) is proportional to $N-1$, which means the external fields in the Ising representation acquire a linear N-dependence that was not present in our 0/1 representation. The two studies that reported the N-dependence of the external fields [7,9] used this representation, and, as predicted by our analysis, the external fields in those studies had a component that was linear in N.

Is there anything wrong with using small time bins?

An outcome of our perturbative approach is that our normalized distance measure, Δ_N, is linear in bin size (see Eq. (10b)). This suggests that one could make the pairwise model look better and better simply by making the bin size smaller and smaller. Is there anything wrong with this? The answer is yes, for reasons discussed above (see the the section "The extrapolation problem"); here we emphasize and expand on this issue, as it is an important one for making sense of experimental results.

The problem arises because what we have been calling the "true" distribution is not really the true distribution of spike trains. It is the distribution assuming independent time bins, an assumption that becomes worse and worse as we make the bins smaller and smaller. (We use this potentially confusing nomenclature primarily because all studies of neuronal data carried out so far have assumed temporal independence, and compared the pairwise distribution to the temporally independent—but still correlated across neurons—distribution [7–9,11]. In addition, the correct name "true under the assumption of temporal independence," is unwieldy.) Here we quantify how much worse. In particular, we show that if one uses time bins that are small compared to the characteristic correlation time in the spike trains, the pairwise model will not provide a good description of the data. Essentially, we show that, when the time bins are too small, the error one makes in ignoring temporal correlations is larger than the error one makes in ignoring correlations across neurons.

As usual, we divide time into bins of size δt. However, because we are dropping the independence assumption, we use $\mathbf{r}^t$, rather than $\mathbf{r}$, to denote the response in bin t. The full probability distribution over all time bins is denoted $\wp\left(\mathbf{r}^1,\ldots,\mathbf{r}^M\right)$. Here M is the number of bins; it is equal to $T/\delta t$ for spike trains of length T. If time bins are approximately independent and the distribution of

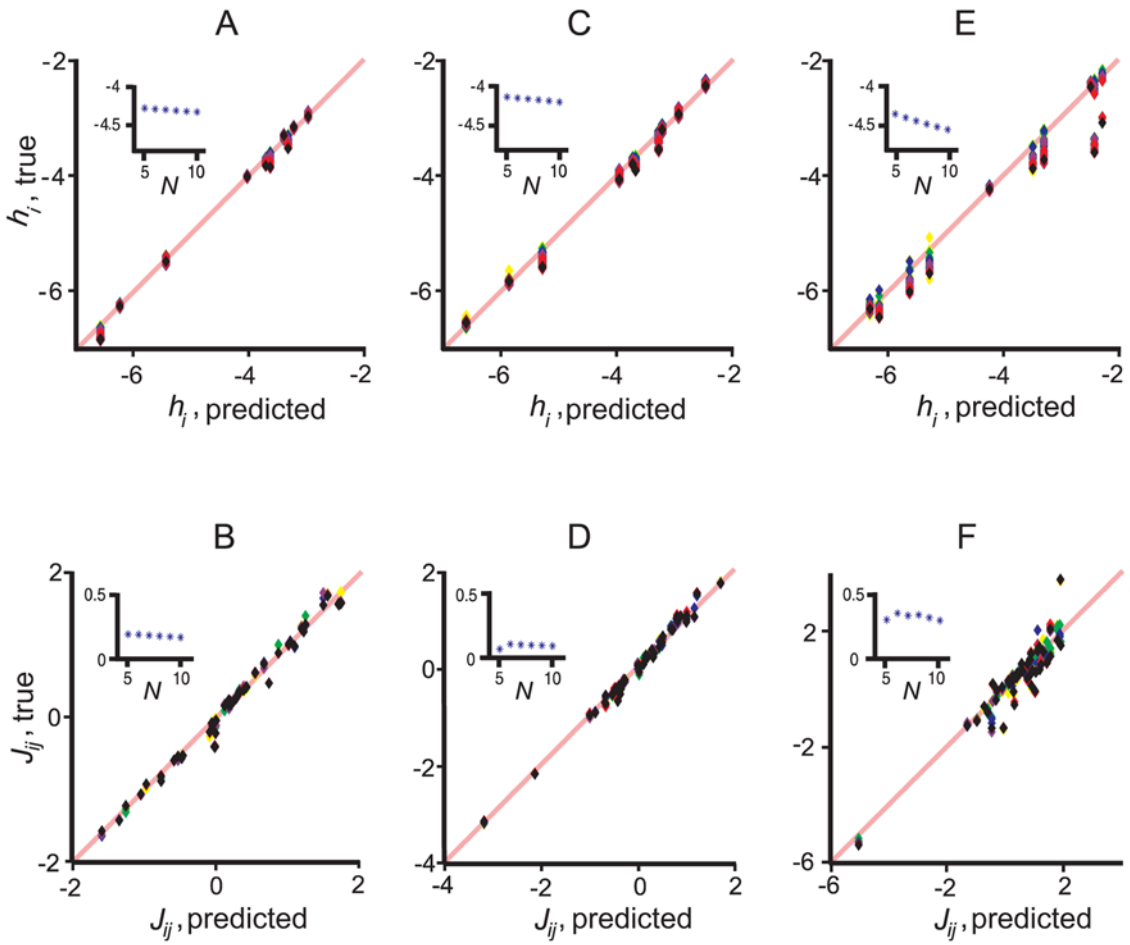

Figure 6. The true external fields and pairwise interactions compared with the predictions of the perturbation expansion. The top row shows the true external fields, h_i, versus those predicted from Eq. (18a), and the bottom row shows the true pairwise interaction, J_{ij}, versus those predicted from Eq. (18b). Values of N ranging from 5 to 10 are shown, with different colors corresponding to different Ns. For each value of N, data is shown for 45 realization of the true distribution. Insets show the N-dependence of the mean external fields (top) and mean pairwise interactions (bottom). The three columns correspond exactly to the columns in Fig. 5. (A, B) ($\overline{\nu}\delta t = 0.024$). There is a very good match between the true and predicted values of both external fields and pairwise interactions. (C, D) ($\overline{\nu}\delta t = 0.029$). Even though $\overline{\nu}\delta t$ has increased, the match is still good. (E, F) ($\overline{\nu}\delta t = 0.037$). The true and predicted external fields and pairwise interactions do not match as well as the cases shown in (A, B, C, D). There is also now a stronger N-dependence in the mean external fields compared to (A) and (B). The N-dependence of the pairwise interactions in (F) is weaker than that of the external fields, but still notably stronger than the ones in (B) and (D). This indicates that the perturbative expansion is starting to break down.
doi:10.1371/journal.pcbi.1000380.g006

$\mathbf{r}^t$ is the same for all t (an assumption we make for convenience only, but do not need; see the section "Extending the normalized distance measure to the time domain," Methods), we can write

$$\wp\left(\mathbf{r}^1,\ldots,\mathbf{r}^M\right) \approx \prod_t p_{true}(\mathbf{r}^t). \tag{22}$$

Furthermore, if the pairwise model is a good one, we have

$$p_{true}(\mathbf{r}^t) \approx p_{pair}(\mathbf{r}^t). \tag{23}$$

Combining Eqs. (22) and Eq. (23) then gives us an especially simple expression for the full probability distribution: $\wp\left(\mathbf{r}^1,\ldots,\mathbf{r}^M\right) \approx \Pi_t\, p_{pair}(\mathbf{r}^t)$.

The problem with small time bins lies in Eq. (22): the right hand side is a good approximation to the true distribution when the time bins are large compared to the spike train correlation time, but it is a bad approximation when the time bins are small (because adjacent time bins become highly correlated). To quantify how bad, we compare the error one makes assuming independence across time to the error one makes assuming independence across neurons. The ratio of those two errors, denoted γ, is given by

$$\gamma = \frac{D_{KL}\left(\wp\left(\mathbf{r}^1,\ldots,\mathbf{r}^M\right)\Big\|\prod_t p_{pair}(\mathbf{r}^t)\right)}{MD_{KL}(p(\mathbf{r})\|p_{ind}(\mathbf{r}))}. \tag{24}$$

It is relatively easy to compute γ in the limit of small time bins (see the section "Extending the normalized distance measure to the time domain," Methods), and we find that

$$\gamma = \Delta_N + (M-1) + \frac{\log_2 M}{g_{ind}(N-1)\delta}. \tag{25}$$

As expected, this reduces to our old result, Δ_N, when there is only one time bin ($M=1$). When M is larger than 1, however, the

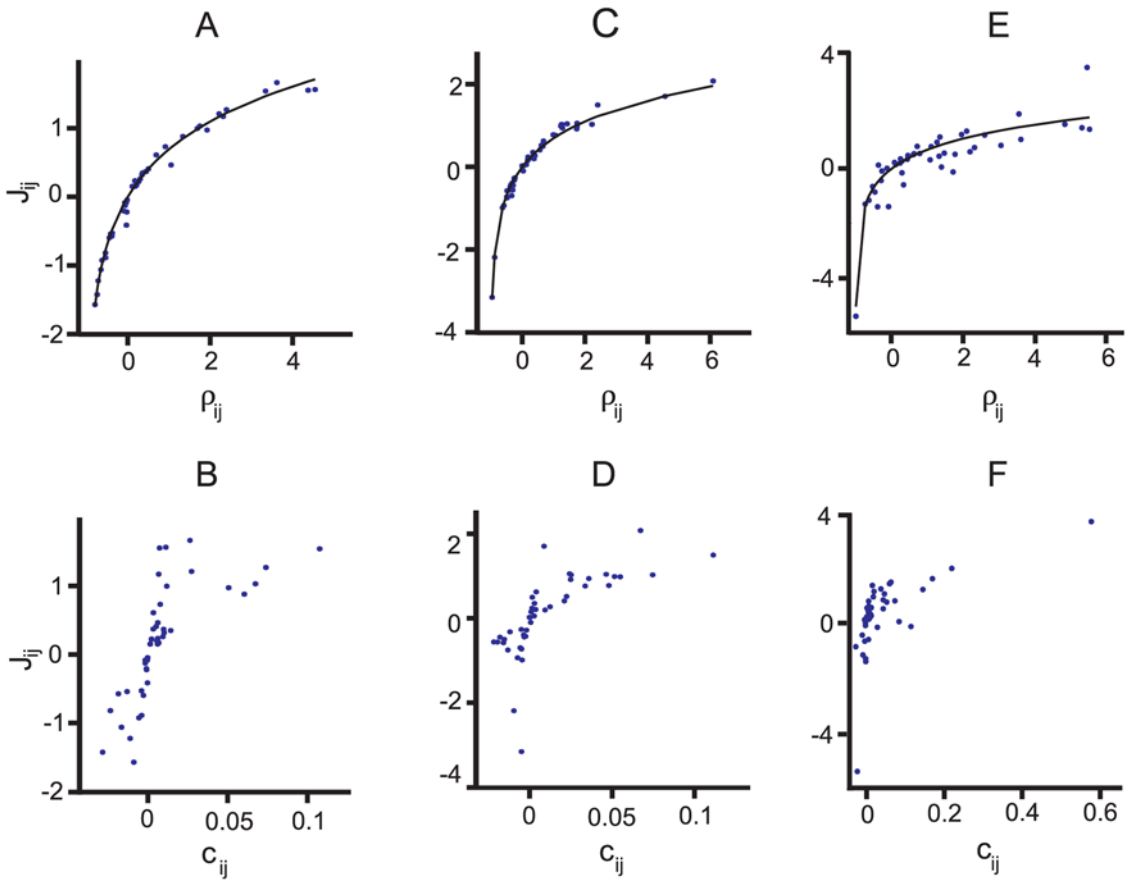

Figure 7. The relation between pairwise couplings and pairwise correlations. This figure shows that there is a simple relation between J_{ij} and ρ_{ij}, but not between J_{ij} and c_{ij}. (A, C, E) J_{ij} versus the normalized coefficients, ρ_{ij} (blue points), along with the predicted relationship, via Eq. (18b) (black line). (B, D, F) J_{ij} versus the Pearson correlation coefficients, c_{ij}, Eq. (26) (blue points). The three columns correspond exactly to the columns in Fig. 5 from left to right; that is, $\overline{\nu}\delta t = 0.024$ for (A, B), $\overline{\nu}\delta t = 0.029$ for (C, D), and $\overline{\nu}\delta t = 0.037$ for (E, F). The prediction in the top row (black line) matches the data well, even in the rightmost column.
doi:10.1371/journal.pcbi.1000380.g007

second term is always at least one, and for small bin size, the third term is much larger than one. Consequently, if we use bins that are small compared to the temporal correlation time of the spike trains, the pairwise model will do a very bad job describing the full, temporally correlated spike trains.

Discussion

Probability distributions over the configurations of biological systems are extremely important quantities. However, because of the large number of interacting elements comprising such systems, these distributions can almost never be determined directly from experimental data. Using parametric models to approximate the true distribution is the only existing alternative. While such models are promising, they are typically applied only to small subsystems, not the full system. This raises the question: are they good models of the full system?

We answered this question for a class of parametric models known as pairwise models. We focused on a particular application, neuronal spike trains, and our main result is as follows: if one were to record spikes from multiple neurons, use sufficiently small time bins and a sufficiently small number of cells, and assume temporal independence, then a pairwise model will almost always succeed in matching the true (but temporally independent) distribution—whether or not it would match the true (but still temporally independent) distribution for large time bins or a large number of cells. In other words, pairwise models in the "sufficiently small" regime, what we refer to as the perturbative regime, have almost no predictive value for what will happen with large populations. This makes extrapolation from small to large systems dangerous.

This observation is important because pairwise models, and in particular pairwise maximum entropy models, have recently attracted a great deal of attention: they have been applied to salamander and guinea pig retinas [7], primate retina [8], primate cortex [9], cultured cortical networks [7], and cat visual cortex [11]. These studies have mainly operated close to the perturbative regime. For example, Schneidman et al. [7] had $N\overline{\nu}\delta t \approx 0.35$ (for the data set described in their Fig. 5), Tang et al. [9] had $N\overline{\nu}\delta t \approx 0.06$ to 0.4 (depending on the preparation), and Yu et al. [11] had $N\overline{\nu}\delta t \approx 0.2$. For these studies, then, it would be hard to justify extrapolating to large populations.

The study by Shlens et al. [8], on the other hand, might be more amenable to extrapolation. This is because spatially localized visual patterns were used to stimulate retinal ganglion cells, for which a nearest neighbor maximum entropy models provided a good fit to their data. (Nearest neighbor means J_{ij} is zero unless neuron i and neuron j are adjacent.) Our analysis still applies, but, since all but the nearest neighbor correlations are zero, many of

the terms that make up g_{ind} and g_{pair} vanish (see Eqs. (42) and (44)). Consequently, the small parameter in the perturbative expansion becomes $K\bar{\nu}\delta t$ (rather than $N\bar{\nu}\delta t$), where K is the number of nearest neighbors. Since K is fixed, independent of the population size, the small parameter will not change as the population size increases. Thus, Shlens et al.may have tapped into the large population behavior even though they considered only a few cells at a time in their analysis. Indeed, they have recently extended their analysis to more than 100 neurons, and they still find that nearest neighbor maximum entropy models provide very good fits to the data [19].

Time bins and population size

That the pairwise model is always good if $N\bar{\nu}\delta t$ is sufficiently small has strong implications: if we want to build a good model for a particular N, we can simply choose a bin size that is small compared to $1/N\bar{\nu}$. However, one of the assumptions in all pairwise models used on neural data is that bins at different times are independent. This produces a tension between small time bins and temporal independence: small time bins essentially ensure that a pairwise model will provide a close approximation to a model with independent bins, but they make adjacent bins highly correlated. Large time bins come with no such assurance, but they make adjacent bins independent. Unfortunately, this tension is often unresolvable in large populations, in the sense that pairwise models are assured to work only up to populations of size $1/(\bar{\nu}\tau_{\rm corr})$ where $\tau_{\rm corr}$ is the typical correlation time. Given that $\bar{\nu}$ is at least several Hz, for experimental paradigms in which the correlation time is more than a few hundred ms, $1/(\bar{\nu}\tau_{\rm corr})$ is about one, and this assurance does not apply to even moderately sized populations of neurons.

These observations are especially relevant for studies that use small time bins to model spike trains driven by natural stimuli. This is because the long correlation times inherent in natural stimuli are passed on to the spike trains, so the assumption of independence across time (which is required for the independence assumption to be valid) breaks badly. Knowing that these models are successful in describing spike trains under the independence assumption, then, does not tell us whether they will be successful in describing full, temporally correlated, spike trains. Note that for studies that use stimuli with short correlation times (e.g., non-natural stimuli such as white noise), the temporal correlations in the spike trains are likely to be short, and using small time bins may be perfectly valid.

The only study that has investigated the issue of temporal correlations in maximum entropy models does indeed support the above picture [9]: for the parameters used in that study ($N\bar{\nu}\delta t = 0.06$ to 0.4), the pairwise maximum entropy model provided a good fit to the data (Δ_N was typically smaller than 0.1), but it did not do a good job modeling the temporal structure of the spike trains.

Other systems—Protein folding

As mentioned in the Introduction, in addition to the studies on neuronal data, studies on protein folding have also emphasized the role of pairwise interactions [2,3]. Briefly, proteins consist of strings of amino acids, and a major question in structural biology is: what is the probability distribution of amino acid strings in naturally folding proteins? One way to answer this is to approximate the full probability distribution of naturally folding proteins from knowledge of single-site and pairwise distributions. One can show that there is a perturbative regime for proteins as well. This can be readily seen using the celebrated HP protein model [20], where a protein is composed of only two types of amino acids. If, at each site, one amino acid type is preferred and occurs with high probability, say $1-\delta$ with $\delta \ll 1$, then a protein of length shorter than $1/\delta$ will be in the perturbative regime, and, therefore, a good match between the true distribution and the pairwise distribution for such a protein is virtually guaranteed.

Fortunately, the properties of real proteins generally prevent this from happening: at the majority of sites in a protein, the distribution of amino acids is *not* sharply peaked around one amino acid. Even for those sites that are sharply peaked (the evolutionarily-conserved sites), the probability of the most likely amino acid, $1-\delta$, rarely exceeds 90% [21,22]. This puts proteins consisting of only a few amino acids out of the perturbative regime, and puts longer proteins—the ones usually studied using pairwise models—well out of it.

This difference is fundamental: because many of the studies that have been carried out on neural data were in the perturbative regime, the conclusions of those studies—specifically, the conclusion that pairwise models provide accurate descriptions of large populations of neurons—is not yet supported. This is not the case for the protein studies, because they are not in the perturbative regime. Thus, the evidence that pairwise models provide accurate descriptions of protein folding remain strong and exceedingly promising.

Open questions

In our analysis, we sidestepped two issues of practical importance: finite sampling and alternative measures for assessing the quality of the pairwise model. These issues are beyond the scope of this paper, but in our view, they are natural next steps in the analysis of pairwise models. Below we briefly expand on them.

Finite sampling refers to the fact that in any real experiment, one has access to only a finite amount of data, and so does not know the true probability distribution of the spike trains. In our analysis, however, we assumed that one did have full knowledge of the true probability distribution. Since a good estimate of the probability distribution is crucial for assessing whether the pairwise model can be extrapolated to large populations, it is important to study how such estimates are affected by finite data. Future work is needed to address this issue, and to find ways to overcome data limitation—for example, by finding efficient methods for removing the finite data bias that affects information theoretic quantities such as the Kullback-Leibler divergence.

There are always many possible ways to assess the quality of a model. Our choice of Δ_N was motivated by two considerations: it is based on the Kullback-Leibler divergence, which is a standard measure of "distance" between probability distributions, and it is a widely used measure in the field [7–10]. It suffers, however, from a number of shortcomings. In particular, Δ_N can be small even when the pairwise model assigns very different probabilities to many of the configurations of the system. It would, therefore, be important to study the quality of pairwise models using other measures.

Methods

The behavior of the true entropy in the large N limit

To understand how the true entropy behaves in the large N limit, it is useful to express the difference of the entropies as a mutual information. Using S_N to denote the true entropy of N neurons and $I(1;N)$ to denote the mutual information between one neuron and the other N neurons in a population of size $N+1$, we have

$$(S_N + S_1) - S_{N+1} = I(1;N) \Rightarrow S_{N+1} - S_N = S_1 - I(1;N). \quad (26)$$

If knowing the activity of N neurons does not fully constrain the firing of neuron $N+1$, then the single neuron entropy, S_1, will exceed the mutual information, $I(1;N)$, and the entropy will be an increasing function of N. For the entropy to be linear in N, all we need is that the mutual information saturates with N. Because of synaptic noise, this is a reasonable assumption for networks of neurons: even if we observed all the spikes from all the neurons, there would still be residual noise associated with synaptic failures, jitter in release time, variability in the amount of neurotransmitter released, stochastic channel dynamics, etc. Consequently, in the large N limit, we may replace $I(1;N)$ by its average, denoted $\langle I(1;\infty)\rangle$. Also replacing S_1 by its average, denoted $\langle S_1\rangle$, we see that for large N, the difference between S_{N+1} and S_N approaches a constant. Specifically,

$$S_N = N[\langle S_1\rangle - \langle I(1;\infty)\rangle] + \text{corrections}, \quad (27)$$

where this expression is valid in the large N limit and the corrections are sublinear in N.

Perturbative expansion

Our main quantitative result, given in Eqs. (8–10), is that the KL divergence between the true distribution and both the independent and pairwise distributions can be computed perturbatively as an expansion in powers of $N\delta$ in the limit $N\delta \ll 1$. Here we carry out this expansion, and derive explicit expressions for the quantities g_{ind} and g_{pair}.

To simplify our notation, here we use $p(\mathbf{r})$ for the true distribution. The critical step in computing the KL divergences perturbatively is to use the Sarmanov-Lancaster expansion [23–28] for $p(\mathbf{r})$,

$$p(\mathbf{r}) = p_{ind}(\mathbf{r})\big(1 + \xi_p(\mathbf{r})\big) \quad (28)$$

where

$$p_{ind}(\mathbf{r}) = \frac{\exp\sum_i \mathcal{H}_i^p r_i}{\Pi_i\left[1+\exp\left(\mathcal{H}_i^p r_i\right)\right]} \quad (29a)$$

$$\xi_p(\mathbf{r}) \equiv \sum_{i<j} \mathcal{J}_{ij}^p \delta r_i \delta r_j + \sum_{i<j<k} \mathcal{K}_{ijk}^p \delta r_i \delta r_j \delta r_k + \ldots \quad (29b)$$

$$\delta r_i = r_i - \bar{r}_i \quad (29c)$$

$$\bar{r}_i \equiv \big(1+\exp\big(-\mathcal{H}_i^p\big)\big)^{-1}. \quad (29d)$$

This expansion has a number of important, but not immediately obvious, properties. First, as can be shown by a direct calculation,

$$\langle r_i\rangle_p = \langle r_i\rangle_{ind} = \bar{r}_i \quad (30)$$

where the subscripts p and *ind* indicate an average with respect to $p(\mathbf{r})$ and $p_{ind}(\mathbf{r})$, respectively. This has an immediate corollary,

$$\langle \delta r_i\rangle_{ind} = 0.$$

This last relationship is important, because it tells us that if a product of $\delta r'$s contains any terms linear in one of the δr_i, the whole product averages to zero under the independent distribution. This can be used to show that

$$\langle \xi_p(\mathbf{r})\rangle_{ind} = 0 \quad (31)$$

from which it follows that

$$\sum_{\mathbf{r}} p(\mathbf{r}) = \langle \big(1+\xi_p(\mathbf{r})\big)\rangle_{ind} = 1.$$

Thus, $p(\mathbf{r})$ is properly normalized. Finally, a slightly more involved calculations provides us with a relationship between the parameters of the model and the moments: for $i \neq j \neq k$,

$$\langle \delta r_i \delta r_j\rangle_p = \bar{r}_i(1-\bar{r}_i)\bar{r}_j\big(1-\bar{r}_j\big)\mathcal{J}_{ij}^p \quad (32a)$$

$$\langle \delta r_i \delta r_j \delta r_k\rangle_p = \bar{r}_i(1-\bar{r}_i)\bar{r}_j\big(1-\bar{r}_j\big)\bar{r}_k(1-\bar{r}_k)\mathcal{K}_{ijk}^p. \quad (32b)$$

Virtually identical expressions hold for higher order moments. It is this last set of relationships that make the Sarmanov-Lancaster expansion so useful.

Note that Eqs. (32a) and (32b), along with the expression for the normalized correlation coefficients given in Eq. (16), imply that

$$(1-\bar{r}_i)\big(1-\bar{r}_j\big)\mathcal{J}_{ij}^p = \rho_{ij}^p \quad (33a)$$

$$(1-\bar{r}_i)\big(1-\bar{r}_j\big)(1-\bar{r}_k)\mathcal{K}_{ijk}^p = \rho_{ijk}^p. \quad (33b)$$

These identities will be extremely useful for simplifying expressions later on.

Because the moments are so closely related to the parameters of the distribution, moment matching is especially convenient: to construct a distribution whose moments match those of $p(\mathbf{r})$ up to some order, one simply needs to ensure that the parameters of that distribution, $\mathcal{H}_i$, $\mathcal{J}_{ij}$, $\mathcal{K}_{ijk}$, etc., are identical to those of the true distributions up to the order of interest. In particular, let us write down a new distribution, $q(\mathbf{r})$,

$$q(\mathbf{r}) = p_{ind}(\mathbf{r})\big(1+\xi_q(\mathbf{r})\big) \quad (34a)$$

$$\xi_q(\mathbf{r}) = \sum_{i<j} \mathcal{J}_{ij}^q \delta r_i \delta r_j + \sum_{i<j<k} \mathcal{K}_{ijk}^q \delta r_i \delta r_j \delta r_k + \ldots. \quad (34b)$$

We can recover the independent distribution by letting $\xi_q(\mathbf{r}) = 0$, and we can recover the pairwise distribution by letting $\mathcal{J}_{ij}^q = \mathcal{J}_{ij}^p$. This allows us to compute $D_{KL}(p\|q)$ in the general case, and then either set ξ_q to zero or set $\mathcal{J}_{ij}^q$ to $\mathcal{J}_{ij}^p$.

Expressions analogous to those in Eqs. (31–33) exist for averages with respect to $q(\mathbf{r})$; the only difference is that p is replaced by q.

The KL divergence in the Sarmanov-Lancaster representation

Using Eqs. (28) and (34a) and a small amount of algebra, the KL divergence between $p(\mathbf{r})$ and $q(\mathbf{r})$ may be written

$$D_{KL}(p\|q)=\frac{1}{\ln 2}\langle f\left(\xi_p(\mathbf{r}),\xi_q(\mathbf{r})\right)\rangle_{ind} \tag{35}$$

where

$$f(x,y)\equiv(1+x)[\ln(1+x)-\ln(1+y)]-(x-y). \tag{36}$$

To derive Eq. (35), we used the fact that $\langle\xi_p\rangle_{ind}=\langle\xi_q\rangle_{ind}=0$ (see Eq. (31)). The extra term $(x-y)$ was included to ensure that $f(x,y)$ and its first derivatives vanish at $x=y$, something that greatly simplifies our analysis later on.

Our approach is to Taylor expand the right hand side of Eq. (35) around $\xi_p=\xi_q=0$, compute each term, and then sum the *whole* series (we do not assume that either ξ_p or ξ_q is small). Using a_{nm} to denote the coefficients of the Taylor series, we have

$$D_{KL}(p\|q)=\frac{1}{\ln 2}\sum_{mn}a_{mn}\langle\xi_p(\mathbf{r})^m\xi_q(\mathbf{r})^n\rangle_{ind}. \tag{37}$$

Note that because $f(x,y)$ and its first derivatives vanish at $x=y=0$, all terms in this sum have $m+n\geq 2$.

Because both ξ_p and ξ_q are themselves sums, an exact calculation of the terms in Eq. (37) would be difficult. However, as we show below, in the section "Averages of powers of ξ_p and ξ_q" (see in particular Eqs. (52) and (54)), they can be computed as perturbation expansions in powers of $N\delta$, and the result is

$$\begin{aligned}&\langle\xi_p(\mathbf{r})^m\xi_q(\mathbf{r})^n\rangle_{ind}=\\&\frac{1}{\ln 2}\sum_{i<j}\overline{r}_i\overline{r}_j\left(\rho_{ij}^p\right)^m\left(\rho_{ij}^q\right)^n+\overline{r}_j\left(-\overline{r}_i\rho_{ij}^p\right)^m\left(-\overline{r}_i\rho_{ij}^q\right)^n\\&+\overline{r}_i\left(-\overline{r}_j\rho_{ij}^p\right)^m\left(-\overline{r}_j\rho_{ij}^q\right)^n\\&+\frac{1}{\ln 2}\sum_{i<j<k}\overline{r}_i\overline{r}_j\overline{r}_k\left(\tilde{\rho}_{ijk}^p\right)^m\left(\tilde{\rho}_{ijk}^q\right)^n+\mathcal{O}\left((N\delta)^4\right)\end{aligned} \tag{38}$$

where $\tilde{\rho}_{ijk}^p$ and $\tilde{\rho}_{ijk}^q$ are given by

$$\tilde{\rho}_{ijk}^x\equiv\rho_{ijk}^x+\rho_{ij}^x+\rho_{ik}^x+\rho_{jk}^x=\frac{\langle r_ir_jr_k\rangle_x-\overline{r}_i\overline{r}_j\overline{r}_k}{\overline{r}_i\overline{r}_j\overline{r}_k}, \tag{39}$$

$x=p,q$. The last equality in Eq. (39) follows from a small amount of algebra and the definition of the correlation coefficients given in Eq. (16). Equation (38) is valid only when $m+n\geq 2$, which is the case of interest to us (since the Taylor expansion of $f(x,y)$ has only terms with $m+n\geq 2$).

The important point about Eq. (38) is that the m and n dependence follows that of the original Taylor expansion. Thus, when we insert this equation back into Eq. (37), we recover our original function—all we have to do is interchange the sums. For example, consider inserting the first term in Eq. (38) into Eq. (37),

$$\begin{aligned}\sum_{m,n}a_{mn}\sum_{i<j}\overline{r}_i\overline{r}_j\left(\rho_{ij}^p\right)^m\left(\rho_{ij}^q\right)^n&=\sum_{i<j}\overline{r}_i\overline{r}_j\sum_{m,n}a_{mn}\left(\rho_{ij}^p\right)^m\left(\rho_{ij}^q\right)^n\\&=\sum_{i<j}\overline{r}_i\overline{r}_jf\left(\rho_{ij}^p,\rho_{ij}^q\right).\end{aligned}$$

Performing the same set of manipulations on all of Eq. (38) leads to

$$\begin{aligned}D_{KL}(p\|q)=&\frac{1}{\ln 2}\sum_{i<j}\overline{r}_i\overline{r}_jf\left(\rho_{ij}^p,\rho_{ij}^q\right)+\overline{r}_jf\left(-\overline{r}_i\rho_{ij}^p,-\overline{r}_i\rho_{ij}^q\right)\\&+\overline{r}_if\left(-\overline{r}_j\rho_{ij}^p,-\overline{r}_j\rho_{ij}^q\right)\\&+\frac{1}{\ln 2}\sum_{i<j<k}\overline{r}_i\overline{r}_j\overline{r}_kf\left(\tilde{\rho}_{ijk}^p,\tilde{\rho}_{ijk}^q\right)+\mathcal{O}\left((N\delta)^4\right).\end{aligned} \tag{40}$$

This expression is true in general (except for some technical considerations; see the section "Averages of powers of ξ_p and ξ_q"); to restrict it to the KL divergences of interest we set $p(\mathbf{r})$ to $p_{true}(\mathbf{r})$ and $q(\mathbf{r})$ to either $p_{ind}(\mathbf{r})$ or $p_{pair}(\mathbf{r})$. In the first case ($q(\mathbf{r})$ set to $p_{ind}(\mathbf{r})$), $\xi_q(\mathbf{r})=0$, which implies that $\mathcal{J}_{ij}^q=0$, and thus $\rho_{ij}^q=0$. Because $f(x,y)$ has a quadratic minimum at $x=y=0$, when $\rho_{ij}^q=0$, the second two terms on the right hand side of Eq. (40) are $\mathcal{O}\left(N^2\delta^3\right)$. We thus have, to lowest nonvanishing order in $N\delta$,

$$D_{KL}(p_{true}\|p_{ind})=\frac{1}{\ln 2}\sum_{i<j}\overline{r}_i\overline{r}_jf\left(\rho_{ij}^p,0\right)+\mathcal{O}\left((N\delta)^3\right), \tag{41}$$

with the $\mathcal{O}\left((N\delta)^3\right)$ correction coming from the last sum in Eq. (40). Defining

$$g_{ind}\equiv\frac{1}{N(N-1)\ln(2)}\sum_{i<j}\frac{\overline{r}_i}{\delta}\frac{\overline{r}_j}{\delta}f\left(\rho_{ij}^p,0\right), \tag{42}$$

where, recall $\delta=\overline{\nu}\delta t$, and inserting Eq. (42) into Eq. (41), we recover Eq. (8a).

In the second case ($q(\mathbf{r})$ set to $p_{pair}(\mathbf{r})$), the first and second moments of $p_{pair}(\mathbf{r})$ and $p_{true}(\mathbf{r})$ are equal. This implies, using Eq. (32), that $\mathcal{J}_{ij}^q=\mathcal{J}_{ij}^p$, and thus $\rho_{ij}^p=\rho_{ij}^q$. Because $f(x,x)=0$ (see Eq. (36)), the first three terms on the right hand side of Eq. (40)—those involving i and j but not k—vanish. The next order term does not vanish, and yields

$$D_{KL}\left(p_{true}\|p_{pair}\right)=\frac{1}{\ln 2}\sum_{i<j<k}\overline{r}_i\overline{r}_j\overline{r}_kf\left(\tilde{\rho}_{ijk}^p,\tilde{\rho}_{ijk}^q\right)+\mathcal{O}\left((N\delta)^4\right). \tag{43}$$

Defining

$$g_{pair}\equiv\frac{1}{N(N-1)(N-2)\ln(2)}\sum_{i<j<k}\frac{r_i}{\delta}\frac{r_j}{\delta}\frac{r_k}{\delta}f\left(\tilde{\rho}_{ijk}^p,\tilde{\rho}_{ijk}^q\right), \tag{44}$$

and inserting this expression into Eq. (43), we recover Eq. (8b).

External fields, pairwise couplings and moments

In this section we derive, to leading order in $N\delta$, expressions relating the external fields and pairwise couplings of the maximum

entropy model, h_i and J_{ij}, to the first and second moments; these are the expressions reported in Eq. (18). The calculation proceeds along the same lines as in the previous section. There is, though, one extra step associated with the fact that the quadratic term in the maximum entropy distribution given in Eq. (14) is proportional to $r_i r_j$, not $\delta r_i \delta r_j$. However, to lowest order in $N\delta$, this doesn't matter. That's because

$$\sum_{i<j} J_{ij} r_i r_j = \sum_{i<j} J_{ij} \delta r_i \delta r_j + r_i \sum_{j \neq i} J_{ij} \overline{r}_j + \text{constants}.$$

where $\overline{r}_i$ is defined as in Eq. (29d) except with $\mathcal{H}_i^p$ replaced by h_i, and we used the fact that $J_{ij} = J_{ji}$. The second term introduces a correction to the external fields, h_i. However, the correction is $\mathcal{O}(N\delta)$, so we drop it. We should keep in mind, though, that our final expression for h_i will have corrections of this order.

Using Eq. (14), but with r_i replaced by δr_i where it appears with J_{ij}, we may write (after a small amount of algebra)

$$p_{maxent}(\mathbf{r}) = p_{ind}(\mathbf{r}) \frac{1 + \xi_2(\mathbf{r}) + \psi(\xi_2(\mathbf{r}))}{1 + \langle \xi_2(\mathbf{r}) + \psi(\xi_2(\mathbf{r})) \rangle_{ind}} \tag{45}$$

where $p_{ind}(\mathbf{r})$ is the same as the function $p_{ind}(\mathbf{r})$ defined in Eq. (29a) except that $\mathcal{H}_i^p$ is replaced by h_i, the subscript "*ind*" indicates, as usual, an average with respect to $p_{ind}(\mathbf{r})$, and the two functions $\xi_2(\mathbf{r})$ and $\psi(x)$ are defined by

$$\xi_2(\mathbf{r}) \equiv \sum_{i<j} J_{ij} \delta r_i \delta r_j \tag{46}$$

and

$$\psi(x) \equiv e^x - 1 - x. \tag{47}$$

Given this setup, we can use Eqs. (55) and (56) below to compute the moments under the maximum entropy model. That's because both $\psi(x)$ and its first derivative vanish at $x = 0$, which are the two conditions required for these equations to be valid. Using also the fact that $\langle \delta r_i \rangle_{ind} = 0$, Eqs. (55) and (56) imply that

$$\langle \xi_2(\mathbf{r}) + \psi(\xi_2(\mathbf{r})) \rangle_{ind} = \sum_{i<j} \overline{r}_i \overline{r}_j \psi(J_{ij}) + \mathcal{O}\left((N\delta)^3\right) \tag{48a}$$

$$\langle r_i \rangle_{maxent} = (1 + \exp(-h_i))^{-1} + \mathcal{O}(N\delta^2) \tag{48b}$$

$$\langle \delta r_i \delta r_j \rangle_{maxent} = \overline{r}_i \overline{r}_j \left[\psi(J_{ij}) + J_{ij}\right] + \mathcal{O}(N\delta^3) \tag{48c}$$

where the first term in Eq. (48b) came from Eq. (29d) with $\mathcal{H}_i^p$ replaced by h_i, the term "$+J_{ij}$" in Eq. (48c) came from $\langle \delta r_i \delta r_j \xi_2(\mathbf{r}) \rangle_{ind}$, and for the second two expressions we used the fact that, to lowest order in $N\delta$, the denominator in Eq. (45) is equal to 1.

Finally, using Eq. (19) for the normalized correlation coefficient, dropping the subscript "*maxent*" (since the first and second moments are the same under the maxent and true distributions), and inverting Eqs. (48b) and (48c) to express the external fields and coupling coefficients in terms of the first and second moments, we arrive at Eq (18).

Averages of powers of ξ_p and ξ_q

Here we compute $\langle \xi_p^m \xi_q^n \rangle_{ind}$, which, as can be seen in Eq. (37), is the key quantity in our perturbation expansion. Our starting point is to (formally) expand the sums that make up ξ_p and ξ_q (see Eqs. (29b) and (34b)), which yields

$$\begin{aligned} &\langle \xi_p(\mathbf{r})^m \xi_q(\mathbf{r})^n \rangle_{ind} \\ &= \sum_{l=2}^{\infty} \sum_{\{m_1,\ldots,m_l\}} \psi^{(l)}_{m_1,\ldots,m_l} \sum_{i_1<\ldots<i_l} \langle \delta r_{i_1}^{m_1} \ldots \delta r_{i_l}^{m_l} \rangle_{ind}. \end{aligned} \tag{49}$$

The sum over $\{m_1,\ldots,m_l\}$ is a sum over all possible configurations of the m_i. The coefficient $\psi^{(l)}_{m_1,\ldots,m_l}$ are complicated functions of the $\mathcal{J}^p_{ij}, \mathcal{J}^q_{ij}, \mathcal{K}^p ijk, \mathcal{K}^q_{ijk}$, etc. Computing these functions is straightforward, although somewhat tedious, especially when l is large; below we compute them only for $l = 2$ and 3. The reason l starts at 2 is that $m + n \geq 2$; see Eq. (37).

We first show that all terms with superscript (l) are $\mathcal{O}(\delta^l)$. To do this, we note that, because the right hand side of Eq. (49) is an average with respect to the independent distribution, the average of the product is the product of the averages,

$$\langle \delta r_{i_1}^{m_1} \delta r_{i_2}^{m_2} \ldots, \delta r_{i_l}^{m_l} \rangle_{ind} = \langle \delta r_{i_1}^{m_1} \rangle_{ind} \langle \delta r_{i_2}^{m_2} \rangle_{ind} \ldots, \langle \delta r_{i_l}^{m_l} \rangle_{ind}. \tag{50}$$

Then, using the fact that $\delta r_i = (1 - \overline{r}_i)$ with probability $\overline{r}_i$ and $-\overline{r}_i$ with probability $(1 - \overline{r}_i)$ (see Eq. (29c)), we have

$$\begin{aligned} \langle \delta r_i^m \rangle_{ind} &= \overline{r}_i (1 - \overline{r}_i)^m + (1 - \overline{r}_i)(-\overline{r}_i)^m \\ &= \overline{r}_i (1 - \overline{r}_i)^m \left[1 - \left(\frac{-\overline{r}_i}{1 - \overline{r}_i}\right)^{m-1}\right]. \end{aligned} \tag{51}$$

The significance of this expression is that, for $m > 1$, $\langle \delta r_i^m \rangle_{ind} \sim \mathcal{O}(r_i) \sim \mathcal{O}(\delta)$, independent of m. Consequently, if all the m_i in Eq. (50) are greater than 1, then the right hand side is $\mathcal{O}(\delta^l)$. This shows that, as promised above, the superscript (l) labels the order of the terms.

As we saw in the section "The KL divergence in the Sarmanov-Lancaster representation", we need to go to third order in δ, which means we need to compute the terms on the right hand side of Eq. (49) with $l = 2$ and 3. Let us start with $l = 2$, which picks out only those terms with two unique indices. Examining the expressions for ξ_p and ξ_q given in Eqs. (29b) and (34b), we see that we must keep only terms involving $\mathcal{J}_{ij}$, since $\mathcal{K}_{ijk}$ has three indices, and higher order terms have more. Thus, the $l = 2$ contribution to the average in Eq. (49), which we denote $\langle \xi_p(\mathbf{r}) \xi_q(\mathbf{r}) \rangle^{(2)}_{ind}$, is given by

$$\langle \xi_p(\mathbf{r})^m \xi_q(\mathbf{r})^n \rangle^{(2)}_{ind} = \sum_{i<j} \langle \left(\mathcal{J}^p_{ij} \delta r_i \delta r_j\right)^m \left(\mathcal{J}^q_{ij} \delta r_i \delta r_j\right)^n \rangle_{ind}.$$

Pulling $\mathcal{J}^p_{ij}$ and $\mathcal{J}^q_{ij}$ out of the averages, using Eq. (33a) to eliminate $\mathcal{J}^p_{ij}$ and $\mathcal{J}^q_{ij}$ in favor of ρ^p_{ij} and ρ^q_{ij}, and applying Eq. (51) (while throwing away some of the terms in that equation that are higher than second order in δ), the above expression may be written

$$\langle \xi_p(\mathbf{r})^m \xi_q(\mathbf{r})^n \rangle^{(2)}_{ind} = \sum_{i<j} \bar{r}_i \bar{r}_j \left(\rho^p_{ij}\right)^m \left(\rho^q_{ij}\right)^n \left[1-(-\bar{r}_i)^{m+n-1}-(-\bar{r}_j)^{m+n-1}\right]. \quad (52)$$

Note that we were not quite consistent in our ordering with respect to δ, in the sense that we kept some higher order terms and not others. We did this so that we could use ρ_{ij} rather than $\mathcal{J}_{ij}$, as the former is directly observable.

For $l=3$ the calculation is more involved, but not substantially so. Including terms with exactly three unique indices in the sum on the right hand side of Eq. (49) gives us

$$\langle \xi_p(\mathbf{r})^m \xi_q(\mathbf{r})^n \rangle^{(3)}_{ind} = \sum_{i<j<k} \langle \left(\mathcal{K}^p_{ijk}\delta r_i\delta r_j\delta r_k + \mathcal{J}^p_{ij}\delta r_i\delta r_j + \mathcal{J}^p_{ik}\delta r_i\delta r_k + \mathcal{J}^p_{jk}\delta r_j\delta r_k\right)^m \left(\mathcal{K}^q_{ijk}\delta r_i\delta r_j\delta r_k + \mathcal{J}^q_{ij}\delta r_i\delta r_j + \mathcal{J}^q_{ik}\delta r_i\delta r_k + \mathcal{J}^q_{jk}\delta r_j\delta r_k\right)^n \rangle_{ind}. \quad (53)$$

This expression is not quite correct, since some of the terms contain only two unique indices—these are the terms proportional to $\left(\mathcal{J}^p_{ij}\right)^m\left(\mathcal{J}^p_{ij}\right)^n$—whereas it should contain only terms with exactly three unique indices. Fortunately, this turns out not to matter, for reasons we discuss at the end of the section.

To perform the averages in Eq. (53), we would need to use multinomial expansions, and then average over the resulting powers of $\delta r'$s. For the latter, we can work to lowest order in δr_i, which means we only take the first term in Eq. (51). This amounts to replacing every δr_i with $1-\bar{r}_i$ (and similarly for j and k), and in addition multiplying the whole expression by an overall factor of $\bar{r}_i\bar{r}_j\bar{r}_k$. For example, if $m=1$ and $n=2$, one of the terms in the multinomial expansion is $\mathcal{K}^p_{ijk}\mathcal{J}^q_{ij}\mathcal{J}^q_{ik}\langle \delta r_i^3\delta r_j^2\delta r_k^2\rangle_{ind}$. This average would yield, using Eq. (51) and considering only the lowest order term, $\bar{r}_i\bar{r}_j\bar{r}_k(1-\bar{r}_i)^3(1-\bar{r}_j)^2(1-\bar{r}_k)^2$.

This procedure also is not quite correct, since terms with only one factor of δr_i, which average to zero, are replaced with $1-\bar{r}_i$. This also turns out not to matter; again, we discuss why at the end of the section.

We can, then, go ahead and use the above "replace blindly" algorithm. Note that the factors of $1-\bar{r}_i$, $1-\bar{r}_j$ and $1-\bar{r}_k$ turn $\mathcal{J}_{ij}$ and $\mathcal{K}_{ijk}$ into normalized correlation coefficients (see Eq. (33)), which considerably simplifies our equations. Using also Eq. (39) for $\tilde{\rho}_{ijk}$, Eq. (53) becomes

$$\langle \xi_p(\mathbf{r})^m \xi_q(\mathbf{r})^n \rangle^{(3)}_{ind} = \sum_{i<j<k} \bar{r}_i\bar{r}_j\bar{r}_k \left(\tilde{\rho}^p_{ijk}\right)^m \left(\tilde{\rho}^q_{ijk}\right)^n. \quad (54)$$

We can now combine Eqs. (52) and (54), and insert them into Eq. (49). This gives us the first two terms in the perturbative expansion of $\langle \xi_p(\mathbf{r})^m \xi_q(\mathbf{r})^n \rangle_{ind}$; the result is written down in Eq. (38) above.

Why can we ignore the overcounting associated with terms in which an index appears exactly zero or one times? We clearly can't do this in general, because for such terms, replacing δr_i with $1-\bar{r}_i$ fails—either because the terms didn't exist in the first place (when one of the indices never appeared) or because they averaged to zero (when an index appeared exactly once). In our case, however, such terms do not appear in the Taylor expansion. To see why, note first of all that, because of the form of $f(x,y)$, its Taylor expansion can be written $(x-y)^2\tilde{f}(x,y)$ where $\tilde{f}(x,y)$ is finite at $x=y$ (see Eq. (36)). Consequently, the original Taylor expansion of $D_{KL}(p\|q)$, Eq. (37), should contain a factor of $\left(\xi_p-\xi_q\right)^2$; i.e.,

$$D_{KL}(p\|q) = \frac{1}{\ln 2}\sum_{m,n} c_{mn}\langle \xi_p(\mathbf{r})^m \xi_q(\mathbf{r})^n \left(\xi_p(\mathbf{r})-\xi_q(\mathbf{r})\right)^2\rangle$$

where the c_{mn} are the coefficients of the Taylor expansion of $\tilde{f}(\xi_p,\xi_q)$. The factor $\left(\xi_p(\mathbf{r})-\xi_q(\mathbf{r})\right)^2$, when expanded, has the form

$$\left(\left(\mathcal{K}^p_{ijk}-\mathcal{K}^q_{ijk}\right)\delta r_i\delta r_j\delta r_k + \left(\mathcal{J}^p_{ij}-\mathcal{J}^q_{ij}\right)\delta r_i\delta r_j + \left(\mathcal{J}^p_{ik}-\mathcal{J}^q_{ik}\right)\delta r_i\delta r_k + \left(\mathcal{J}^p_{jk}-\mathcal{J}^q_{jk}\right)\delta r_j\delta r_k\right)^2.$$

As we saw in the previous section, we are interested in the third order term only to compute $D_{KL}\left(p_{true}\|p_{pair}\right)$, for which $\mathcal{J}^p_{ij}=\mathcal{J}^q_{ij}$. Therefore, the above multiplicative factor reduces to $\left(\mathcal{K}^p_{ijk}-\mathcal{K}^q_{ijk}\right)^2\left(\delta r_i\delta r_j\delta r_k\right)^2$. It is that last factor of $\left(\delta r_i\delta r_j\delta r_k\right)^2$ that is important, since it guarantees that for every term in the Taylor expansion, all indices appear at least twice. Therefore, although Eq. (53) is not true in general, it is valid for our analysis.

We end this section by pointing out that there is a very simple procedure for computing averages to second order in δ. Consider a function $\phi(\xi_p,\xi_q)$ that has a minimum at $\xi_p=\xi_q=0$ and also obeys $\phi(0,0)=0$. Then, based on the above analysis, we have

$$\langle \phi(\xi_p,\xi_q)\rangle_{ind} = \sum_{i<j}\bar{r}_i\bar{r}_j\phi\left(\mathcal{J}^p_{ij},\mathcal{J}^q_{ij}\right)+\mathcal{O}\left((N\delta)^3\right). \quad (55)$$

Two easy corollaries of this are: for k and l positive integers,

$$\langle \delta r_i^k \phi(\xi_p,\xi_q)\rangle_{ind} = \sum_{j\neq i}\bar{r}_i\bar{r}_j\phi\left(\mathcal{J}^p_{ij},\mathcal{J}^q_{ij}\right)+\mathcal{O}\left(N^2\delta^3\right) \quad (56a)$$

$$\langle \delta r_i^k \delta r_j^l \phi(\xi_p,\xi_q)\rangle_{ind} = \bar{r}_i\bar{r}_j\phi\left(\mathcal{J}^p_{ij},\mathcal{J}^q_{ij}\right)+\mathcal{O}\left(N\delta^3\right) \quad (56b)$$

where the sum in Eq. (56a) runs over j only, and we used the fact that both $\mathcal{J}^p_{ij}$ and $\mathcal{J}^q_{ij}$ are symmetric with respect to the interchange of i and j.

Generating synthetic data

As can be seen in Eq. (13), the synthetic data depends on three sets of parameters: h^{true}_i,J^{true}_{ij}, and K^{true}_{ijk}. Here we describe how they were generated.

To generate the h^{true}_i, we draw a set of firing rates, $r^*_1,r^*_2,\ldots,r^*_{N^*}$, from an exponential distribution with mean 0.02 (recall that N^*, which we set to 15, is the number of neurons in our base distribution). From this we chose the external field according to Eq. (18a),

$$h_i^{true} = -\log\left(\frac{1}{r_i^*} - 1\right).$$

In the perturbative regime, a distribution generated with these values of the external fields has firing rates approximately equal to the r_i^*; see Eq. (18a) and Fig. 6.

To generate the J_{ij}^{true} and K_{ijk}^{true}, we drew them from Gaussian distributions with means equal to 0.05 and 0.02 and standard deviations of 0.8 and 0.5, respectively. Using non-zero values for $\mathcal{K}_{ijk}$ means that the true distribution is not pairwise.

Bin size and the correlation coefficients

One of our main claims is that Δ_N is linear in bin size, δt. This is true, however, only if g_{ind} and g_{pair} are independent of δt, as can be seen from Eq. (10b). In this section we show that independence is satisfied if δt is smaller than the typical correlation time of the responses. For δt larger than such correlation times, g_{ind} and g_{pair} do depend on δt, and Δ_N is no longer linear in δt. Note, though, that the correlation time is always finite, so there will always be a bin size below which the linear relationship, $\Delta_N \sim \delta t$, is guaranteed.

Examining Eqs. (42) and (44), we see that g_{ind} and g_{pair} depend on the normalized correlation coefficients, ρ_{ij} and $\tilde{\rho}_{ijk}$ (we drop superscripts, since our discussion will be generic). Thus, to understand how g_{ind} and g_{pair} depend on bin size, we need to understand how the normalized correlation coefficients depend on bin size. To do that, we express them in terms of standard cross-correlograms, as the cross-correlograms contain, in a very natural way, information about the temporal timescales in the spike train.

We start with the second order correlation coefficient, since it is simplest. The corresponding cross-correlogram, which we denote $C_{ij}(\tau)$, is given by

$$C_{ij}(\tau) = \frac{1}{\nu_i \nu_j} \lim_{T\to\infty} \frac{1}{T} \sum_{kl} \delta\left(t_i^k - t_j^l - \tau\right) \tag{57}$$

where t_i^k is the time of the k^{th} spike on neuron i (and similarly for t_j^l), and $\delta(\cdot)$ is the Dirac δ-function. The normalization in Eq. (57) is slightly non-standard—more typical is to divide by something with units of firing rate (ν_i, ν_j or $(\nu_i \nu_j)^{1/2}$), to give units of spikes/s. The normalization we use is convenient, however, because in the limit of large τ, $C_{ij}(\tau)$ approaches one.

It is slightly tedious, but otherwise straightforward, to show that when δt is sufficiently small that only one spike can occur in a time bin, ρ_{ij} is related to $C_{ij}(\tau)$ via

$$\rho_{ij} = \frac{1}{\delta t} \int_{-\delta t}^{\delta t} d\tau (1 - |\tau|/\delta t)\left(C_{ij}(\tau) - 1\right). \tag{58}$$

The (unimportant) factor $(1-|\tau|/\delta t)$ comes from the fact that the spikes occur at random locations within a bin.

Equation (58) has a simple interpretation: ρ_{ij} is the average height of the central peak of the cross-correlogram relative to baseline. How strongly ρ_{ij} depends on δt is thus determined by the shape of the cross-correlogram. If it is smooth, then ρ_{ij} approaches a constant as δt becomes small. If, on the other hand, there is a sharp peak at $\tau = 0$, then $\rho_{ij} \sim 1/\bar{\nu}\delta t = 1/\delta$ for small δt, so long as δt is larger than the width of the peak. (The factor of $\bar{\nu}$ included in the scaling is approximate; it is a placeholder for an effective firing rate that depends on the indices i and j. It is, however, sufficiently accurate for our purposes.) A similar relationship exists between the third order correlogram and the correlation coefficient. Thus, $\tilde{\rho}_{ijk}$ is also independent of δt in the small δt limit, whereas if the central peak is sharp it scales as $1/\delta^2$.

The upshot of this analysis is that the shape of the cross-correlogram has a profound effect on the correlation coefficients and, therefore, on Δ_N. What is the shape in real networks? The answer typically depends on the physical distance between cells. If two neurons are close, they are likely to receive common input and thus exhibit a narrow central peak in their cross-correlogram. Just how narrow depends on the area. Early in the sensory pathways, such as retina [29–31] and LGN [32], peaks can be very narrow—on the order of milliseconds. Deeper into cortex, however, peaks tend to broaden, to at least tens of milliseconds [33,34]. If, on the other hand, the neurons are far apart, they are less likely to receive common input. In this case, the correlations come from external stimuli, so the central peak tends to have a characteristic width given by the temporal correlation time of the stimulus, typically 100 s of milliseconds.

Although clearly both kinds of cross-correlograms exist in any single population of neurons, it is convenient to analyze them separately. We have already considered networks in which the cross-correlograms were broad and perfectly flat, so that the correlation coefficients were strictly independent of bin size. We can also consider the opposite extreme: networks in which the the cross-correlograms (both second and higher order) among nearby neurons exhibit sharp peaks while those among distant neurons are uniformly equal to 1. In this regime, the correlation coefficients depend on δt: as discussed above, the second order ones scale as $1/\delta$ and the third as $1/\delta^2$. This means that the arguments of $f(\rho_{ij},0)$ and $f\left(\tilde{\rho}_{ijk}^{true},\tilde{\rho}_{ijk}^{pair}\right)$ are large (see Eqs. (42) and (44)). From the definition of $f(x,y)$ in Eq. (36), in this regime both are approximately linear in their arguments (ignoring log corrections). Consequently, $f(\rho_{ij},0) \sim 1/\delta$ and $f\left(\tilde{\rho}_{ijk}^{true},\tilde{\rho}_{ijk}^{pair}\right) \sim 1/\delta^2$. This implies that g_{ind} and g_{pair} scale as $N\delta$ and $N^2\delta$, respectively, and so $\Delta_N \sim N$, independent of δ. Thus, if the bin size is large compared to the correlation time, Δ_N will be approximately independent of bin size.

Extending the normalized distance measure to the time domain

In this section we derive the expression for γ given in Eq. (25). Our starting point is its definition, Eq. (24). It is convenient to define $\mathbf{R}$ to be a concatenation of the responses in M time bins,

$$\mathbf{R} \equiv \left(\mathbf{r}^1, \mathbf{r}^2, \ldots, \mathbf{r}^M\right) \tag{59}$$

where, as in the section "Is there anything wrong with using small time bins?", the superscript labels time, so $\wp(\mathbf{R})$ is the full, temporally correlated, distribution.

With this definition, we may write the numerator in Eq. (24) as

$$\begin{aligned} &D_{KL}\left(\wp(\mathbf{R}) \| \prod_t p_{pair}(\mathbf{r}^t)\right) = \\ &-S_{true}^M - \sum_t \sum_{\mathbf{r}} p_{true}^t(\mathbf{r}) \log_2 p_{pair}(\mathbf{r}) \end{aligned} \tag{60}$$

where S_{true}^M is the entropy of $\wp(\mathbf{R})$, the last sum follows from a marginalization over all but one element of $\wp(\mathbf{R})$, and $p_{true}^t(\mathbf{r})$ is the true distribution at time $\mathbf{r}$ (unlike in the section "Is there

anything wrong with using small time bins?", here we do not assume that the true distribution is the same in all time bins). Note that $p_{pair}(\mathbf{r})$ is independent of time, since it is computed from a time average of the true distribution. That time average, which we call $p_{true}(\mathbf{r})$, is given in terms of $p^t_{true}(\mathbf{r})$ as

$$p_{true}(\mathbf{r}) = \frac{1}{M}\sum_t p^t_{true}(\mathbf{r}).$$

Inserting this definition into Eq. (60) eliminates the sum over t, and replaces it with $Mp_{true}(\mathbf{r})$. For simplicity we consider the maximum entropy pairwise model. In this case, because $p_{pair}(\mathbf{r})$ is in the exponential family, and the first and second moments are the same under the true and maximum entropy distributions, we can replace $p_{true}(\mathbf{r})$ with $p_{maxent}(\mathbf{r})$. Consequently, Eq. (60) becomes

$$D_{KL}\left(\wp(\mathbf{R})\|\prod_t p_{pair}(\mathbf{r}^t)\right) = MS_{maxent} - S^M_{true}.$$

This gives us the numerator in the expression for γ (Eq. (24)); using Eq. (4) to write $D_{KL}(p_{true}\|p_{ind}) = S_{ind} - S_{true}$, the full expression for γ becomes

$$\gamma = \frac{M(S_{maxent} - S_{true})}{M(S_{ind} - S_{true})} + \frac{MS_{true} - S^M_{true}}{M(S_{ind} - S_{true})}. \quad (61)$$

where we added and subtracted MS_{true} to the numerator.

The first term on the right hand side of Eq. (61) we recognize, from Eq. (6), as Δ_N. To cast the second into a reasonable form, we define S^M_{ind} to be the entropy of the distribution that retains the temporal correlations within each neuron but is independent across neurons. Then, adding and subtracting this quantity to the numerator in Eq. (61), and also adding and subtracting MS_{ind}, we have

$$\gamma = \Delta_N + \frac{\left(S^M_{ind} - S^M_{true}\right) - M(S_{ind} - S_{true}) + \left(MS_{ind} - S^M_{ind}\right)}{M(S_{ind} - S_{true})}. \quad (62)$$

The key observation is that if $M\delta \ll 1$, then

$$S^M_{ind} - S^M_{true} = g_{ind}N(N-1)(M\delta)^2.$$

Comparing this with Eqs. (8a) and (9a), we see that $S^M_{ind} - S^M_{true}$ is a factor of M^2 times larger than $S_{ind} - S_{true}$. We thus have

$$\gamma = \Delta_N + (M-1) + \frac{MS_{ind} - S^M_{ind}}{M(S_{ind} - S_{true})}. \quad (63)$$

Again assuming $M\delta \ll 1$, and defining $h(x) \equiv -x\log_2 x - (1-x)\log_2(1-x)$, the last term in this expression may be written

$$\begin{aligned} MS_{ind} - S^M_{ind} &= M\sum_i h(r_i) - \sum_i h(Mr_i) \\ &\approx M\sum_i r_i \log_2 M = N\delta \times M\log_2 M. \end{aligned} \quad (64)$$

Inserting this into Eq. (63) and using Eqs. (4), (8a) and (9a) yields Eq. (25).

We have assumed here that $M\delta \ll 1$; what happens when $M\delta \sim 1$, or larger? To answer this, we rewrite Eq. (61) as

$$\gamma = \frac{S_{maxent} - S^M_{true}/M}{S_{ind} - S_{true}}. \quad (65)$$

We argue that in general, as M increases, S^M_{true}/M becomes increasingly different from S_{maxent}, since the former was derived under the assumption that the responses at different time bins were independent. Thus, Eq. (25) should be considered a lower bound on γ.

Author Contributions

Analyzed the data: YR PEL. Wrote the paper: YR SN PEL. Conceived and designed the study: YR SN PEL. Performed the experiments/simulations/mathematical derivations: YR PEL.

References

1. Rieke F, Warland D, de Ruyter van Steveninck R, Bialek W (1997) Spikes: exploring the neural code. Cambridge, MA: MIT Press.
2. Russ W, Lowery D, Mishra P, Yaffe M, Ranganathan R (2005) Natural-like function in artificial WW domains. Nature 437: 579–583.
3. Socolich M, Lockless S, Russ W, Lee H, Gardner K, et al. (2005) Evolutionary information for specifying a protein fold. Nature 437: 512–518.
4. Oates J (1987) Food distribution and foraging behavior. In: Smuts B, Cheney D, Seyfarth R, Wrangham R, Struhsaker T, eds. Primate societies. Chicago: University of Chicago Press. pp 197–209.
5. Wrangham R (1987) Evolution of social structure. In: Smuts B, Cheney D, Seyfarth R, Wrangham R, Struhsaker T, eds. Primate societies. Chicago: University of Chicago Press. pp 282–298.
6. Eisenberg J, Muckenhirn N, Rundran R (1972) The relation between ecology a social structure in primates. Science 176: 863–874.
7. Schneidman E, Berry M, Segev R, Bialek W (2006) Weak pairwise correlations imply strongly correlated network states in a neural population. Nature 440: 1007–1012.
8. Shlens J, Field G, Gauthier J, Grivich M, Petrusca D, et al. (2006) The structure of multi-neuron firing patterns in primate retina. J Neurosci 26: 8254–8266.
9. Tang A, Jackson D, Hobbs J, Chen W, Smith J, et al. (2008) A maximum entropy model applied to spatial and temporal correlations from cortical networks in vitro. J Neurosci 28: 505–518.
10. Bethge M, Berens P (2008) Near-maximum entropy models for binary neural representations of natural images. In: Platt J, Koller D, Singer Y, Roweis S, eds. Advances in Neural Information Processing Systems 20. Cambridge, MA: MIT Press. pp 97–104.
11. Yu S, Huang D, Singer W, Nikolic D (2008) A small world of neuronal synchrony. Cereb Cortex 18: 2891–2901.
12. Kullback S, Leibler R (1951) On information and sufficiency. Ann Math Stat 22: 79–86.
13. Friedman N, Mosenzon O, Slonim N, Tishby N (2001) Multivariate information bottleneck. In: Proc. of Uncertainty in Artificial Intelligence (UAI-17). San Mateo, CA: Morgan Kaufmann Publishers. pp 152–161.
14. Slonim N, Friedman N, Tishby N (2006) Multivariate information bottleneck. Neural Comput 18: 1739–1789.
15. Shannon C, Weaver W (1949) The mathematical theory of communication. Urbana, Illinois: University of Illinois Press.
16. Cover T, Thomas J (1991) Elements of information theory. New York, NY: John Wiley & Sons.
17. Sessak V, Monasson R (2009) Small-correlation expansions for the inverse ising problem. J Phys A 42: 055001.
18. Amari S (2009) Measure of correlation orthogonal to changing in firing rate. Neural Comput 21: 960–972.
19. Shlens J, Field G, Gauthier J, Greschner M, Sher A, et al. (2009) Spatial organization of large-scale concerted activity in the primate retina. J Neurosci. In Press.
20. Dill K (1985) Theory for the folding and stability of globular proteins. Biochemistry 24: 1501–1509.

21. Lockless S, Ranganathan R (1999) Evolutionarily conserved pathways of energetic connectivity in protein families. Science 286: 295–299.
22. Vargas-Madrazo E, Lara-Ochoa F, Jiménez-Montaño M (1994) A skewed distribution of amino acids at recognition sites of the hypervariable region of immunoglobulins. J Mol Evol 38: 100–104.
23. Sarmanov O (1962) Maximum correlation coeffcient (nonsymmetric case). In: Selected Translations in Mathematical Statistics and Probability. Amer. Math. Soc. Volume 2. pp 207–210.
24. Sarmanov O (1963) Maximum correlation coefficient (nonsymmetric case). In: Selected Translations in Mathematical Statistics and Probability. Amer. Math. Soc. Volume 4. pp 271–275.
25. Lancaster H (1958) The structure of bivariate distributions. Ann Math Stat 29: 719–736.
26. Lancaster H (1963) Correlation and complete dependence of random variables. Ann Math Stat 34: 1315–1321.
27. Bahadur R (1961) A representation of the joint distribution of responses to n dichotomous items. In: Solomon H, ed. Studies in Item Analysis and Prediction. Stanford University Press. pp 158–168.
28. Johnson D, Goodman I (2008) Inferring the capacity of the vector Poisson channel with a Bernoulli model. Network 19: 13–33.
29. Mastronarde D (1983) Correlated firing of cat retinal ganglion cells. I. spontaneously active inputs to X- and Y-cells. J Neurophysiol 49: 303–324.
30. DeVries S (1999) Correlated firing in rabbit retinal ganglion cell. J Neurophysiol 81: 908–920.
31. Nirenberg S, Carcieri S, Jacobs A, Latham P (2001) Retinal ganglion cells act largely as independent encoders. Nature 411: 698–701.
32. Dan Y, Alonso J, Usrey W, Reid R (1998) Coding of visual information by precisely correlated spikes in the lateral geniculate nucleus. Nat Neurosci 1: 501–507.
33. Ts'o D, Gilbert C, Wiesel T (1986) Relationships between horizontal interactions and functional architecture in cat striate cortex as revealed by cross-correlation analysis. J Neurosci 6: 1160–1170.
34. Nelson J, Salin P, Munk M, Arzi M, Bullier J (1992) Spatial and temporal coherence in cortico-cortical connections: a cross-correlation study in areas 17 and 18 in the cat. Vis Neurosci 9: 21–37.