Fakultät für Elektrotechnik und Informationstechnik

Learning about Neural Computation from Sparse Recordings

Marcel Nonnenmacher

Vollständiger Abdruck der von der

Fakultät für Elektrotechnik und Informationstechnik

der Technischen Universität München zur Erlangung des akademischen Grades eines Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender: Prof. Dr. Bernhard Wolfrum

Prüfer der Dissertation:

1. Prof. Dr. Jakob H. Macke
2. Prof. Julijana Gjorgjieva, Ph.D.

Die Dissertation wurde am 26.6.2019 bei der Technischen Universität München eingereicht und durch die Fakultät für Elektrotechnik und Informationstechnik am 02.03.2020 angenommen.

Acknowledgements

The work presented in this thesis certainly would not have been possible without the many people that supported me over the years. I first and foremost want to thank Prof. Macke for the chance to join his then young research group, for his continued guidance and support, and in particular for his patience with me. He gave me the freedom to pursue my own ideas, and the opportunity to travel to many exciting places and meet excellent scientists to exchange ideas. It was a pleasure to work together with him on our scientific questions.

My thanks also go to Prof. Bethge, Prof. Berens and Christian Behrens for the fruitful collaboration during the my first project during our time in Tübingen, and especially for their improvements on our manuscript on thermodynamic criticality. I particularly wish to thank Srinivas Turaga for his advice and direct contributions to my second publication on fitting models to subsampled data. It was a joy to listen to and learn from his many great ideas for our project. Our results would also not have been possible without the discussions with Lars Buesing, William Bishop and Prof. Yu that Dr. Turaga made possible at the Janelia Research Campus. I have to acknowledge my predecessors Pedro Goncalves, Jan-Matthis Lückmann and Kaan Öcal for laying the strong basis of our project on likelihood-free inference. Kaan Öcal greatly helped me find my way into the project, and especially Dr. Goncalves continued to provide valuable counsel on the project, as well as on managing my life as a PhD student. It was a grand fortune for the last year of my studies that David Greenberg joined our lab. His resourcefulness and determination enabled our joint project on likelihood-free inference to succeed, and I was very glad to be a part of it. It was a great achievement of Prof. Macke to bring together such a great collection of people and keep us together across various cities and institutions. Over all this time I enjoyed a warm and open atmosphere that I believe I would have found in few other labs, and by now I consider all of them family. I am very thankful to my colleagues Giacomo Bassetto and Alexandre Rene for the extensive and often lively discussions we had over lunch and on so many other occasions, and to Poornima Ramesh to help me keep up my excitement for our work through her shining example. My special gratitude goes to my office mate Artur Speiser for always lending me an ear for my often half-baked ideas.

I should not forget to thank all those scientists and students I met at the conferences, workshops and summer schools that I was lucky to be able to attend over the years. I also experienced outstanding support from HR staff in Tübingen, Bonn and Munich. This especially holds for Katrin Prax and Jacqueline Matzkeit who always had our backs.

Lastly, I have to thank my family for their unwavering trust in me though all these years, which at times felt more than I earned, and Daniel, May-Li and Federica for the time together in Bonn.

Abstract

The impressive capabilities of the brain result from the orchestrated activity of many individual cells. For most neuroscientific model systems, the vast numbers of neurons is still far beyond what modern population recording techniques can simultaneously record at single-cell resolution. Scientists trying to understand neural computations thus regularly deal with very incomplete views on neural circuits. Many relevant neurons are missing from the recordings, and the input they provide to the recorded neurons is also unknown. The implications of this subsampling for the analysis of neural data have gained increased attention, but are far from fully understood. We need data analysis methods which are adapted to severe subsampling of neural population recordings.

The results presented in this publication-based thesis span both statistical modeling and the interpretation of data analysis. My first publication demonstrates the effects of subsampling on the outcome of a test for criticality in neural systems. A recent series of studies targeted the scaling properties of neural codes when neural populations grow in size, and found signatures of thermodynamic criticality in the spiking activity. We re-investigated their analyses, and could relate the observation of criticality to commonly studied quantities such as neuronal firing rates and pair-wise correlations. Using numerical simulations and theoretical analysis, we found that the reported signatures of criticality can be explained by subsampling effects alone.

My second publication develops new methods for statistical modeling of neural data that explicitly address subsampling. New recording techniques allow us to study large neural populations by recording one subpopulation at a time. We can assemble a full picture of the neural activity by fitting a model of the full population to the sequence of partial recordings. Exploiting the structure of the resulting observation patterns allows us to scale this approach to systems with millions of variables. We show empirically that our methods can read out population activity spread over thousands of neurons from severely subsampled recordings when combining several of them.

Our results stress the importance of actively incorporating knowledge about the incompleteness of the neural population recordings when trying to make sense of the observed population activity, and show how new recording techniques can be used to greatly alleviate the problem.

Keywords: neural population activity, statistical modeling, missing data, neural code

Contents

1 Publication Record

During the course of this thesis, two peer-reviewed publications were published with me as first author:

1. **Nonnenmacher M**, Behrens C, Berens P, Bethge M, Macke JH. Signatures of criticality arise from random subsampling in simple population models. PLoS Comput Biol. 2017;(13)10:e1005718

2. **Nonnenmacher M**, Turaga SC, Macke JH. Extracting low-dimensional dynamics from multiple large-scale neural population recordings by learning to predict correlations. In: Advances in Neural Information Processing Systems; 2017. p. 5706–5716.

This publication-based dissertation rests on the work presented in these two papers. Following an introduction explaining the general context of my thesis, both publications are included in ths manuscript alongside a short summary.

Both first-author publications directly dealt with the effects of subsampling on our understanding of neural computations. For both studies, the application of mathematical models to data made up a considerable part of the work. In the later stages of my doctoral studies, I then focused more on this aspect of applying the models to data. This resulted in two more peer-reviewed papers published on conferences for machine learning:

3. Lueckmann JM, Gonçalves PJ, Bassetto B, Öcal K, **Nonnenmacher M**, Macke JH. Flexible statistical inference for mechanistic models of neural dynamics. In: Advances in Neural Information Processing Systems; 2017. p. 1289-1299

4. Greenberg DS, **Nonnenmacher M**, Macke JH. Automatic Posterior Transformation for Likelihood-free inference. Proceedings of the 36th International Conference on Machine Learning, in PMLR 97, p. 2404-2414

These latter publications contain applications of our newly developed methods to models from computational neuroscience in a general context, but we think that these methods will also be useful for the particular context of sparse neural recordings. For this reason, I included short segments on this work in introduction and discussion.

2 Introduction

2.1 Modeling in neuroscience

How the brain works has been a marvel for a very long time. Ramon y Cajal's 'neuron doctrine' [29, 158] famously described the brain as being made up from individual neurons, and raised the question how large networks of neurons perform information processing, decision making and action planning. A long series of improvements in scientific recording methodology allowed to study the activity of neurons under better control and in ever greater detail.

This development in recording techniques was accompanied by mathematical modeling from an early time on. The aim of that modeling is to understand the principles of neural processing that underly the recorded data. A mathematical model is a tool that allows us to reason about what is important to explain the observations and what is not. Desgning a model requires us to define model variables, and thus forces us to think about which entities and concepts we believe to play a role. The model also expresses a hypothesis of how those entities interact through one or several mathematical equations involving those variables. The choice of included variables and form of equations can be motivated by insights into the (bio-)physical processes, by flexibility and expressive power, or by mere ease of solving them. We can compare model predictions against experimental measurements. Through predictions that do no match the data, we come to question, improve or discard the model and ultimately design a new one. Predictions that do match the data prompt us to investigate why the model produced these results — which aspects are essential, how would the model react in different situations, and what do we learn about the modelled processes?

Early mathematical models used in neuroscience were desined to help understand the activity of individual neurons, as it was only possible to measure activity in individual neurons. The Hodgkin-Huxley model results from a seminal study of action potential generation [59]. Action potentials are short and stereotypical events of current flow across the charged neural membrane. They are also referred to as 'spikes', and sequences of action potentials as spike trains. Action potentials are generally considered the primary means of communivation within the brain used by neurons to convey information within the neural network. Through voltage-clamp measurements [33] with a micropipette on the squid giant axon, Hodgkin and Huxley established which ion flows across the membrane and controlling gating mechanisms are central to the formation of action potentials and the temporal evolution of voltage changes. Their model comprised a set of equations for a dynamical system describing voltage-gated ion currents across the cell membrane. With its detailed representation of biophysical quantities, the Hodgkin-

Huxley model is an early example for a biophysical model. Biophysical models are designed from insights into physical or chemical processes understood to underlie the observed activity. They can be very useful to study these mechanisms, but are often hard to adjust to a specific dataset. Hodgkin and Huxley set the free model parameters such that the model under the experimental protocol would reproduce the empirically recorded data.

Another classical model for single-cell activity is the perceptron [125]. Unlike the Hodgkin-Huxley model, the perceptron does not model any biophysical processes involved in shaping the form of action potentials, and instead focuses on modeling on how incoming activity from other cells functionally influences the output activity of a neuron [51]. The perceptron emits an action potential in a given short time interval if the weighted sum of incoming activity within that time interval exceeds a pre-set threshold. Thus unlike the Hodgkin-Huxley model, the output here consists of a sequence of binary variables, i.e. a spike train. The parameters of this model are the input weights—one per input neuron—and the threshold level. As a model for a single artificial neurons, the perceptron formed the building block for feed-forward artificial neural networks [126] and the field of deep learning. Its linear summation assumption was later verified to hold to some degree in real neurons [30], but the perceptron proved more important for further innovations in theoretical neuroscience than for applied data analysis. Unlike the Hodgkin-Huxley model, the perceptron is an example for a phenomenologically motivated model. In contrast to biophysical models, phenomenological models are not necessarily derived from first principles or insights from physics or biology, and their variables and parameters often do not directly correspond to real-world objects or processes. Phomenologocial models are instead often grounded in probability theory and designed for expressiveness and/or ease of use. They consequently focus on closely capturing the observed data and often have high predictive power.

Knowledge about specific processes underlying the data generation can be worked into these models through statistical dependencies between model variables [65]—the probability of a spike to occur at a given point in time can for instance be made depedendent on the recent history of spiking activity through a temporal filter [100, 117]. Phenomenological models however are generally hard to interpret in terms of the biophysical mechanisms that drive the observed activity.

The Hodgkin-Huxley equations and perceptron were important examples of early models to understand the activity of individual neurons. With further developments on recording techniques that allow to simultaneously record the activity of large populations of interconnected neurons, also the mathematical modeling of neuroscientific data saw a shift towards high-dimensional dynamics [141]. Contemporary neuroscientific recording techniques produce vast amounts of data, and modeling has become an integral tool to make sense of these large datasets [107].

2.1.1 Neural recordings

The perceptron aims to describe neural activity given as action potentials, whereas the Hodkgin-Huxley equations model the membrane voltage of neuron, which is a graded

signal. Which type of data we deal with in a neural recording depends on the recording technique that was used to obtain the data, but also on the amount of data (pre-) processing.

Action potentials are typically extracted from recordings through well-established pre-processing steps on the raw recorded signals. The least pre-processing is necessary for intracellular recordings, where electrodes record voltage changes across the membrane directly from within the neurons [80, 60, 99]. Extracellular electrical recordings, where electrodes are held near cell bodies, require more pre-processing [1]. Since the electrodes can pick up signals from several close-by cells, identified action potentials have to assigned to the correct cell [78].

In recent years, optical recording methods became important tools to study neural activity [72, 53, 70]. The most widespread form of optical imaging is calcium imaging [153]. Calcium imaging records changes in fluorescence caused by the changes in intracellular calcium concentration during and after an action potential [19]. The calcium signal is slow compared to the voltage changes underlying the action potentials (response times of several dozen milliseconds for the commonly used calcium indicator GCamp6f [9]), and graded rather than binary. Much recent work [157, 118] focuses on extracting binary spike signals from calcium imaging signals, but models often also work directly on the calcium signal (or on the derived $\Delta F/F$ [31], which highlights changes in fluorescence relative to a moving baseline).

The work presented in this thesis considers both binary spike trains and graded neural output signals such as result from optical recording methods. We will later introduce different families of statistical models that are designed for either binary or graded data.

2.1.2 Neural population recordings

The focus of this thesis lies on neural population recordings, i.e. activity data that is recorded simultaneously from a potentially large number of neurons. Techniques for recording neural activity vary in temporal and spatial resolution, from single-cell intracellular patch-clamp recordings [98] to large-scale methods like electroencephalography [101] or functional magnetic resonance tomography [18]. A drawback of the two mentioned large-scale techniques is that they lack the temporal and/or spatial resolution to resolve the temporal activity of individual cells. For the study of network function and principles of neural information processing, we will focus on methods that allow to identify individual action potentials. Several techniques allow to study the activity of neural populations at the required resolution.

Electrodes pick up the signals of several nearby neurons, which led to early recordings of groups of neurons, albeit with very little control over which neurons are recorded from, or if and how they are interconnected [55]. Multi-shank electrodes and multi-electrode arrays extended the covered area and the coverage of neurons within that area [146, 13]. In some areas, multi-electrode arrays can be used to sample a local population of cells densely, i.e. cover close to 100% of cells within a confined area. One such area is the layer of retinal ganglion cells in a patch of retina [88], and we will cover this example in more detail in chapter 2.2.1.

More recently, advances in optical imaging [73, 88] largely extended the number of simultaneously recorded neurons up to several thousand [144, 108]. This enabled to empirically study the activity of much larger neuronal populations than were previously accessible with electrode recordings.

Optical methods such as single- and two-photon imaging [153, 143] currently obtain the largest population recordings at cell-resolution with up to tens of thousands of neurons [144]. Optical methods allow tracking population activity with single-cell resolution for entire nervous systems e.g. for the nematode C. Elegans and larval zebrafish [3, 2]. These two particular cases however heavily rely on specifics of the respective species—the small diameter of the C. Elegans body and the translucency of the larval zebrafish brain, respectively. For most other important experimental animals such as drosophila, mice or primates, whole-brain imaging with optimal methods is not yet available.

An important concept for optical methods is their field of view (FOV), which limits the size of recorded populations. Fields of view for optical techniques such as single- and two-photon imaging are planar. Laser-scanning approaches move the focal point of the scanning laser in planar patterns, typically in a grid of scan lines with a fast and slow axis. With line-scanning speeds being typically fixed, the density of these grid constitutes a trade-off between frame rates and spatial resolution.

Light-sheet microscopy is a technique that simultaneously illuminates a plane of tissue rather than a line [71, 3]. Many brain structures of interest of interest cover a three-dimensional volume and are not fully captured by a single imaging plane, such as the 6-layered neocortex in mammals. To cover volumes, laser-scanning techniques can move the imaging plane through several imaging depths [34]. The cost of this is that the imaging frequency becomes proportional to the inverse of the number of imaging depths—again a direct trade-off between temporal and spatial resolution. The acquisition time for individual planes influences how sharp the resulting images are and thus constitutes another trade-off [47].

Optical methods require an extra step of data processing to identify individual neurons within the high-dimensional image frames. This can be done by eye and manual annotations of the data videos with so-called regions of interest, or in semi- or fully automated [46, 108] fashion with dedicated algorithms. For simplicity, we will ignore this for most of this thesis and refer to the processed data as neural population activity, with one signal dimension per neuron.

2.1.3 Modeling for neural population recordings

The availability of neural population activity datasets shifted the focus of neuroscientific studies from single-cell activity to understanding how neurons jointly represent and process incoming stimuli through their complex interactions at the population level. This led to new analysis methods that aim to better understand these high-dimensional time-varying signals: Recent modeling tools for neural population activity [82, 36, 149] were developed to examine the statistics of large neural populations and search for principles underlying their collective dynamics [32, 27, 92, 48]. The availability of large

datasets has led to new ways of thinking about theoretical neuroscience [111] that was referred to as 'golden age of computational neuroscience' [112].

A major distinction among mathematical models used in neuroscience remains between phenomenological and biophysical models. In the recent past, many relevant models of population data have been phenomenological. An important reason for this is that we have principled methods to apply phenomenological models to data, as we will see in the next section. Another distinction between models is whether they operate on continuous neuron outputs or on binary signals such as action potentials. This distinction is relevant when it comes to applying a model to data. The next two sections will introduce two families for population analysis models that were fundamental for the work presented in this thesis.

2.1.3.1 Maximum entropy models

One basic goal in the study of neural population activity is to characterise the 'vocabulary' of neural populations, and study the prevalence e.g. of synchronous firing events within the population [105, 149] from finite amounts of recorded data. This can be done in a very targeted fashion with maximum entropy models [63], which describe multivariate probability distributions $P(\mathbf{x})$ that capture selected aspects of the data $\mathbf{x}$, but are otherwise as 'unspecific' as possible. Maximum entropy models are an important model class for neural population analysis on spiking data [128, 148, 127, 149]. We will in the following focus on models for multivariate binary data, although maximum entropy models also exist for continuous signals [64].

A maximum entropy models is designed to capture selected aspects of the data. The selected aspects are expressed through a feature function $F(\mathbf{x})$ applied to data vectors $\mathbf{x}$. The model is constraint by the expected values of these features F attaining specific (measured) values μ, i.e. $\mathrm{E}[F(\mathbf{x})] = \sum_{\mathbf{x}} F(\mathbf{x})P(\mathbf{x}) = \mu$. Being 'unspecific' here is quantified by the entropy of the described probability distribution — for this model class the entropy is maximal given the chosen constraints. Typical constraints chosen to describe neural population activity are the firing rates (first statistical moments), and the pairwise correlations (second statistical moments) between neurons [128, 135]. For binary (spiking) data, these constraints lead to a model sometimes simply called pairwise maximum entropy model. In statistical physics, it is also studied as the spin glass [42, 132], and goes back to the well-known Ising model [62]. For continuous data, the corresponding maximum entropy model with these constraints would be the multivariate Normal distribution.

More formally, a binary maximum entropy model for neural population data assigns a probability $P(\mathbf{x})$ to each spike-pattern $\mathbf{x} \in \{0, 1\}^n$, and can be written as

$$P(\mathbf{x}|\theta) = \frac{1}{Z} \exp\left(-\mathcal{E}(\mathbf{x}|\theta)\right), , \tag{2.1}$$

where the energy function $\mathcal{E}(\mathbf{x}|\theta) = \theta^\top F(\mathbf{x})$ constitutes a linear combination between the parameter vector θ of the model and the feature vector $F(\mathbf{x})$. The recorded neural

population activity $\{\mathbf{x}_t\}_{t=1}^T$ forms a length-T discrete-time sequence of n-dimensional binary vectors.

The selected data features that we want the maximum entropy distribution to match become the components of the feature vector:

$$E_\theta[F(\mathbf{x})] = \sum_{\mathbf{x}} F(\mathbf{x})\ P(\mathbf{x}|\theta) = \langle F(\mathbf{x})\rangle, \tag{2.2}$$

where $\langle F(\mathbf{x})\rangle = \frac{1}{T}\sum_{t=1}^T F(x_t)$ is the average value of the feature across the data.

Adding components to the feature vector allows to extend the maximum entropy model to match additional statistics of the data distribution. The 'K-pairwise' maximum entropy model [149] for spike patterns $\mathbf{x}$ is an extension of pairwise maximum entropy models which reproduce the firing rates and pairwise covariances, and has additional terms to capture population spike-counts [149],

$$P(\mathbf{x}|\theta) = \frac{1}{Z(\theta)} \exp\left(h^\top \mathbf{x} + \mathbf{x}^\top J \mathbf{x} + \sum_{k=0}^{n} V_k \delta\left(K(\mathbf{x}) = k\right)\right). \tag{2.3}$$

The model parameters are h, J and V. The vector $h \in \mathbb{R}^n$ and the upper-triangular matrix $J \in \mathbb{R}^{n\times n}$ correspond to the bias terms and interaction terms in a pairwise maximum entropy model [128], respectively. V is a vector whose entries control the distribution over so-called population spike-counts. Population spike-counts $K(\mathbf{x}) = \sum_{i=1}^n x_i$ are given by the total number of spikes across the population within a single time bin. The indicator-term $\delta\left(K = k\right)$ equals to 1 if the population spike-count is k, and is 0 otherwise. The term $\sum_{k=0}^n V_k \delta\left(K = k\right)$ was introduced [149] to ensure that the model precisely captures the population spike-count distribution of the data using n additional free parameters. The partition function $Z(\theta) = \sum_{\mathbf{x}} \exp\left(h^\top \mathbf{x} + \mathbf{x}^\top J \mathbf{x} + \sum_{k=0}^n V_k \delta\left(K(\mathbf{x}) = k\right)\right)$ ensures that the probabilities given by the model sum to 1. The K-pairwise model is a model of instantaneous population activity and doe not model temporal dependencies in the data. Maximum entropy models can also capture dynamics e.g. by including time-lagged moments as features in $F(\mathbf{x})$ [96].

Maximum likelihood estimation Applying high-dimensional maximum entropy models to data is a challenging computational problem [45, 39, 137, 130]. A common method is to use maximum likelihood estimation (MLE) and Markov chain Monte Carlo (MCMC) [25].

Optimizing the parameters $\theta = \{h, J, V\}$ of the K-pairwise model using maximum likelihood means minimizing a loss L [40, 6] over the data $D = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\}$ with respect to the parameters,

$$L(h, J, V) := -\sum_{t=1}^{T} \log P(\mathbf{x}_t|h, J, V)$$
$$. \tag{2.4}$$

The loss can additionally be regularized, i.e. terms are added to eq. 2.4 which control the magnitudes of parameters h, J, that favour sparse coupling matrices J, or that ensures that the variables V_k controlling the spike-count distribution vary smoothly in k. A smoothness prior on V can be helpful for very large spike counts—these are rarely observed in limited recording times, and hence the model has to interpolate between parameters for spike-counts for which the number of observations is small.
Markov chain Monte Carlo is necessary to estimate statistical moments of $P(\mathbf{x}|\theta)$. The (K-)pairwise maximum entropy model requires the moments $\mathrm{E}_\theta[F(\mathbf{x})]$ given current θ to calculate gradients, as seen from the gradient of the l-th paramter component

$$\begin{aligned}
\frac{\delta}{\delta\theta_l}\sum_{t=1}^{T}\log P(\mathbf{x}_t|\theta) &= \frac{\delta}{\delta\theta_l}\sum_{t=1}^{T}\left(\theta^\top F(\mathbf{x}_t) - \log Z(\theta)\right) \\
&= \sum_{t=1}^{T}\frac{\delta}{\delta\theta_l}\theta^\top f(\mathbf{x}_t) - \frac{\delta}{\delta\theta_l}T\log\sum_{\mathbf{x}}\exp\left(\theta^\top F(\mathbf{x})\right) \\
&= \sum_{t=1}^{T}F_l(\mathbf{x}_t) - T\frac{\sum_{\mathbf{x}}\theta_l\exp\left(\theta^\top F(\mathbf{x})\right)}{\sum_{\mathbf{x}}\exp\left(\theta^\top F(\mathbf{x})\right)} \\
&= T\left(\frac{1}{T}\sum_{t=1}^{T}F_l(\mathbf{x}_t) - \mathrm{E}_\theta[F_l(\mathbf{x})]\right). \qquad (2.5)
\end{aligned}$$

For the K-pairwise model, the components $F_l(\mathbf{x})$ of the feature vector include $x^{(i)}$, $x^{(i)}x^{(j)}$ and the indicator functions $\delta\left(\sum_i x^{(i)}\right) = k\big)$. Thus we need to estimate all first- and second-order moments under the model, as well as the distribution of population spike-counts $P(K(\mathbf{x}) = k) = E[\delta\left(\sum_i x^{(i)}\right) = k)]$, for population spike-count $\sum_i x^{(i)} = K(\mathbf{x})$ and $k = 0, \ldots, n$. These expected values in turn require summations over all 2^n possible states $\mathbf{x} \in \{0,1\}^n$. Since this is prohibitively expensive to compute for $n > 20$, MCMC is commonly used to approximate the moments. Previous work [25] established Gibbs sampling as a simple and reliable tool to approximate the relevant expectations. Gibbs sampling here updates the activity of one neuron i at a time by re-sampling its state from the conditional distribution given the state of the other $n-1$ neurons in the population, $P(x^{(i)}|x^{\sim i}, \theta = \{h, J, V\})$. In practice the MCMC sampling is by far the computationally most expensive part of the algorithm, and within MCMC the estimates of the quadratically many pairwise covariances tend to take the longest to converge.

Flat models Due to the quadratic scaling of the number of second-order moments, applying the full K-pairwise maximum entropy model to data becomes cumbersome beyond a few dozen neurons. Thus previous work studied a simplified nested model [150] that does not explicitly capture first- and second-order moments. We refer to it as 'flat' maximum entropy model, as it effectively assumes that all modeled neurons have identical mean firing rates, pairwise correlations and higher-order correlations [7, 84, 163, 12]. Such a model is fully specified by the population spike-count distribution $P(K = k)$, and all spike words with the same spike count are equally probable. As a

result, the probabilities of individual patterns $\mathbf{x}$ can be read off from the spike-count distribution by

$$P(\mathbf{x}) = \binom{n}{k}^{-1} P(K = k) \tag{2.6}$$

whenever $\sum_{i=1}^{n} x_i = k$. This model can be obtained from the K-pairwise model by setting $h_i = 0$ and $J_{ij} = 0$ for all $i, j \in \{1, \ldots, n\}$ and only optimising entries of V. This in particular means that the quadratically many model parameters in matrix J do no longer need to be adjusted to data. One can fix $V_0 = 0$ [150] without loss of generality, resulting in n degrees of freedom for the model.
Flat model allow the explicit construction of a limit $n \to \infty$, which assumes the existence of a spike-count density $f(r)$, $r \in [0, 1]$ describing the population spike-count distribution of an infinitely large population. $f(r)$ denotes the probability density of a fraction of r neurons spiking simultaneously. Finite-size populations of n cells are then obtained as random subsamples out of this infinitely large system.

2.1.3.2 State-space models

The ongoing exchange of information between individual neurons within a network means that neural population activity develops sequentially over time. Studying population dynamics is hence important for the understanding neural population activity. State-space models [66, 21] are a prominent class of models for time-varying signals with temporal dependencies between observations. In recent years, they became an important tool for the study of neural population dynamics [110]. One reason for the increased popularity of these models is their ability to do *dimensionality reduction*, i.e. to explain the high-dimensional neural population activity through a much smaller amount of variables that are easier to interpret.

State-space models assume that the observed activity $\mathbf{x}_t$ can be described as the result of an unobserved state trajectory $\mathbf{z}_t$ that evolves over time and gives rise to the observations through some probabilistic mapping from the so-called state space onto the observed space,

$$\mathbf{x}_t = g(\mathbf{z}_t, \varepsilon_t), \qquad \varepsilon_t \sim p(\varepsilon_t), \tag{2.7}$$

$$\mathbf{z}_{t+1} = f(\mathbf{z}_t, \eta_t), \qquad \eta_t \sim p(\eta). \tag{2.8}$$

We again denote recorded neural population activity by $\{\mathbf{x}_t\}_{t=1}^{T}$, a length-$T$ discrete-time sequence of n-dimensional real-valued vectors [1] The mean of the observed activity is usually assumed to be zero for simplicity, $\mathrm{E}[\mathbf{x}] = 0$, which can be ensured through a simple pre-processing step of subtracting the empirical average activity.

An important degree of freedom of state-space models with continuous latents $\mathbf{x}$ is the dimensionality of the latent space of $\mathbf{z}_t$, which can be chosen much smaller than the dimensionality of the observed data $\mathbf{x}_t$. This aspect of dimensionality reduction

[1] State-space models can just as well be applied on the level of individual pixels or voxels, which are often several of orders more numerous than the neurons. We in fact used these models in voxel space in our own work on light-sheet microscopy data [104].

makes state-space models interesting as ways to summarize complex time-varying systems through the evolution of only a handful of variables [36, 28, 82, 115, 50]. In this view, individual components of the latent vector $\mathbf{z}$ can be thought of as prominent themes of the population dynamics, and were e.g. mapped to functional subpopulations [69]. Since dynamics are confined to the latent variables, a low-dimensional latent space can greatly simplify the description of the population dynamics in cases where they indeed can be described by a handful of variables. For successful dimensionality reduction, models have to exploit structured correlations in neural activity across both neurons and time [32]. Dimensionality reduction has been used to identify low-dimensional state trajectories that are informative about both stimuli and behaviour, and has yielded important insights into neural computations [92, 24, 27, 131, 85, 49, 79].

A simple example for a state-space model with discrete states is the Hidden-Markov Model [14, 122], which has a univariate discrete hidden state $\mathbf{z}_t \in \{1, \ldots, K\}$ and hidden dynamics governed by a discrete $K \times K$ state-transition table. For continuous states, the linear dynamical system (LDS) [66] with both linear state dynamics $f(\mathbf{x}_t, \eta_t) = A\mathbf{x}_t + \eta_t$, $\eta_t \sim \mathcal{N}(0, Q)$ and a linear mapping from states onto observations with additive noise, $g(\mathbf{x}, \epsilon) = C\mathbf{x} + \epsilon$, is a very important example. The latent state trajectory $\{\mathbf{z}_t\}_{t=1}^T$ in this case is a length-T discrete-time sequence of n-dimensional vectors. In signal processing and engineering, the linear dynamical system is also known as the multi-input multi-output (MIMO) linear time-invariant (LTI) mode , and extensively studied in control theory [68].

State-space models assume that the underlying latent state trajectory $\mathbf{z}_t$ modulates the observed variables $\mathbf{x}_t$ through the mapping f, and that observed temporal structure in $\mathbf{x}_t$ is mediated by the latent trajectory and their dynamics defined by f. The Markov assumption on the latent dynamics expressed through eq. 2.8 is very common when working with discretized time, and can be extended to involve more than one time step in the past when determining the next latent state. Many different variants of state-space models and the linear dynamical system were applied to neural population recordings. These however typically require specialised methods of applying the models to data. Gaussian Process Factor Analysis [28] assumes that the latent states change smoothly over time, which is achieved through a Gaussian process prior assumption over the latent states. For spiking data, the linear dynamical system with Poisson-distributed observations is known as Poisson Linear Dynamics System (PLDS) [82]. Other variants also exist which represent dynamics f and mappings g using flexible (recurrent) neural networks [75, 67, 145].

Expectation Maximization and spectral methods Applying state-space models to data via maximum likelihood is challenging due to their latent variables. For MLE, we want to maximize the likelihood of the data $p(\mathbf{x}|\theta)$ with respect to θ. The model eqs. 2.7 and 2.7 however only give us $p(\mathbf{x}, \mathbf{z}|\theta)$, from which we still need to marginalize over all possible latent trajectories $\mathbf{z}$.

One algorithm that is commonly used for state-space models is Expectation Maximization [37]. EM increases the likelihood over several iterations of two alternat-

ing steps. The E- (expectation-) step consists of finding the expected log-likelihood $E_{p(\mathbf{z}|\mathbf{x},\theta)}([\log p(p(\mathbf{x},\mathbf{z}|\theta))]$ under the distribution of latents given the data and current parameter estimates. For the linear dynamical system, the E-step consists of applying the Kalman filter [66] with the current parameter estimates to the recorded data $\mathbf{x}$. The M- (maximization-) step finds the parameters that maximize the current expected log-likelihood. For the linear dynamical system, the optima of the expected log-likelihood can be found in closed form [52]. Expectation Maximization is prone to getting stuck in local optima, and the final results can strongly depend on the initialization. It is thus common for state-space models to initialize EM with a parameter estimate obtained from some computationally cheaper class of algorithms [83].

An alternative class of algorithms used primarily for linear state-state models is subspace identification (SSID) [156, 68]. SSID algorithms comprise several methods based on linear algebra for identifying the parameters of the model from data. A subclass of SSID algorithms [58, 26] is based on matching the moments of the model with those of the empirical data. The idea here is to first express the time-lagged covariances predicted by the model as a function of the model parameters. Due to the linearity of the mapping from latent states to observed neural activity and the Gaussian-distributed additive noise, the (time-lagged) covariance matrix under the model is simply

$$\mathrm{Cov}[\mathbf{x}_t] = C\mathrm{Cov}[\mathbf{z}_t]C^\top + R, \tag{2.9}$$

$$\mathrm{Cov}[\mathbf{x}_{t+s}, \mathbf{x}_t] = C\mathrm{Cov}[\mathbf{z}_{t+s}, \mathbf{z}_t]C^\top, \tag{2.10}$$

for time-lag $s = 1, \ldots$. From eqs. 2.9, 2.10, spectral methods such as singular value decomposition are used to solve for the model parameters after substituting the observed-variable covariances by empirically measured values. Information from several time-lags $s = 1, \ldots, S$ can be combined in a space-time Hankel matrix, which is a $nS \times nS$ block matrix with individual blocks given by the time-lagged covariances $\mathrm{Cov}[\mathbf{x}_{t+s}, \mathbf{x}_t]$. In this case we require a singular value decomposition of the space-time Hankel matrix [58]. An drawback of these methods is the quadratic scaling of pairwise covariances with population size n, which makes this subclass of SSID algorithms inapplicable in very high-dimensional settings, where one might not even be able to compute the full (time-lagged) covariance matrix $\mathrm{Cov}[\mathbf{x}_t] \in \mathbb{R}^{n\times n}$.

2.2 Subsampling effects

Recording errors, errors in storing data and loss of stored data are basic nuisances of scientific measurement. In consequence, many methods and practices were developed to deal with incomplete recordings and missing data.

A large body of research in statistics and engineering deals with missing data for specific models and model fitting algorithms. Specifically in the context of SSID methods for linear dynamical systems, Markovsky [87, 86] derived conditions for the reconstruction of missing data from deterministic univariate linear time-invariant signals, and Liu et al. [81] use a nuclear norm-regularized SSID to reconstruct partially missing data vectors. Balzano et al. [11, 56] presented a scalable dimensionality reduction approach for large fractions of missing data. Also the Expectation Maximization algorithm for training models with latent variables is robust to a limited amount of missing data, as it can treat individual missing data entries as latent variables. Importantly though, most of these approaches assume that data is missing at random.

In neuroscience, however, missing data is commonly *not* missing at random: Access to neural tissue is often very limited, and for many experimental animals, a skull first has to be opened. The brain is heavily vascularized and an important blood vessels may occlude a neuron group of interest. Deeper brain structures are obstructed by more superficial tissue. And even within a clear field of view, a fluorescent dye might not be expressed in every single neuron.

As will be discussed in further detail chapter 2.2.2, another limitation relevant in all but the simplest and smallest experimental animals is the sheer size of the brain and of functional areas within it, which are still too large to be fully recorded at high rates and at single-cell resolution.

These factors mean that in many situations, data will be missing not for some random neurons at isolated points in time, but that we systematically lack data from many neurons either during the entire recording session or over large contiguous chunks of time. We will refer to this form of missing data as spatio-temporal subsampling. That is, these recordings are spatially subsampled because only a subset of neurons is captured, or they are spatio-temporally subsampled because some neurons are recorded only for a subset of timepoints and at limited temporal resolution. In the remainder of this chapter, we will introduce spatial and spatio-temporal subsampling in the context of neural data analysis. In the next section, we will illustrate effects of spatial subsampling on the outcome and interpretation of data analysis on the example of a recent series of studies on neural coding in the early visual system. In the section thereafter, we will introduce causes and possible solutions to spatio-temporal subsampling in large-scale optical imaging.

Since we now deal with full and subsampled populations, we will denote the size of the full population with N, and the size of subpopulations will be denoted as n, with $n \leq N$. Since different recordings may cover a different amount of neurons, we will in principle deal with multiple different subpopulation sizes n. For the sake of simplicity and to avoid cluttering subindices, we will in the following however only regard a single subpopulation at a time that we compare against the full population.

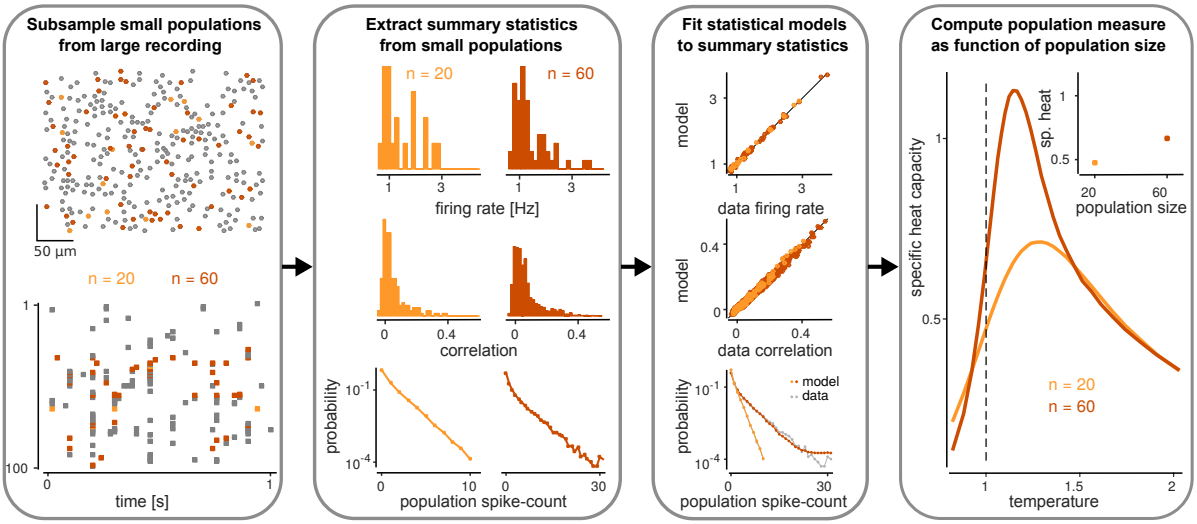

Figure 2.1: The role of subsampling for the study of thermodynamic 'criticality'. Recent publications [151, 96, 61] investigate thermodynamic criticality in neural population activity. For this, they compute a population statistic (specific heat capacity) for subpopulations with varying population size n. Neural subsampling is used here to obtain subpopulations of different size n from a single fixed population recordings (left). Figure reproduced from our paper [103].

2.2.1 Case study: apparent criticality in the early visual system

Which questions about properties of neural population activity can we answer from strongly subsampled population data? Subsampling effects have previously attracted interest for neural coding theories [89]. Not accounting for subsampling effects can profoundly confound the results of data analysis techniques: In particular spatial subsampling can have misleading effects on population-level statistics such as degree of connectivity [138], clustering and spread of activity [77] within subsampled population.

Subsampling effects can also occur where neural population recordings themselves are almost complete. Some statistical analyses require random subsampling of activity from a neural population [139]. With resampling [43, 44], analyses can be averaged over multiple subsamples to gain generality over a single fixed dataset and to increase the statistical robustness of the results.

A recent series of publications [151, 96, 61] used the resampling to study whether neural population recordings of retinal ganglion cells exhibit thermodynamic criticality. Several random subsamples of the same size n are made from amongst all recorded neurons. This procedure furthermore allows to make statements over population 'growth' when repeated for several increasing population sizes n, The scaling behavior of neural network properties is an important aspect of computational modeling: Most neural network models used in theoretical neuroscience were orders of magnitudes smaller than actual biological circuits, which makes it relevant to know how this scaling gap influences the usefulness of model predictions [35].

Critical phenomena have been very helpful in revealing principles underlying the behavior of thermodynamic systems, since the behaviour of a system at a critical point is informative about its intrinsic properties. The recent studies indeed found that the statistics of neural population activity from large-scale multielectrode array recordings [88] resemble those of physical systems at a critical point, more specifically a second-order phase transition. At a phase transition, media qualitatively change their properties by transitioning from one state of matter into another, e.g. a liquid becomes gaseous at the boiling point.

An important signature of a second-order phase transitions the divergence of the specific heat capacity c, a normalized variance of log-probabilities, with population size n [151]. The specific heat capacity is given by

$$c(n) = \frac{1}{n}\mathrm{Var}[\log P(\mathbf{x})], \tag{2.11}$$

where $\mathbf{x} \in \{0,1\}^n$ are activity vectors of a neural population of size n, and $P(\mathbf{x})$ is a probability distribution such as the maximum entropy models introduced in chapter 2.1.3.1. The data used for the analysis however consisted of a dense recording of a single large population of retinal ganglion cells [88]. The size of the recorded population was hence fixed, at about 120 neurons. To investigate a range of different population sizes n, these studies constructed subpopulations by randomly subsampling different amounts of neurons from this full recording. To the data of each subpopulation, the authors fit a K-pairwise maximum entropy model that assigns probabilities to every possible pattern of action potentials. The authors then used MCMC methods to estimate the variance of log-probabilities from the K-pairwise models, and computed specific heat capacities from that. When repeating this process of subsampling and model fitting for increasing population sizes, the authors found that the specific heat capacity seems to diverge as a function of n over the accessible range of $n = 20$ to $n = 120$.

The finding of such thermodynamic criticality could help us better understand how the activity of large neural population is organized [17, 164] and how neural populations in sensory areas such as the retina may process sensory information. Specifically, systems at or close to a thermodynamic critical point were argued to be highly sensitive to external perturbations [151, 96]. To operate at thermodynamic criticality might be hence beneficial for early sensory areas to minimize loss of incoming information. More generally, the occurrence of criticality immediately raises the question of how it came about. Classically, critical phenomena would only be observable in a small area of the space of possible systems. Thus, observing that a system operating at a critical point would be surprising, and we could expect an underlying mechanism that poses the system at this point. Moreover, in biological systems we might expect homeostatic mechanisms that keep the system at this point despite constant micro-level changes [95, 151]. This in turn would give new possible interpretations to known mechanisms of adaptation in the early visual system [93, 76, 134] and alternative mechanisms of self-organization [10]. Several studies [97, 20, 140] also made comparable observations in other biological systems, and it was speculated that criticality might reflect a more general principle of neural circuit organization [95].

Several other studies raised criticism against the claims of thermodynamic criticality in neuroscientific data. A first line of criticism [91] raised that telling whether a system is truly 'close' to an interesting critical point is generally difficult from finite data. Several groups [84, 129, 4, 5] argued that signatures of criticality in high-dimensional datasets are by far not as surprising as previously suggested, and that effects often seen in models applied to neuroscientific data, such as common input, can often account for findings of criticality. Specifically, studies [129, 4, 5] showed that the occurrence of the critical phenomenon of 'Zipf's law' [165, 166] is unsurprising for high-dimensional latent-variable models under a wide range of circumstances. Zip's law and and the divergence of the specific heat are closely related [5]. Empirical studies of salamander retina recordings [155] showed that simple feedforward models of information processing can display Zipf's law without particular tuning or tailored criticality-inducing mechanisms. Even before, simple population model of common input were shown [84] to exhibit signs of diverging specific heat capacity.

A direct connection between signatures of thermodynamic criticality and subsampling had previously not been made (Aitchison et al. [5] discussed subsampling effects on Zipf's law in response to the pre-print version [102] of our work). Part of the reason why subsampling may not been recognized as an important factor of the thermodynamic criticality findings is that it was only introduced at the data analysis stage: The used neural population recordings represent almost the entire RGC population within a small patch of retina [88]. A related field of neural criticality studies is that of neuronal avalanches [57, 16], which studies the occurrence and temporal sequence of bursts in population activity as e.g. captured by the population spike-count. Studies of neuronal avalanche typically rely on multi-electrode recordings, which have sparser coverage than the RGC recordings used for thermodynamic criticality studies. Here, the resulting strong spatial subsampling has been identified as an important potential cause for spurious findings of criticality [120, 161].

For thermodynamic criticality, we suspected that spatial subsampling similarly is a major factor in explaining the finding of diverging specific heat capacity in neural population recordings. Intuitively, the construction of the population-size limit based on random spatial subsampling differs a lot from those studied in statistical physics: In statistical physics, different population sizes typically correspond to systems of different total size, and system properties are thought to scale as a deterministic function, such as spin-glass parameters being drawn from a Gaussian distribution with variance proportional to $1/n$ [133, 94]). As we show in Nonnenmacher et al. (2017a) [103], the key difference is that random spatial subsampling (i.e. without any reference to the spatial location of the neurons) transfers basic statistical properties of the full recorded population onto any large enough subsample—here these were average firing rates and average pairwise correlations. This has important consequences for the behavior of specific heat capacity in the large-population limit of almost any system—the specific heat capacity of almost any realistic system will diverge with population size. The simplicity of the'flat' maximum entropy model was of great help in showing this, since it allows to analytically construct population-size limits of the specific heat capacity.

2.2.2 Spatio-temporal subsampling in neural population recordings

Recording the complete central nervous system is still out of reach for all but the simplest organisms. In consequence, most neural population recordings with single-cell resolution are spatio-temporally subsampled.

Multi-electrode arrays are limited to record individual cell activity only from the direct vicinity of the individual recordings. Thus the maximal spatial extent of the recorded population is limited by the spatial extent of the chip that carries the electrodes, and the sampling density of the neural population activity depends on the spatial density of the electrodes. Sampling rates are typically very high, allowing sub-millisecond resolution, but slow electrode drift can add a form of spatio-temporal subsampling of neural activity: Some neurons are only identified at the beginning of the session, while activity of others only appears after some time into the session [38].

Already for important model systems such as drosophila, contemporary recording techniques cannot provide well-resolved whole-brain activity. For important mammalian models such as rats and mice, new approaches allow to simultaneously record from large parts of the cortical surface [152, 142], albeit at the cost of very low sampling rates. New techniques for fast volumetric imaging such as light-field microscopy [119] are becoming available, but are still limited in their field of view and require extensive algorithmic post-processing of the recorded video data.

In particular if the focus of interest lies on neural *dynamics*, the presence of spatio-temporal subsampling has important consequences for subsequent analysis of the recorded data. Spatio-temporal subsampling leads to problems such as the well-known common input problem [116], which affects a wide range of neural data analyses. The measured neural population activity is thought to come about both from local interactions within the recorded population, and from input from other neurons that were not recorded [106]. Population activity measurements may reflect both local computations and common input.

2.2.3 Filling in missing gaps by combining multiple incomplete recordings

Several neural recording techniques allow to densely sample neural activity within a limited volume of tissue, by recording activity from the majority of neurons within a limited field of view. Sequentially moving this field of view [136, 147] gives rise to multiple spatio-temporally subsampled recordings that together represent a larger subsample of the full population than any individual recording. How can we, from multiple partial recordings, gain insights into dynamics distributed across entire circuits or multiple brain areas? Previous work showed that population activity in principle can be reconstructed from only tens of neurons [160]. While different neurons will be recorded in different recordings, we can in many cases expect the underlying dynamics to be preserved across subpopulations and recordings. This gives us a chance to assemble the big picture of the population dynamics from the pieces found in individual recordings.

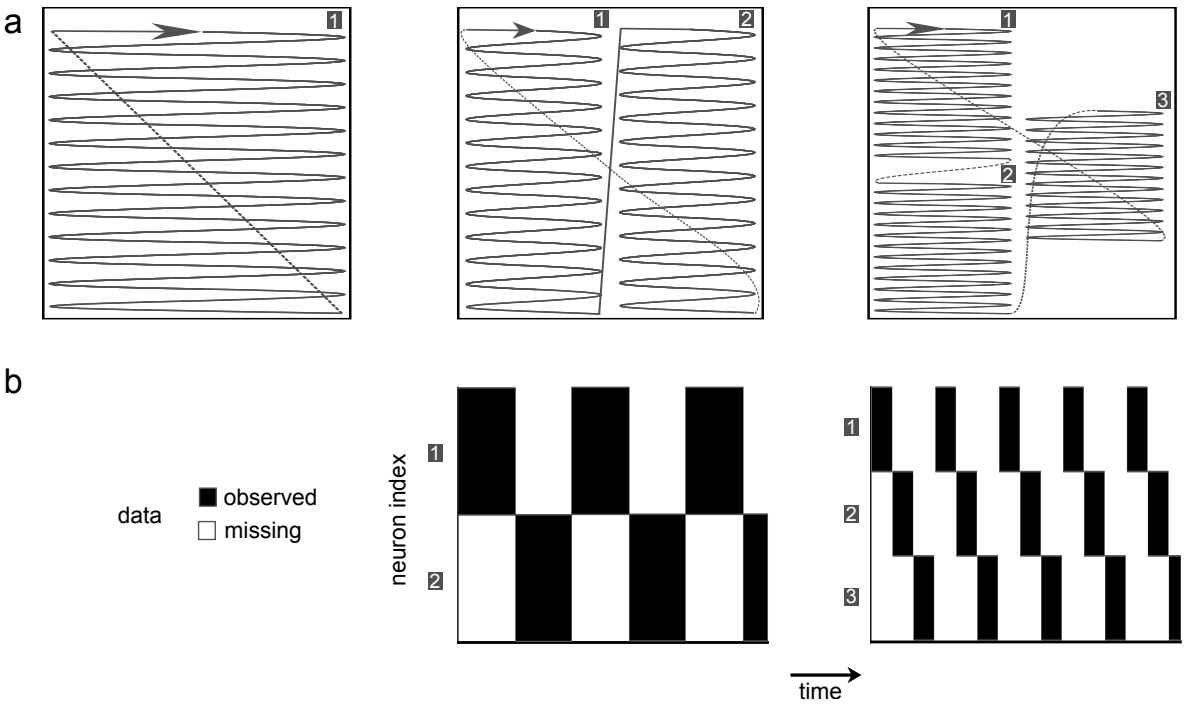

Figure 2.2: Structure of spatio-temporal subsampling in modern optical imaging. Microscopes used e.g. for two-photon imaging record from an area by scanning it. The laser that illuminates the tissue is moved in an adjustable pattern (simplified sketch for illustration). Novel microscopes [136] allow to structure the scanning pattern to cover multiple regions. **a)** The classical scanning pattern is a stack of line scans, with a 'flyback' that restarts the image cycle (left). Some microscopes allow to alternate the scanning between several regions, creating multiple fields of view (centre). Patterns can be complex, and regions can be designed with a high degree of freedom (right). **b)** Different scanning patterns lead to different patterns of missing data. Structured scanning patterns lead to highly structured patterns of missing data, with blocks of missing and observed data.

Combining multiple recordings requires methods that can adapt statistical models to multiple incomplete datasets. In recent years, several lines of work [154, 138] suggested to mitigate the problem of subsampling by combining multiple population recordings into a single global model of population activity. In the following, we will again denote the full population size as N and the size of a subpopulation as n. We here refer with 'full' population to the union of all neurons recorded within any of the individual partial recordings.

Soudry et al. (2015) [138] sought to tackle the common input problem by applying a global model of population activity to a sequence of partial recordings. They approached population subsampling from the perspective of inferring functional connectivity. Starting from a recurrent network model for the full size-N population and using several simplifying assumptions valid for large populations, they arrived at a model with a functional connectivity parameter for each pair of neurons. To be able infer their associated connectivity parameter, their approach requires each pair of neurons within the population to be co-observed at some point. The authors discuss several possible scanning patterns for optical imaging methods to achieve this, but ultimately require multiple independently-moving scanning devices. It should also be noted that there are N^2 many connectivity parameters to be learned from the data, so their approach in practice is limited to small or medium population sizes.

Bishop & Yu (2014) [22] studied the conditions under which a covariance-matrix can be reconstructed from multiple partial measurements. Unlike Soudry et al, they explicitly regarded the case where not all neurons are co-observed over some interval of time. They assumed a dimensionality reduction model without dynamics (as given by eq. 2.7 and ignoring eq. 2.8). Their approach is not based on learning all model parameters jointly. Instead, the parameters governing the covariances between the n neurons in each partial recording are estimated separately. A fundamental problem caused by this is that the parameters identified from different recordings may require a post-hoc alignment, since they may correspond to different latent coordinate systems $\mathbf{z}' = Q\mathbf{z}$ with an unknown change of coordinates between them given by an unknown (invertible) matrix $Q \in \mathbb{R}^{n \times n}$. The authors established rigorous conditions under which the change of coordinate Q can be identified alongside the model parameters, and find that neurons overlapping between different recordings play a crucial role—essentially, every neuron in overlap between two neural subpopulations observed in different recordings reveals another row of Q.

Turaga et al. (2014) [154] use linear state-space models, i.e. incorporate both eqs. 2.7 and 2.8 and hence also dynamics. The main idea behind the use of state-space models is that not all neural population activity has to be observed at any point in time in order to keep track of the latent dynamics. This idea, studied as 'observability' [68] in control theory, states that under certain conditions the parameters of a linear dynamical system can be retrieved from only a subset of the observed variables. The authors avoid the problems of explicitly aligning the model parameters recovered from different partial recordings by fitting a single global dynamical model. They use the term 'stitching' to describe the act of combining multiple partial population recordings. Unlike Bishop & Yu, they do not explicitly investigate the conditions under which the model parameters can be recovered, but also notice that overlap between subpopulations clearly helps to

recover the model parameters. Turaga et al. used Expectation Maximization to learn the parameters of the global high-dimensional linear dynamical system form multiple recordings. Their model assumed one latent variable per neuron, assigned to a specific neuron by the choice of the linear emission matrix $C = I_N$ being fixed to be the identity matrix. Intuitively, each latent variable 'takes over' whenever its assigned neuron is not observed. Because the number of parameters for the latent linear dynamics between N latent variables scales with N^2, this high latent dimensionality in practice limited their approach to small or medium population sizes ($N \approx 100$).

Modern optical imaging methods can simultaneously record thousands of neurons, and we want to integrate multiple such recordings into a single model to identify dynamics shared across multiple subpopulations. As we have seen, several methods to combine multiple partial recordings exist [154, 22, 138]. We have however also seen that they have strong limitations, in particular in their applicability to the immense data dimensionalities that result from combining multiple high-dimensional neural population recordings.

In the work presented in Nonnenmacher et al. (2017b) [104], we tried to remedy these limitations and thus help close the gap between currently available data analysis techniques and the potential of modern neuroscientific datasets.

In particular, we wanted to combine the strengths of the state-space approach from Turaga et al. (2014) with those of the dimensionality-reduction approach of Bishop & Yu (2014). In order to combine temporal dynamics with scaling to large population sizes and good analytical access, we turned to covariance-based subspace identification methods (chapter 2.1.3.2). SSID methods are typically written in terms of linear algebra operations, but can also be expressed as a optimization problems iteratively solvable with gradient descent [81]. Having an explicit loss function allowed us to handle missing data much more easily, and gave us some freedom of choice of how to minimize that loss—using fast stochastic gradient descent greatly improved the scaling of our approach to large datasets.

Another special challenge to applying covariance-based SSID methods is the quadratic scaling of the number of pairwise covariances with the total population size N. Here, again the formulation of model training as stochastic gradient descent provided the solution. By careful reordering of the gradient terms, we could avoid explicit computation of pairwise covariances, and in fact of any terms that scale with N^2 or even just n^2.

One central question to us was how including dynamics can improve the conditions for successful model recovery established by Bishop & Yu. In unpublished work on the idealized case of infinite data (i.e. without estimation errors on the model parameters), we found that correctly identified latent linear dynamics can provide constraints on the unknown latent coordinate systems in a similar way as Bishop & Yu found for overlap between subpopulations. These constraints can reduce the need for overlap, and for rich enough dynamics can replace the overlap requirements almost completely—up to a single bit flip. Our work on these updated stitching conditions is added as an appendix to this thesis.

2.2.4 Model-agnostic parameter learning for spatio-temporal subsampling

In the previous chapter, we sketched how to apply dimensionality-reduction models to large-scale recordings that were spatio-temporally subsampled. Our method to apply these models was heavily tailored towards the linear dynamical system, making use of covariance equations 2.9 and 2.10 and the specific structure of the stochastic gradients resulting from the model loss. On the positive side, using model structure allows to optimize the algorithmic implementations for speed and efficiency. A general downside of model-specific application methods is that even small changes to the model can render them inappropriate.

More recently, work in machine learning focused on 'black-box' fitting methods [123, 8] that can robustly adapt large classes of models to data. This allows flexible data analysis with models that are tailored to specific experimental datasets and support rapid testing of hypotheses. A type of black-box inference that is interesting in the context of subsampled data is likelihood-free or simulation-based inference [15, 54]. Such methods have previously been used to infer the parameters of state-space models from complete data [90].

Likelihood-free methods infer model parameters from data without explicit knowledge of structure of the model or its governing equations. A classical for a model used in neuroscience that does not easily permit a tractable likelihood is the Hodgkin-Huxley model introduced in chapter 2.1.

Everything these likelihood-free methods require from a model is to be able to generate outputs given the model input and a set model parameters θ. Probabilistic models—as previously with eqs. 2.7, 2.8 and 2.1—define a conditional density $p(D|\theta)$ over data $D = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\}$ given model parameters θ.

With a prior distribution over parameters $p(\theta)$, Bayes' theorem allows to get the posterior distribution of parameters given the data as

$$p(\theta|D) \propto p(D, \theta) = p(D|\theta)p(\theta). \tag{2.12}$$

Bayesian inference is typically much harder than maximum likelihood estimation even if the likelihood is analytically accessible. Several approaches were developed over the past decades that are still applicable even if we cannot or do not wish to access the full likelihood. One promising approach [113] uses the the prior and model as a joint distribution $p(D, \theta) = p(D|\theta)p(\theta)$ to generate many pairs (D_i, θ_i) of parameters and 'synthetic' datasets by first sampling a parameter set $\theta_i \sim p(\theta)$ and then generating a dataset from the model $D_i \sim p(D|\theta)$. The core idea of the approach is to then learn the conditional density $p(\theta|D)$ through a regression from datasets D_i into distributions $p(\theta|D_i)$ and then afterwards to 'plug in' the actual recorded dataset for D_i to obtain the posterior. Many real-world datasets will be way too large to learn a distribution $p(\theta|D_i)$ for every possible dataset D_i, in particular for large-scale neural population recordings. It is thus common to compress the datasets by so-called summary features [162] $s(D) \in \mathbb{R}^d$—typically only a handful of hand-selected features—and instead regress from summary statistics $s_i = s(D_i)$ into distributions $p(\theta|s_i)$. To do this, one optimizes

the free parameters ψ of a flexible conditional density estimator $q_\psi(\theta|s)$ via maximum likelihood.

$$q_\psi(\theta|s) \approx p(\theta|s) \propto p(s,\theta) = p(s|\theta)p(\theta) \tag{2.13}$$

The idea to perform likelihood-free inference through regression was first implemented with linear-Gaussian conditional densities $p(\theta|s)$ [15], i.e.

$$q_\psi(\theta|s) = \mathcal{N}(\theta|\mu(s), \Sigma), \tag{2.14}$$

where $\mu(s) = Ms$ was a linear function of the summary statistics and $\psi = \{M, \Sigma\}$. Since then, several lines of work aimed to improve the approach: Multiple studies developed methods for the design of better summary features. The intersection between statistical modeling and deep learning introduced new classes of flexible conditional densities [124, 114, 74].

A major concern for the regression approach lies with how many synthetic data pairs (s_i, θ_i) we need to learn the conditional density $p(\theta|s)$. While the effort of generating a single synthetic dataset $D_i \sim p(D|\theta)$ will generally pale against the effort it took to record the actual neural dataset in a neuroscientific experiment, the generation of hundreds of thousands D_i may still pose a considerable computational burden. To reduce the number of synthetic datasets needed, several studies developed methods to iteratively refine the conditional density estimate $q_\psi(\theta|s)$ [23, 113] and use the current-best posterior estimate to propose model parameters θ_i rather than to sample them from the prior $p(\theta)$. The underlying idea is to early on identify which regions of parameter space are likely to generate the actual dataset and focus the resources of the regression there.

The iterative refinement of the conditional density estimate has shown to be very useful, especially for high-dimensional model parameters and narrow posteriors (as often occur for large datasets and good summary statistics). It however also introduced a new challenge: sampling model parameters from a proposal rather than from the prior leads to synthetic data pairs (s_i, θ_i) that no longer follow $p(s, \theta) \propto p(\theta|s)$ for fixed s. To still learn a conditional density estimator for $p(\theta|s)$ from such synthetic data, we have to correct for the mismatch between prior and proposal. A recent study [113] elegantly solved this correction issue for a flexible class of conditional density estimators, but their solution turned out numerically unstable in practice [41]. To explore the use of likelihood-free inference for large but subsampled neuroscientific datasets, we first had to ensure that the methods are reliably applicable.

We addressed the numerical instability issue in two consecutive publications that each attempted to solve the issue with a different approach. In the first publication Lueckmann et al. (2017), we also demonstrated the use of likelihood-free inference on the Hodgkin-Huxley model. The method established in this first publication however still suffered from a high variance of resulting conditional density estimates and consequently from the need for a large number of synthetic datasets. Recent innovations in the field of normalizing flows [124, 114, 74] furthermore made us want to extend the class of conditional density estimators for which we can correct for parameter proposals and

apply the iterative refinement. In the second publication Greenberg, Nonnenmacher & Macke (2019), we not only resolved the stability issues, but also removed most of the previously existing constraints on the class of the parameter proposals and the class of conditional density estimators.

3 First-author publications

This chapter contains the two first-author publications that form the basis of this publication-based dissertation. A one-page summary for each of the two publications states my personal contributions.

3.1 Signatures of criticality arise from random subsampling in simple population models.

Previous studies found signatures of thermodynamic criticality in neural population activity, but it was unclear how informative about neural computations these findings actually were. We could indeed show the same signatures of criticality also in synthetic data from a simplistic simulation of retinal ganglion cells. We were able to relate the rate of specific heat divergence to basic statistics of the population activity—the stronger the average correlations between the neurons are, the faster the increase of specific heat increase with population size. For the K-pairwise model we verified this empirically with our simulated data, and for the simpler 'flat' model we derived analytic relationships that hold in the limit of large populations. This connection between divergence rates and correlation strength then led us to spatial subsampling as a sufficient condition for the divergence of specific heat capacity: The criticality analysis used random subsampling of neurons to artificially generate neural populations with different sizes from a single recorded population. Simple derivations and numerical experiments showed that random subsampling of a neural population preserves basic statistical quantities such as average firing rate and average correlations. Since the growth rate of specific heat capacity depends on these basic statistical quantities, the specific heat capacity keeps growing at a constant rate for subsampled populations, no matter how large the populations become. Non-random subsampling schemes do not have this issue: subsequently adding neurons to a subpopulation according to their spatial proximity decreases the average correlation strength, since retinal ganglion cells that are further apart tend to have weaker pairwise correlations. The specific heat capacity in this case visibly saturates with increasing population size.

Prof. Macke has extensive experience with maximum entropy modelling, and had studied specific heat capacity before in simpler population models. An important methodological requirement for the simulation study were fast and accurate algorithms for fitting maximum entropy models to several synthetic datasets. For this, I both implemented and state-of-the art algorithms and substantially improved them (which were at least a factor of 3 (e.g. reducing the time needed for MCMC sampling by about a third through Rao-Blackwellization). I did most of the coding, and programmed all of the model fitting

and fit evaluation. Prof. Macke and I contributed equally to the analytical results of the paper. More specifically, I

- led the design and evaluation of the numerical simulation of retinal ganglion cell activity. The implementation of the simulation was written by Christian Behrens under supervision of Prof. Philipp Berens.
- implemented the code for applying the K-pairwise and 'flat' models to data, and the code for evaluating specific heat capacities at different population sizes and population 'temperatures'. The code package resulting from this study is available at `https://github.com/mackelab/CorBinian`.
- provided new analytical results for the 'flat model', e.g. the formula for the asymptotic rate of specific heat divergence as function of correlation strength (eq. 3).

- performed the analytical and numerical studies on the effects of subsampling on average correlations within subsampled populations. The results are summarized in chapter 4 of the supplementary information of the publication.
- performed all data analysis for this project, including all visualizations and figures. In particular, I suggested, performed and analyzed the numerical studies on non-random subsampling strategies to avoid spurious divergence of specific heat capacity.

- wrote the first draft of the paper, which was subsequently revised by all co-authors.

3.2 Extracting low-dimensional dynamics from multiple large-scale neural population recordings.

Modern optical recording techniques allow to sequentially record activity from multiple neural populations, but how to combine these recordings and draw conclusions about the entire system remains an open problem. One promising approach is to apply a global state-space model to all the partial recordings together. However, no existing method scaled to the system sizes that result from combining several large-scale datasets obtained via optical imaging. With concepts from system identification and techniques from machine learning, we developed a method to quickly apply a state-space model at scale to several partial recordings. The state-space model here not only serves to 'stitch' multiple partial recordings together, but also does dimensionality reduction for the high-dimensional data through the identification of low-dimensional latent dynamics. Our algorithm S3ID (Stitching SubSpace IDentification) identifies the parameters of the state-space model by learning to predict those pairwise covariances between neurons that are observed within the recorded data. This allows to identify latent dynamics even in the presence of severe subsampling and small overlap between recordings, as we showed both on simulated data and a whole-brain imaging recording from larval zebrafish.

The main contributions of this paper were of technical nature. S3ID is an iterative algorithm based on stochastic gradient descent, and the main challenge was to efficiently implement the update equations for the estimate of model parameters. To scale well to high-dimensional data, S3ID has to exploit the particular structure of missing data that results from sequentially recording multiple neural populations. S3ID also exploits the linear structure of covariances to keep the computational complexity of stochastic gradient updates linear in population size, which is crucial for the population sizes we aim for. For this project, I was the sole programmer and implemented S3ID. I also performed all numerical simulations and data analysis for this project. More specifically, my contributions to this study included:

- derivation of stochastic gradient updates (equations 7-9 in the publication).
- implementation of the S3ID algorithm, which is publicly available under `https://github.com/mackelab/S3ID`.
- design and implementation of a 'stitching' variant of the Expectation Maximization algorithm for linear state-space models, called 'sEM' (stitching-EM). Unlike the previously published stitching algorithm from Turaga et al. (2013), sEM also supports dimensionality reduction. sEM was used for performance comparisons in figures 2 and 5 of the publication.
- design and execution of the numerical experiments (with advice from Prof. Macke and Dr. Turaga).
- writing the first draft of the paper, which was subsequently revised by Prof. Macke and Dr. Turaga. I made all visualizations and figures of the publication.
- study of necessary and sufficient conditions for successful stitching (section 2.4 in the publication). The work on stitching conditions was greatly helped by discussions with Dr. Turaga and Lars Buesing (then at Columbia University, NY).

4 Discussion

In order to realise the potential of large-scale recordings of neural activity in the search of a theory of neural computation, we need data analysis methods which are adapted to the specific properties of biological data, and in particular the fact that neural population activity is highly subsampled.

To this end, we need to understand the implications of subsampling for data analysis. From the example of the studies on thermodynamic criticality, we see that sometimes we have to take a very close look. In this case the data analysis technique itself introduced spatial subsampling, as a seemingly unproblematic procedure. Whereas several other studies raised criticism against the criticality findings, we note that we were the first to identify the random subsampling as being potentially problematic.

We eventually found the random spatial subsampling to be very problematic indeed, and to actually be sufficient to find diverging specific heat capacity in most studied systems (only a neural population without any correlations would not show diverging specific heat capacity). We want to stress that our findings by themselves do not rule out that critical phenomena play any important role for the organization of neural population activity, be it in the early visual system or elsewhere. We share with Tkacik et al. that finding thermodynamic criticality in neural population activity would be an intriguing scientific result due to the possible insights on the organization of the neural code. Yet a data analysis method that inherently declares any system with non-zero correlations to be in a 'critical' state is hardly useful. We do not exclude that in the future an adapted or entirely new method for the study of thermodynamic criticality becomes a useful tool for the study of neural population activity.

In simple simulated control studies, we did find the spurious signs of specific heat divergence to disappear when using an adapted subsampling strategy that considers the spatial location of neurons: we simply added neurons to the studied (sub-)population by incrementally including neurons according their spatial position along the length of the recorded area. This 'spatial' growth led to specific heat divergences that clearly saturated as the subpopulation size approached the full N. The spatially-informed subsampling strategy is much more in line with system growth as studied in thermodynamics, as it corresponds to an increase of the spatial extent of the neural population, rather than in increase in physical density. So far, we are not aware of the results when applying the 'spatial' subsampling strategy to the RGC population recordings used for the original studies of thermodynamic criticality. Our work on thermodynamic criticality illustrates that subsampling has to be specifically addressed when transferring existing population analyses from other fields to neuroscience.

As in the case of thermodynamic criticality, we generally expect implications of subsampling to be analysis-specific [121]. Thus it is desirable to establish analysis methods

that allow to avoid or at least mitigate subsampling effects. A major source of subsampling effects is given by the limited field of few of electrical and optical recording techniques. A new generation of optical recording techniques allow close-to-complete coverage of neural activity in a incrementally movable field of view [136]. This opens up way to scale up the empirical study of neural dynamics by sequentially recording from multiple neural subpopulations. We here investigated how recordings from these existing techniques can be analysed—*after* the data is acquired—to effectively increase the field of view and hence reduce severeness of subsampling. Our work focused on methods that scale to the large neural populations that are accessible with these extended fields of view, and may eventually open up applying models to neural populations spanning entire neural circuits. In one application, we demonstrated the good scaling properties of our methods to data from whole-brain imaging in fictively behaving larval zebrafish [3] at sub-cellular resolution. The light-sheet microscopy recordings in this case were not subsampled, and we assumed a scenario of spatio-temporally subsampling for illustration purposes. As is shared with other recent large-FOV recording techniques [152, 142], these recordings were however marked by a fairly low frame rate of $1.15Hz$, limiting our understanding of the fast neural dynamics underlying for instance perception and decision making. We believe that our methods can also allow combining the resulting neural activity to yield a high-spatial and temporal resolution portrait of brain-wide population dynamics: as previously discussed (chapter 2.1.2), light-sheet microscopy images volumes by stepping through a sequence of imaging planes. The recorded activity from the different imaging planes are then usually treated as a single image frame recorded all at once. More accurately however, the activity represents a set of sequentially imaged neural subpopulations, with one subpopulation per imaging plane. By treating this data as heavily spatio-temporally subsampled, one can in principle increase the framerate from the time it takes to image a full frame to the time it takes to image a single imaging plane (see figure 4.1a). This 'frame-rate stitching' may allow to quantitatively identify specific neural sub-populations involved in decision-making and both their causal and computational roles by analysing the structure of these fast-scale neural dynamics. Several pitfalls for frame-rate stitching exist, such as the potentially low or even non-existent overlap between imaging planes. A central question of stitching that may also play a role here is whether the neurons spread across all imaging planes truly share the same underlying population dynamics. In many cases, we can be optimistically expect that external stimuli and experimentally induced behavior impose brain-wide structure in the neural activity. Another challenge is that in experimental reality, observation patterns are not always as structured as we so far assumed in our work. One particular example for this is given by long-term electrical recordings as in Dhawale et al. (2015). Which neurons fall in or out of recording in subsequent days depends strongly on the their exact locations and hard-to-control movements of the electrode.

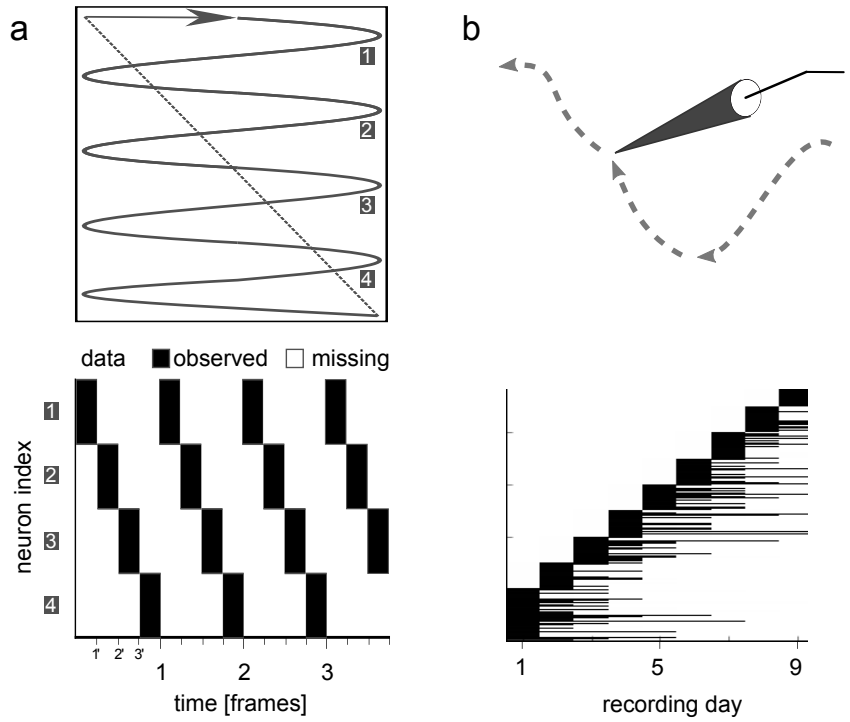

Figure 4.1: Other forms of spatio-temporal subsampling in neural population data. **a)** Scanning techniques take time to scan the field of view for each frame (sketched for two-photon imaging, but this also holds for light-sheet microscopy). A frame is usually treated as an instantaneous measurement, but in reality is better approximated as a sequence of scans over different subregion and their neural populations. Then the framerate is given only by the time it takes to scan a subregion (here about 4x faster). **b)** Electrode-drift in electrical recordings. Over time, the exact position of the electrode moves. **c)** Patterns of observed and missing data for sub-frame resolved recordings. (**d**) For electrode drift, the dependence on the exact location means that observation patterns will be less clearly structured (cf. figure 4C in [38]).

Our work focused on the well-understood linear dynamical system, but our methods are also applicable to more recent studies with powerful nonlinear models [109] [1]. We expect that our work on likelihood-free inference will eventually further broaden the applicability of stitching methods. We can hope that future work improves the scope and fidelity of neural recording techniques [159] to the point where subsampling eventually stops being a concern for the analysis of most neuroscientific experiments.

[1] The state-space model used in Pandarinath et al. (2018) [109] is based on recurrent and feed-forward neural networks, which themselves feature linear-nonlinear function cascades. We believe the initialization of the linear function to be very well possible, and to improve upon the initialization used in their stitching application.

Bibliography

[1] Moshe Abeles and Moise H Goldstein. Multispike train analysis. *Proceedings of the IEEE*, 65(5):762–773, 1977.

[2] Misha B Ahrens and Florian Engert. Large-scale imaging in small brains. *Current opinion in neurobiology*, 32:78–86, 2015.

[3] Misha B Ahrens, Michael B Orger, Drew N Robson, Jennifer M Li, and Philipp J Keller. Whole-brain functional imaging at cellular resolution using light-sheet microscopy. *Nature methods*, 10(5):413–420, 2013.

[4] Laurence Aitchison, Nicola Corradi, and Peter E Latham. Zipf's law arises naturally in structured, high-dimensional data. *arXiv preprint*, 1407.7135v4, 2014.

[5] Laurence Aitchison, Nicola Corradi, and Peter E Latham. Zipf's law arises naturally when there are underlying, unobserved variables. *PLoS Comput Biol*, 12(12):e1005110, Dec 2016.

[6] Yasemin Altun and Alex Smola. Unifying divergence minimization and statistical inference via convex duality. In *Learning theory*, pages 139–153. Springer, 2006.

[7] Shun-ichi Amari, Hiroyuki Nakahara, Si Wu, and Yutaka Sakai. Synchronous firing and higher-order interactions in neuron pool. *Neural Computation*, 15(1):127–142, 2003.

[8] Evan Archer, Il Memming Park, Lars Buesing, John Cunningham, and Liam Paninski. Black box variational inference for state space models. *arXiv:1511.07367*, 2015.

[9] Aleksandra Badura, Xiaonan R Sun, Andrea Giovannucci, Laura A Lynch, and Samuel S H Wang. Fast calcium sensor proteins for monitoring neural activity. *Neurophotonics*, 1(2):025008, 2014.

[10] Per Bak, Chao Tang, and Kurt Wiesenfeld. Self-organized criticality: An explanation of the 1/f noise. *Physical review letters*, 59(4):381, 1987.

[11] Laura Balzano, Robert Nowak, and Benjamin Recht. Online identification and tracking of subspaces from highly incomplete information. In *Communication, Control, and Computing (Allerton), 2010 48th Annual Allerton Conference on*, pages 704–711. IEEE, 2010.

[12] Andrea K Barreiro, Julijana Gjorgjieva, Fred Rieke, and Eric Shea-Brown. When do microcircuits produce beyond-pairwise correlations? *Front Comput Neurosci*, 8:10, 2014.

[13] M Baudry and M Taketani. *Advances in network electrophysiology: Using multi-electrode arrays*. Springer, 2006.

[14] Leonard E Baum and Ted Petrie. Statistical inference for probabilistic functions of finite state markov chains. *The annals of mathematical statistics*, 37(6):1554–1563, 1966.

[15] Mark A Beaumont, Jean-Marie Cornuet, Jean-Michel Marin, and Christian P Robert. Adaptive approximate bayesian computation. *Biometrika*, page asp052, 2009.

[16] John M Beggs and Dietmar Plenz. Neuronal avalanches in neocortical circuits. *The Journal of neuroscience*, 23(35):11167–11177, 2003.

[17] John M Beggs and Nicholas Timme. Being critical of criticality in the brain. *Frontiers in physiology*, 3, 2012.

[18] JW Belliveau, DN Kennedy, RC McKinstry, BR Buchbinder, RMt Weisskoff, MS Cohen, JM Vevea, TJ Brady, and BR Rosen. Functional mapping of the human visual cortex by magnetic resonance imaging. *Science*, 254(5032):716–719, 1991.

[19] Michael J Berridge, Peter Lipp, and Martin D Bootman. The versatility and universality of calcium signalling. *Nature reviews Molecular cell biology*, 1(1):11, 2000.

[20] William Bialek, Andrea Cavagna, Irene Giardina, Thierry Mora, Edmondo Silvestri, Massimiliano Viale, and Aleksandra M Walczak. Statistical mechanics for natural flocks of birds. *Proceedings of the National Academy of Sciences*, 109(13):4786–4791, 2012.

[21] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

[22] William E Bishop and Byron M Yu. Deterministic symmetric positive semidefinite matrix completion. In *Advances in Neural Information Processing Systems*, pages 2762–2770, 2014.

[23] M G B Blum and O François. Non-linear regression models for approximate bayesian computation. *Statistics and Computing*, 20(1), 2010.

[24] K L Briggman, H D I Abarbanel, and W B Kristan, Jr. Optical imaging of neuronal populations during decision-making. *Science*, 307(5711):896–901, 2005.

[25] Tamara Broderick, Miroslav Dudik, Gasper Tkacik, Robert E Schapire, and William Bialek. Faster solutions of the inverse pairwise ising problem. *arXiv*, 0712.2437v2, 2007.

[26] L. Buesing, J. H. Macke, and M. Sahani. Spectral learning of linear dynamics from generalised-linear observations with application to neural population data. In *Advances in Neural Information Processing Systems*, pages 1682–1690, 2012.

[27] D. V. Buonomano and W. Maass. State-dependent computations: spatiotemporal processing in cortical networks. *Nat Rev Neurosci*, 10(2):113–125, 2009.

[28] M Yu Byron, John P Cunningham, Gopal Santhanam, Stephen I Ryu, Krishna V Shenoy, and Maneesh Sahani. Gaussian-process factor analysis for low-dimensional single-trial analysis of neural population activity. In *Advances in neural information processing systems*, pages 1881–1888, 2009.

[29] SR Cajal. Histologia del sistema nervioso. *Madrid: Moya*, 1899.

[30] Sydney Cash and Rafael Yuste. Linear summation of excitatory inputs by ca1 pyramidal neurons. *Neuron*, 22(2):383–394, 1999.

[31] Tsai-Wen Chen, Trevor J Wardill, Yi Sun, Stefan R Pulver, Sabine L Renninger, Amy Baohan, Eric R Schreiter, Rex A Kerr, Michael B Orger, Vivek Jayaraman, et al. Ultrasensitive fluorescent proteins for imaging neuronal activity. *Nature*, 499(7458):295, 2013.

[32] Mark M Churchland, John P Cunningham, Matthew T Kaufman, Justin D Foster, Paul Nuyujukian, Stephen I Ryu, and Krishna V Shenoy. Neural population dynamics during reaching. *Nature*, 487(7405):51, 2012.

[33] Kenneth S Cole and Howard J Curtis. Electric impedance of the squid giant axon during activity. *The Journal of general physiology*, 22(5):649–670, 1939.

[34] Lee Cossell, Maria Florencia Iacaruso, Dylan R Muir, Rachael Houlton, Elie N Sader, Ho Ko, Sonja B Hofer, and Thomas D Mrsic-Flogel. Functional organization of excitatory synaptic strength in primary visual cortex. *Nature*, 518(7539):399–403, 2015.

[35] JR Cotton, AS Ecker, E Froudarakis, Philipp Berens, M Bethge, P Saggau, and AS Tolias. Scaling of information in large sensory neuronal populations. In *45th Annual Meeting of the Society for Neuroscience (Neuroscience 2015)*, 2015.

[36] John P Cunningham and M Yu Byron. Dimensionality reduction for large-scale neural recordings. *Nature neuroscience*, 17(11):1500–1509, 2014.

[37] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

[38] Ashesh K Dhawale, Rajesh Poddar, Steffen BE Wolff, Valentin A Normand, Evi Kopelowitz, and Bence P Ölveczky. Automated long-term recording and analysis of neural activity in behaving animals. *eLife*, 6, 2017.

[39] Miroslav Dudik, Steven J Phillips, and Robert E Schapire. Performance guarantees for regularized maximum entropy density estimation. In *Learning Theory*, pages 472–486. Springer, 2004.

[40] Miroslav Dudík and Robert E Schapire. Maximum entropy distribution estimation with generalized regularization. In *Learning Theory*, pages 123–138. Springer, 2006.

[41] Conor Durkan, George Papamakarios, and Iain Murray. Sequential neural methods for likelihood-free inference. *arXiv preprint arXiv:1811.08723*, 2018.

[42] Samuel Frederick Edwards and Phil W Anderson. Theory of spin glasses. *Journal of Physics F: Metal Physics*, 5(5):965, 1975.

[43] Bradley Efron. *The jackknife, the bootstrap, and other resampling plans*, volume 38. Siam, 1982.

[44] Bradley Efron. Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics*, pages 569–593. Springer, 1992.

[45] Alan M Ferrenberg and Robert H Swendsen. New monte carlo technique for studying phase transitions. *Physical review letters*, 61(23):2635, 1988.

[46] Michael Francis, Xun Qian, Chimène Charbel, Jonathan Ledoux, James C Parker, and Mark S Taylor. Automated region of interest analysis of dynamic ca2+ signals in image sequences. *American Journal of Physiology-Cell Physiology*, 303(3):C236–C243, 2012.

[47] Johannes Friedrich, Weijian Yang, Daniel Soudry, Yu Mu, Misha B Ahrens, Rafael Yuste, Darcy S Peterka, and Liam Paninski. Multi-scale approaches for high-speed imaging and analysis of large neural populations. *PLoS computational biology*, 13(8):e1005685, 2017.

[48] Peiran Gao and Surya Ganguli. On simplicity and complexity in the brave new world of large-scale neuroscience. *Current opinion in neurobiology*, 32:148–155, 2015.

[49] Peiran Gao and Surya Ganguli. On simplicity and complexity in the brave new world of large-scale neuroscience. *Curr Opin Neurobiol*, 32:148–55, 2015.

[50] Yuanjun Gao, Lars Busing, Krishna V Shenoy, and John P Cunningham. High-dimensional neural spike train analysis with generalized count linear dynamical systems. In *Advances in Neural Information Processing Systems*, pages 2044–2052, 2015.

[51] Wulfram Gerstner, Werner M Kistler, Richard Naud, and Liam Paninski. *Neuronal dynamics: From single neurons to networks and models of cognition*. Cambridge University Press, 2014.

[52] Zoubin Ghahramani and Geoffrey E Hinton. Parameter estimation for linear dynamical systems. Technical report, Technical Report CRG-TR-96-2, University of Totronto, Dept. of Computer Science, 1996.

[53] Christine Grienberger and Arthur Konnerth. Imaging calcium in neurons. *Neuron*, 73(5):862–885, 2012.

[54] Michael U Gutmann and Jukka Corander. Bayesian optimization for likelihood-free inference of simulator-based statistical models. *The Journal of Machine Learning Research*, 17(1):4256–4302, 2016.

[55] Kenneth D Harris, Rodrigo Quian Quiroga, Jeremy Freeman, and Spencer L Smith. Improving data quality in neuronal population recordings. *Nature neuroscience*, 19(9):1165, 2016.

[56] Jun He, Laura Balzano, and John Lui. Online robust subspace tracking from partial information. *arXiv preprint arXiv:1109.3827*, 2011.

[57] Andreas VM Herz and John J Hopfield. Earthquake cycles and neural reverberations: collective oscillations in systems with pulse-coupled threshold elements. *Physical review letters*, 75(6):1222, 1995.

[58] BL HO and Rudolph E Kalman. Editorial: Effective construction of linear state-variable models from input/output functions. *at-Automatisierungstechnik*, 14(1-12):545–548, 1966.

[59] Alan L Hodgkin and Andrew F Huxley. A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of physiology*, 117(4):500–544, 1952.

[60] Alan L Hodgkin, Andrew F Huxley, and B Katz. Measurement of current-voltage relations in the membrane of the giant axon of loligo. *The Journal of physiology*, 116(4):424–448, 1952.

[61] Mark L Ioffe and J Berry II, Michael. The structuredlow temperature'phase of the retinal population code. *arXiv preprint arXiv:1608.05751*, 2016.

[62] Ernst Ising. Beitrag zur theorie des ferromagnetismus. *Zeitschrift für Physik A Hadrons and Nuclei*, 31(1):253–258, 1925.

[63] Edwin T Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620, 1957.

[64] Edwin T Jaynes. *Probability theory: The logic of science*. Cambridge university press, 2003.

[65] Michael I Jordan et al. Graphical models. *Statistical Science*, 19(1):140–155, 2004.

[66] Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. *Journal of basic Engineering*, 82(1):35–45, 1960.

[67] Maximilian Karl, Maximilian Soelch, Justin Bayer, and Patrick van der Smagt. Deep variational bayes filters: Unsupervised learning of state space models from raw data. *arXiv preprint arXiv:1605.06432*, 2016.

[68] Tohru Katayama. *Subspace methods for system identification*. Springer Science & Business Media, 2006.

[69] Saul Kato, Harris S Kaplan, Tina Schrödel, Susanne Skora, Theodore H Lindsay, Eviatar Yemini, Shawn Lockery, and Manuel Zimmer. Global brain dynamics embed the motor command sequence of caenorhabditis elegans. *Cell*, 163(3):656–669, 2015.

[70] Philipp J Keller and Misha B Ahrens. Visualizing whole-brain activity and development at the single-cell level using light-sheet microscopy. *Neuron*, 85(3):462–83, Feb 2015.

[71] Philipp J Keller, Annette D Schmidt, Joachim Wittbrodt, and Ernst HK Stelzer. Reconstruction of zebrafish early embryonic development by scanned light sheet microscopy. *science*, 322(5904):1065–1069, 2008.

[72] J. N. D. Kerr and W. Denk. Imaging in vivo: watching the brain in action. *Nat Rev Neurosci*, 9(3):195–205, 2008.

[73] J. N. D. Kerr and W. Denk. Imaging in vivo: watching the brain in action. *Nature Reviews Neurosci*, 9(3):195–205, 2008.

[74] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems*, pages 10236–10245, 2018.

[75] Rahul G Krishnan, Uri Shalit, and David Sontag. Deep kalman filters. *arXiv preprint arXiv:1511.05121*, 2015.

[76] Anna Levina, J Michael Herrmann, and Theo Geisel. Phase transitions towards criticality in a neural system with adaptive interactions. *Physical Review Letters*, 102(11):118110, 2009.

[77] Anna Levina and Viola Priesemann. Subsampling scaling. *Nature Communications*, 8:15140, 2017.

[78] Michael S Lewicki. A review of methods for spike sorting: the detection and classification of neural action potentials. *Network: Computation in Neural Systems*, 9(4):R53–R78, 1998.

[79] Nuo Li, Kayvon Daie, Karel Svoboda, and Shaul Druckmann. Robust neuronal dynamics in premotor cortex during motor planning. *Nature*, 532(7600):459–64, 2016.

[80] G Ling and RW Gerard. The normal membrane potential of frog sartorius fibers. *Journal of cellular and comparative physiology*, 34(3):383–396, 1949.

[81] Zhang Liu, Anders Hansson, and Lieven Vandenberghe. Nuclear norm system identification with missing inputs and outputs. *Systems & Control Letters*, 62(8):605–612, 2013.

[82] J. H. Macke, L. Buesing, J. P. Cunningham, B. M. Yu, K. V. Shenoy, and M. Sahani. Empirical models of spiking in neural populations. In *Advances in Neural Information Processing Systems*, volume 24, 2012.

[83] J. H. Macke, L. Buesing, and M. Sahani. *Advanced State Space Methods for Neural and Clinical Data*, chapter Estimating state and model parameters in state-space models of spike trains. Cambridge University Press, 2015.

[84] Jakob H Macke, Manfred Opper, and Matthias Bethge. Common input explains higher-order correlations and entropy in a simple model of neural population activity. *Physical Review Letters*, 106(20):208102, 2011.

[85] Valerio Mante, David Sussillo, Krishna V Shenoy, and William T Newsome. Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature*, 503(7474):78–84, 2013.

[86] Ivan Markovsky. A missing data approach to data-driven filtering and control. *IEEE Transactions on Automatic Control*, 2016.

[87] Ivan Markovsky. The most powerful unfalsified model for data with missing values. *Systems & Control Letters*, 2016.

[88] Olivier Marre, Dario Amodei, Nikhil Deshmukh, Kolia Sadeghi, Frederick Soo, Timothy E Holy, and Michael J Berry. Mapping a complete neural population in the retina. *The Journal of Neuroscience*, 32(43):14859–14873, 2012.

[89] Matteo Marsili, Iacopo Mastromatteo, and Yasser Roudi. On sampling and modeling complex systems. *Journal of Statistical Mechanics: Theory and Experiment*, 2013(09):P09003, 2013.

[90] GM Martin, BP McCabe, W Maneesoonthorn, and CP Robert. Approximate bayesian computation in state space models. *arXiv preprint*, page arXiv:1409.8363, 2014.

[91] Iacopo Mastromatteo and Matteo Marsili. On the criticality of inferred models. *Journal of Statistical Mechanics: Theory and Experiment*, 2011(10):P10012, 2011.

[92] Ofer Mazor and Gilles Laurent. Transient dynamics versus fixed points in odor representations by locust antennal lobe projection neurons. *Neuron*, 48(4):661–73, 2005.

[93] Christian Meisel and Thilo Gross. Adaptive self-organization in a realistic neural network model. *Physical Review E*, 80(6):061917, 2009.

[94] M Mezard, G Parisi, and MA Virasoro. *Spin Glass Theory and Beyond (Singapore: Word Scientific)*. 1987.

[95] Thierry Mora and William Bialek. Are biological systems poised at criticality? *Journal of Statistical Physics*, 144(2):268–302, 2011.

[96] Thierry Mora, Stéphane Deny, and Olivier Marre. Dynamical criticality in the collective activity of a population of retinal neurons. *Physical review letters*, 114(7):078105, 2015.

[97] Thierry Mora, Aleksandra M Walczak, William Bialek, and Curtis G Callan. Maximum entropy models for antibody diversity. *Proceedings of the National Academy of Sciences*, 107(12):5405–5410, 2010.

[98] Erwin Neher and Bert Sakmann. Single-channel currents recorded from membrane of denervated frog muscle fibres. *Nature*, 260(5554):799, 1976.

[99] Erwin Neher and Bert Sakmann. The patch clamp technique. *Scientific American*, 266(3):44–51, 1992.

[100] John Ashworth Nelder and Robert WM Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3):370–384, 1972.

[101] Ernst Niedermeyer and FH Lopes da Silva. *Electroencephalography: basic principles, clinical applications, and related fields*. Lippincott Williams & Wilkins, 2005.

[102] Marcel Nonnenmacher, Christian Behrens, Philipp Berens, Matthias Bethge, and Jakob H Macke. Signatures of criticality arise in simple neural population models with correlations. *arXiv:1603.00097*, 2016.

[103] Marcel Nonnenmacher, Christian Behrens, Philipp Berens, Matthias Bethge, and Jakob H Macke. Signatures of criticality arise from random subsampling in simple population models. *PLoS computational biology*, 13(10):e1005718, 2017.

[104] Marcel Nonnenmacher, Srinivas C Turaga, and Jakob H Macke. Extracting low-dimensional dynamics from multiple large-scale neural population recordings by learning to predict correlations. In *Advances in Neural Information Processing Systems*, pages 5702–5712, 2017.

[105] Michael Okun, Nicholas A Steinmetz, Lee Cossell, M Florencia Iacaruso, Ho Ko, Péter Barthó, Tirin Moore, Sonja B Hofer, Thomas D Mrsic-Flogel, Matteo Carandini, and Kenneth D Harris. Diverse coupling of neurons to populations in sensory cortex. *Nature*, 521(7553):511–5, May 2015.

[106] Timothy O'leary and Eve Marder. Mapping neural activation onto behavior in an entire animal. *Science*, 344(6182):372–373, 2014.

[107] Timothy O'Leary, Alexander C Sutton, and Eve Marder. Computational models in the age of large datasets. *Current opinion in neurobiology*, 32:87–94, 2015.

[108] Marius Pachitariu, Carsen Stringer, Mario Dipoppa, Sylvia Schröder, L Federico Rossi, Henry Dalgleish, Matteo Carandini, and Kenneth D Harris. Suite2p: beyond 10,000 neurons with standard two-photon microscopy. *Biorxiv*, page 061507, 2017.

[109] Chethan Pandarinath, Daniel J O'Shea, Jasmine Collins, Rafal Jozefowicz, Sergey D Stavisky, Jonathan C Kao, Eric M Trautmann, Matthew T Kaufman, Stephen I Ryu, Leigh R Hochberg, et al. Inferring single-trial neural population dynamics using sequential auto-encoders. *Nature methods*, page 1, 2018.

[110] Liam Paninski, Yashar Ahmadian, Daniel Gil Ferreira, Shinsuke Koyama, Kamiar Rahnama Rad, Michael Vidne, Joshua Vogelstein, and Wei Wu. A new look at state-space models for neural data. *Journal of computational neuroscience*, 29(1-2):107–126, 2010.

[111] Liam Paninski and John Cunningham. Neural data science: accelerating the experiment-analysis-theory cycle in large-scale neuroscience. *bioRxiv*, page 196949, 2017.

[112] Liam Paninski and Emily Singer. Why neuroscience needs data scientists, November 2018.

[113] George Papamakarios and Iain Murray. Fast ε-free inference of simulation models with bayesian conditional density estimation. In *Advances in Neural Information Processing Systems*, pages 1028–1036, 2016.

[114] George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for density estimation. In *Advances in Neural Information Processing Systems*, pages 2338–2347, 2017.

[115] David Pfau, Eftychios A Pnevmatikakis, and Liam Paninski. Robust learning of low-dimensional dynamics from large neural ensembles. In *Advances in neural information processing systems*, pages 2391–2399, 2013.

[116] Jonathan W Pillow and Peter E Latham. Neural characterization in partially observed populations of spiking neurons. In *Advances in Neural Information Processing Systems*, pages 1161–1168, 2008.

[117] Jonathan W Pillow, Jonathon Shlens, Liam Paninski, Alexander Sher, Alan M Litke, E J Chichilnisky, and Eero P Simoncelli. Spatio-temporal correlations and visual signalling in a complete neuronal population. *Nature*, 454(7207):995–9, 2008.

[118] Eftychios A Pnevmatikakis, Daniel Soudry, Yuanjun Gao, Timothy A Machado, Josh Merel, David Pfau, Thomas Reardon, Yu Mu, Clay Lacefield, Weijian Yang, et al. Simultaneous denoising, deconvolution, and demixing of calcium imaging data. *Neuron*, 89(2):285–299, 2016.

[119] Robert Prevedel, Young-Gyu Yoon, Maximilian Hoffmann, Nikita Pak, Gordon Wetzstein, Saul Kato, Tina Schrödel, Ramesh Raskar, Manuel Zimmer, Edward S Boyden, et al. Simultaneous whole-animal 3d imaging of neuronal activity using light-field microscopy. *Nature methods*, 11(7):727, 2014.

[120] Viola Priesemann, Matthias HJ Munk, and Michael Wibral. Subsampling effects in neuronal avalanche distributions recorded in vivo. *BMC neuroscience*, 10(1):40, 2009.

[121] Viola Priesemann, Michael Wibral, and Jochen Triesch. Learning more by sampling less: subsampling effects are model specific. *BMC neuroscience*, 14(1):P414, 2013.

[122] Lawrence R Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[123] Rajesh Ranganath, Sean Gerrish, and David Blei. Black box variational inference. In *Artificial Intelligence and Statistics*, pages 814–822, 2014.

[124] Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. *arXiv preprint arXiv:1505.05770*, 2015.

[125] Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

[126] Frank Rosenblatt. Principles of neurodynamics. perceptrons and the theory of brain mechanisms. Technical report, Cornell Aeronautical Lab Inc Buffalo NY, 1961.

[127] Yasser Roudi, Sheila Nirenberg, and Peter E Latham. Pairwise maximum entropy models for studying large biological systems: when they can work and when they can't. *PLoS Comput Biol*, 5(5):e1000380, 2009.

[128] E. Schneidman, M. J. 2nd Berry, R. Segev, and W. Bialek. Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature*, 440(7087):1007–12, 2006.

[129] David J Schwab, Ilya Nemenman, and Pankaj Mehta. Zipf's law and criticality in multivariate data without fine-tuning. *Physical review letters*, 113(6):068102, 2014.

[130] Greg Schwartz, Jakob Macke, Dario Amodei, Hanlin Tang, and Michael J Berry, 2nd. Low error discrimination using a correlated population code. *J Neurophysiol*, 108(4):1069–88, 2012.

[131] K. V. Shenoy, M. Sahani, and M. M. Churchland. Cortical control of arm movements: a dynamical systems perspective. *Annu Rev Neurosci*, 36:337–59, 2013.

[132] David Sherrington and Scott Kirkpatrick. Solvable model of a spin-glass. *Physical review letters*, 35(26):1792, 1975.

[133] David Sherrington and Scott Kirkpatrick. Solvable model of a spin-glass. *Physical review letters*, 35(26):1792, 1975.

[134] Woodrow L Shew, Wesley P Clawson, Jeff Pobst, Yahya Karimipanah, Nathaniel C Wright, and Ralf Wessel. Adaptation to sensory input tunes visual cortex to criticality. *Nature Physics*, 11(8):659–663, 2015.

[135] Jonathon Shlens, Greg D Field, Jeffrey L Gauthier, Matthew I Grivich, Dumitru Petrusca, Alexander Sher, Alan M Litke, and E J Chichilnisky. The structure of multi-neuron firing patterns in primate retina. *J Neurosci*, 26(32):8254–66, 2006.

[136] Nicholas James Sofroniew, Daniel Flickinger, Jonathon King, and Karel Svoboda. A large field of view two-photon mesoscope with subcellular resolution for in vivo imaging. *eLife*, 5, 2016.

[137] Jascha Sohl-Dickstein, Peter B Battaglino, and Michael R DeWeese. New method for parameter estimation in probabilistic models: minimum probability flow. *Physical review letters*, 107(22):220601, 2011.

[138] Daniel Soudry, Suraj Keshri, Patrick Stinson, Min-hwan Oh, Garud Iyengar, and Liam Paninski. Efficient" shotgun" inference of neural connectivity from highly sub-sampled activity data. *PLoS Comput Biol*, 11(10):e1004464, 2015.

[139] Eran Stark and Moshe Abeles. Applying resampling methods to neurophysiological data. *Journal of neuroscience methods*, 145(1-2):133–144, 2005.

[140] Greg J Stephens, Thierry Mora, Gašper Tkačik, and William Bialek. Statistical thermodynamics of natural images. *Phys Rev Lett*, 110(1):018701, Jan 2013.

[141] Ian H Stevenson and Konrad P Kording. How advances in neural recording affect data analysis. *Nature neuroscience*, 14(2):139, 2011.

[142] Jeffrey N Stirman, Ikuko T Smith, Michael W Kudenov, and Spencer L Smith. Wide field-of-view, multi-region, two-photon imaging of neuronal activity in the mammalian brain. *Nature biotechnology*, 34(8):857, 2016.

[143] Christoph Stosiek, Olga Garaschuk, Knut Holthoff, and Arthur Konnerth. In vivo two-photon calcium imaging of neuronal networks. *Proceedings of the National Academy of Sciences*, 100(12):7319–7324, 2003.

[144] Carsen Stringer, Marius Pachitariu, Nicholas Steinmetz, Charu Bai Reddy, Matteo Carandini, and Kenneth D Harris. Spontaneous behaviors drive multidimensional, brain-wide population activity. *BioRxiv*, page 306019, 2018.

[145] David Sussillo, Rafal Jozefowicz, LF Abbott, and Chethan Pandarinath. Lfads-latent factor analysis via dynamical systems. *arXiv:1608.06315*, 2016.

[146] CA Thomas Jr, PA Springer, GE Loeb, Y Berwald-Netter, and LM Okun. A miniature microelectrode array to monitor the bioelectric activity of cultured cells. *Experimental cell research*, 74(1):61–66, 1972.

[147] CH Tischbirek, T Noda, M Tohmi, A Birkner, I Nelken, and A Konnerth. In vivo functional mapping of a cortical column at single-neuron resolution. *Cell reports*, 27(5):1319–1326, 2019.

[148] G. Tkacik, E. Schneidman, M. J. Berry, II, and W. Bialek. Spin glass models for a network of real neurons. *arXiv:q-bio/0611072v2*, 2009.

[149] Gašper Tkačik, Olivier Marre, Dario Amodei, Elad Schneidman, William Bialek, and Michael J Berry, 2nd. Searching for collective behavior in a large network of sensory neurons. *PLoS Comput Biol*, 10(1):e1003408, 2014.

[150] Gašper Tkačik, Olivier Marre, Thierry Mora, Dario Amodei, Michael J Berry II, and William Bialek. The simplest maximum entropy model for collective behavior in a neural network. *Journal of Statistical Mechanics: Theory and Experiment*, 2013(03):P03011, 2013.

[151] Gašper Tkačik, Thierry Mora, Olivier Marre, Dario Amodei, Stephanie E. Palmer, Michael J. Berry, and William Bialek. Thermodynamics and signatures of criticality in a network of neurons. *Proceedings of the National Academy of Sciences*, 112(37):11508–11513, 2015.

[152] Philbert S Tsai, Celine Mateo, Jeffrey J Field, Chris B Schaffer, Matthew E Anderson, and David Kleinfeld. Ultra-large field-of-view two-photon microscopy. *Optics express*, 23(11):13833–13847, 2015.

[153] Roger Y Tsien. Fluorescence measurement and photochemical manipulation of cytosolic free calcium. *Trends in neurosciences*, 11(10):419–424, 1988.

[154] Srini Turaga, Lars Buesing, Adam M Packer, Henry Dalgleish, Noah Pettit, Michael Hausser, and Jakob Macke. Inferring neural population dynamics from multiple partial recordings of the same neural circuit. In *Advances in Neural Information Processing Systems*, pages 539–547, 2013.

[155] Joanna Tyrcha, Yasser Roudi, Matteo Marsili, and John Hertz. The effect of non-stationarity on models inferred from neural data. *Journal of Statistical Mechanics: Theory and Experiment*, 2013(03):P03005, 2013.

[156] Peter Van Overschee and BL De Moor. *Subspace identification for linear systems: Theory—Implementation—Applications*. Springer Science & Business Media, 2012.

[157] Joshua T Vogelstein, Brendon O Watson, Adam M Packer, Rafael Yuste, Bruno Jedynak, and Liam Paninski. Spike inference from calcium imaging using sequential monte carlo methods. *Biophysical journal*, 97(2):636–655, 2009.

[158] Wilhelm Waldeyer. Ueber einige neuere forschungen im gebiete der anatomie des centralnervensystems1. *DMW-Deutsche Medizinische Wochenschrift*, 17(44):1213–1218, 1891.

[159] Siegfried Weisenburger and Alipasha Vaziri. A guide to emerging technologies for large-scale and whole-brain optical imaging of neuronal activity. *Annual review of neuroscience*, 41:431–452, 2018.

[160] Ryan C Williamson, Benjamin R Cowley, Ashok Litwin-Kumar, Brent Doiron, Adam Kohn, Matthew A Smith, and M Yu Byron. Scaling properties of dimensionality reduction for neural populations and network models. *PLOS Computational Biology*, 12(12):e1005141, 2016.

[161] Jens Wilting and Viola Priesemann. Inferring collective dynamical states from widely unobserved systems. *Nature communications*, 9(1):2325, 2018.

[162] Simon N Wood. Statistical inference for noisy nonlinear ecological dynamic systems. *Nature*, 466(7310):1102, 2010.

[163] S. Yu, H. Yang, H. Nakahara, G. S. Santos, D. Nikolic, and D. Plenz. Higher-order interactions characterized in cortical activity. *J Neurosci*, 31(48):17514–17526, 2011.

[164] Shan Yu, Hongdian Yang, Oren Shriki, and Dietmar Plenz. Universal organization of resting brain activity at the thermodynamic critical point. *Front Syst Neurosci*, 7:42, 2013.

[165] George Kingsley Zipf. *The psycho-biology of language*. Houghton, Mifflin, 1935.

[166] George Kingsley Zipf. Human behavior and the principle of least effort. 1949.

B Conditions for Stitching

A central question to stitching asks under which circumstance we can idenfify the parameters of the state-space model from multiple partial recordings. For concreteness, we assume the following parametrization of a linear dynamical system:

$$\mathbf{x}_t = C\mathbf{z}_t + \varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}(0, R), \tag{B.1}$$

$$\mathbf{z}_{t+1} = A\mathbf{z}_t + \eta_t, \qquad \eta_t \sim \mathcal{N}(0, Q). \tag{B.2}$$

The approach followed by Bishop & Yu (2014) considers the situation of two subpopulations with a given amount of overlap, which are observed one after the other. We will assume that the latent dynamics are observable from either subpopulation, leading to two independent system reconstructions $\theta = \{A, Q, C, R\}$, $\tilde{\theta} = \{\tilde{A}, \tilde{Q}, \tilde{C}, \tilde{R}\}$. For our target application of dimensionality reduction with very high-dimensional observations and comparatively low-dimensional dynamics, this assumption is typically not very problematic in practice. These two sets of model parameters $\theta, \tilde{\theta}$ for the same linear dynamical system are related via an unknown change of coordinate system. For easier comparison with Bishop & Yu (2014) and the notation in my own publication on stitching, we here denote this change of coordinates as complex-valued matrix $M \in \mathbb{C}^{n\times n}$, where for n is the *latent* dimensionality.

In the following, we quantify the degrees of freedom on the choice of latent coordinate system M eliminated by dynamics and overlap by collecting (sparsity) constraints on a matrix $n \times n$ matrix S. S allows to quantify how to trade overlap against dynamics-related conditions. We find identification of S without overlap is possible up to one bit flip. The approach is constructive and extends the algorithm of Bishop & Yu (2014). We also find surprisingly simple failure cases, showing that stitching from dynamics is far from being 'magical'.

We assume both θ and $\tilde{\theta}$ to be valid system reconstructions of the same underlying linear dynamics system. This leaves the n^2-many degrees of freedom of having different latent representations $\tilde{x} = Mx$, described by some change of basis $M \in \mathbb{C}^{n\times n}$. We can stitch if we know M, and thus seek conditions that allow to uniquely identify this matrix. We first focus on the conditions imposed on the dynamics-related matrices by

$$\tilde{A} = MAM^{-1} \tag{B.3}$$

$$\tilde{Q} = MQM^{\top}. \tag{B.4}$$

The key observation is that M by B.3 defines one canonical basis of $\tilde{A}$, i.e. columns of M define the (generalized) eigenspaces of $\tilde{A}$. When obtaining a second canonical

basis V from $\tilde{A} = V\tilde{J}V^{-1}$, V inherits the same eigenspace structure, up to the choice of representation within each generalized eigenspace. We however know one set of spanning vectors for each generalized eigenspace from columns of V, and hence only need to identify the unique linear combinations (seperature for each generalized eigenspace) to find corresponding columns of M. Depending on the spectrum of A, we can thus uniquely map between V and M using potentially much fewer degrees of freedom (down to n) than without dynamics (n^2). In some cases, Q can be used to fill in the remaining n degrees of freedom, up to an overall sign flip.

Notation: The notation for the eigen-analysis of general matrices $A \in \mathbb{R}^{n\times n}$ can become very cluttered. We will try to avoid excessive sub-indexing where possible. We will generally use j to denote column indices. When working with repeated eigenvalues, we will use subindex i to refer to each distinct eigenvalue $\lambda_i, i = 1, \ldots, r$, $r \leq n$, and subindex j to refer to the repeated λ_j associated with column $j = 1, \ldots, n$.

B.1 Dynamics matrix A

We start with B.3, which states that A and $\tilde{A}$ are similar matrices that share the same eigenvalues. Typically when studying matrix similarity, one is given M and A and then seeks to find $\tilde{A}$. Here we go the other way around, starting out with $(A, \tilde{A})$ and seek constraints on M. Note that we cannot expect $(A, \tilde{A})$ to help identify M in general. This is most easily seen for the example $A = I_n$, which enforces $\tilde{A} = MM^{-1} = I_n$ under any M, and thus does not allow us to learn anything about M. We will see that the problem with the identity I is not so much that all its eigenvalues are non-distinguishable (repeated eigenvalue $\lambda_1 = 1$ has algebraic multiplicity $\mu_1 = n$), but that this sole eigenvalue has a geometric multiplicity $\rho_1 = n > 1$. When $\rho_i \leq 1$ for all r-many distinct eigenvalues $\lambda_i, i = 1, \ldots, r \leq n$, we will see that eq. B.3 reduces the degrees of freedom on M to one free scale per latent dimension—or one free bit flip when columns of M are additionally known to be normalized.

Without loss of generality, we assume $A = J$ to be in Jordan normal form. If this was not the case, we could first apply an initial change of basis to both A and $\tilde{A}$ to ensure A being in Jordan normal form. Denote the column vectors of $M = [m_1, m_2, \ldots m_n]$. We rewrite B.3 to

$$\begin{aligned} \tilde{A}M &= MJ, \\ \tilde{A}m_j &= \sum_k J_{kj} m_k, \end{aligned}$$

where the second line holds for each column index $j = 1, \ldots, n$. Columns of J are highly sparse with only one ($J_{jj} = \lambda_j$) or two (additionally $J_{j-1,j} = 1$) non-zero entries, depending on whether (i) j denotes the first column index within a Jordan block of J,

or (ii) not. Accordingly,

$$\begin{aligned} \tilde{A}m_j &= \lambda_j m_j \qquad \text{if (i)}, \\ (\tilde{A} - \lambda_j I)m_j &= m_{j-1} \qquad \text{if (ii)}. \end{aligned}$$

These conditions establish the columns m_j as generalized eigenvectors of $\tilde{A}$ with eigenvalues λ_i. Thus columns m_j in particular are elements of the respective generalized eigenspaces $\mathrm{eig}_{\lambda_i}(\tilde{A})$ and can be uniquely expressed as a linear combination $m_j = V^{(i)}\tilde{s}_j$ of any valid basis $V^{(i)} = [v_1^{(i)}, v_2^{(i)}, \dots v_{\mu_i}^{(i)}]$ for $\mathrm{eig}_{\lambda_i}(\tilde{A})$.

We obtain one possible basis $V^{(i)}$ for each of the generalized eigenspaces from computing the Jordan normal form of $\tilde{A} = V\tilde{J}V^{-1}$. V here is a canonical basis for $\tilde{A}$ formed from eigenvectors and generalized eigenvectors of $\tilde{A}$, and the $V^{(i)}$ are given by those columns v_j corresponding to distinct eigenvalue λ_i. The number of Jordan blocks corresponding to any distinct eigenvalue λ_i is given by its geometric multiplicity ρ_i. We can reorder the Jordan blocks of both J, $\tilde{J}$ such that $\tilde{J} = J$, i.e.

$$MJM^{-1} = VJV^{-1}.$$

We next need to find coefficients $\tilde{s}_j \in \mathbb{R}^{\mu_i}$. V forms a valid basis for $\mathbb{C}^n$. Without loss of generality, we write

$$M = VS, \tag{B.5}$$

for $S \in \mathbb{C}^{n\times n}$, i.e. the columns m_j are linear combinations $m_j = Vs_j$ of the columns $[v_1, v_2, \dots v_n] = V$, with up-to-here unknown coefficients s_j given by columns of S. The division of $\mathbb{C}^n$ into the generalized eigenspaces of $\tilde{A}$ enforces a potentially strong sparsity on S: Because the block-structures of $J = \tilde{J}$ match, column pairs v_j and m_j are part of the (mutually expressable) respective bases for the same generalized eigenspace, and S is block-diagonal with r-many $\mu_i \times \mu_i$ diagonal blocks. The μ_i-many columns m_j corresponding to one distinct eigenvalue λ_i form ρ_i-many Jordan chains (see condition (ii) above). Due to the recursive definition of Jordan chains from a single 'generator' vector per chain, all μ_i-many columns m_j for any distinct eigenvalue can be computed from ρ_i-many basis vectors. Once the linear coefficients $\tilde{s}_j$ for the generator vector are known, we can compute the coefficients for all other generalized eigenvectors for this distinct eigenvalue as $\mu_i \times \mu_i$ coefficient block matrix S_i:

$$S_i = \begin{bmatrix} S_i^{(11)} & S_i^{(12)} & \dots & S_i^{(1\rho_i)} \\ S_i^{(21)} & S_i^{(22)} & \dots & S_i^{(2\rho_i)} \\ \vdots & \vdots & \ddots & \vdots \\ S_i^{(\rho_i 1)} & S_i^{(\rho_i 2)} & \dots & S_i^{(\rho_i \rho_i)} \end{bmatrix}, \tag{B.6}$$

where

$$S_i^{(kl)} = \begin{bmatrix} \dots & \tilde{s}_{i,n_k-2}^{(kl)} & \tilde{s}_{i,n_k-1}^{(kl)} & \tilde{s}_{i,n_k}^{(kl)} \\ & \vdots & \vdots & \vdots \\ \dots & \tilde{s}_{i,1}^{(kl)} & \tilde{s}_{i,2}^{(kl)} & \tilde{s}_{i,3}^{(kl)} \\ \dots & 0 & \tilde{s}_{i,1}^{(kl)} & \tilde{s}_{i,2}^{(kl)} \\ \dots & 0 & 0 & \tilde{s}_{i,1}^{(kl)} \end{bmatrix} \in \mathbb{C}^{n_k \times n_l}, \tag{B.7}$$

n_k is the size of the k-th Jordan block, and $S = \mathrm{diag}\,(S_1, S_2, \dots, S_r)$. The Toeplitz form of B.7 results from V being a canonical basis, i.e. the columns v_j associated with distinct eigenvalue λ_i also form ρ_i-many Jordan chains $(\tilde{A} - \lambda_i I_n)v_j = v_{j-1}$.

In summary, to construct S we require $\rho_i\mu_i$-many unique coefficients $\tilde{s}_i$ for each distinct λ_i, which over all distinct eigenvalues sums up to $n \leq \sum_{i=1}^{r} \mu_i\rho_i \leq n^2$-many remaining degrees of freedom on $M = VS$.

Pairwise distinct eigenvalues: In the special case of pairwise distinct eigenvalues λ_i, we have $\rho_i \leq \mu_i = 1$ for all $i = 1, \dots, n$ and $S = \mathrm{diag}\,(S_1, S_2, \dots, S_n)$ is a diagonal matrix holding a single free scale per latent dimension. Since for diagonal S, the entries S_{jj} in $M = VS$ just scale the norm of m_j, we for column-normalized matrices M obtain $S_{jj} \in \{|v_j|^{-1}, -|v_j|^{-1}\}$, i.e. each $S_{jj}, j = 1, \dots, n$ is immediately known up to a single bit.

B.2 Latent covariance matrix Q

Under a change of basis V^{-1} that sets $\tilde{A} = V\tilde{J}V^{-1}$ to be in Jordan normal form, we have

$$\tilde{Q} = MQM^\top = VSQS^\top V^\top, \tag{B.8}$$

and

$$V^{-1}\tilde{Q}V^{-T} = SQS^\top,$$

and with $S = V^{-1}M$ as in the previous section. We denote $V^{-1}\tilde{Q}V^{-T} = \bar{Q}$ and obtain

$$\bar{Q}_{ki} = \sum_{l,j} S_{kl}Q_{lj}S_{ij},$$

i.e. the parameter reconstructions Q and $\bar{Q}$ contain information only on the pairwise products of entries of S. As we will see, this is sufficient to resolve remaining scale and relative-sign ambiguities on diagonal S that were not resolved from A and $\tilde{A}$. A flip of all mathematical signs for all latent coordinates $\mathbf{x}_i$ however jointly cannot be resolved from $Q, \tilde{Q}$.

Pairwise distinct eigenvalues: As we have seen, $A, \tilde{A}$ in this case ensure that S is diagonal and it immediately follows from the above identity that

$$S_{ii} = \pm\sqrt{\frac{\bar{Q}_{ii}}{Q_{ii}}}$$
$$\operatorname{sign}(S_{ii}S_{jj}) = \operatorname{sign}\left(\frac{\bar{Q}_{ij}}{Q_{ij}}\right) \qquad \forall (i,j) : Q_{ij} \neq 0$$

Note how we require non-zero latent covariances $Q_{ij}, \bar{Q}_{ij}$ to identify pairwise products of mathematical signs. If the latent Gaussian noise is independent under any coordinate system under which the dynamics matrix is diagonalized, i.e. $Q = I$ (by assumption) or $\bar{Q} = I$ (by construction of $\bar{Q}$), we cannot resolve sign ambiguities. This is somewhat intuitive, as in the case where both A and Q jointly diagonalize, the latent dynamics fully decouple.

Note on the general case: We omit the general case here. The hierarchical structure on S (with upper triangular Toeplitz matrices forming the blocks of diagonal blocks of S) quickly makes even just the notation cumbersome. Intuitively, this case appears very challenging, also because the eigenvalue spectra of A and Q are not necessarily related in any way.

B.3 Emission matrix C

We have

$$\tilde{C} = CM^{-1} = \left(CV^{-1}\right) S^{-1} \tag{B.9}$$

with unknown target change of basis M and a second change of basis V^{-1} that sets $\tilde{A} = MJM^{-1} = V\tilde{J}V^{-1}$ to be in Jordan normal form $\tilde{J}$. We denote $\bar{C} = CV^{-1}$. The structure on S depends on the geometry of the eigenspaces of $(\tilde{A}, A)$. As we will show, this structure in turn has direct consequences on required overlap between rows observed both in $\bar{C}$ and C.

Without incorporating dynamics, Bishop & Yu (2014) required at least n rows of overlap $\tilde{C}_{(k,:)}$, $C_{(k,:)}$ to identify $M \in \mathbb{R}^{n\times n}$. Expressing $M = VS$ through the latent basis V obtained from dynamics matrix $\tilde{A} = V\tilde{J}V^{-1}$ allows to work with block-diagonal matrix S instead of potentially dense M. For the $r \leq n$ distinct eigenvalues $\lambda_i, i = 1, \ldots, r$ with $1 < \rho_i \leq \mu_i$, the matrix S most generally consists of r-many $\mu_i \times \mu_i$ diagonal blocks with in total $\sum_{i=1}^{r} \mu_i\rho_i$ unknown values.

Right-multiplication with S (or its inverse S^{-1} with same block-structure) divides the columns of $\bar{C}$ into non-overlapping groups corresponding to the distinct eigenvalues λ_i. This decouples the identification of the rows of S^{-1} corresponding to different λ_i,

$$\bar{C}_{(:,I_i)} = C_{(:,I_i)}S_i^{-1},$$

where $I_i \subseteq \{1, \ldots, n\}$ collects all columns j associated with distinct eigenvalue λ_i. This effectively reduces the problem from a single matrix idenfitication with dimensionality n to r-many matrix identifications with dimensionalities μ_i that can be solved in parallel. Hence the

$$\sum_{i=1}^{r} \mu_i \rho_i \leq \sum_{i=1}^{r} (\mu_i \rho_{\max}) = \rho_{\max} \sum_{i=1}^{r} \mu_i = \rho_{\max} n$$

many unknown entries of S can be idenfitied from $\rho_{\max} := \max_i\{\rho_i\} \leq n$ many rows in overlap.

We fall back onto the demand of n variables in overlap only if $A = \lambda_1 I_n$ is similar to a scaled version of the identity (which subsumes the case of 'no dynamics' $\lambda_1 = 0$). A second distinct eigenvalue $\lambda_2 \neq \lambda_1$ reduces the required overlap $\rho_{\max} \leq n - \min(\rho_1, \rho_2)$, and so on.

Pairwise distinct eigenvalues: In the special case of pairwise distinct eigenvalues S is a diagonal matrix which via B.9 can potentially be idenfitied from a single pair of corresponding non-zero rows $\bar{C}_{(k,:)}$, $C_{(k,:)}$ from

$$S_{ii} = \frac{C_{(k,i)}}{\bar{C}_{(k,i)}} \tag{B.10}$$

If $\bar{C}_{(k,i)} = 0$ for any $i = 1, \ldots, n$ for this particular row k, we cannot identify S and require to observe additional rows. This issue stems from solving r-many matrix identifications with dimensionalities $\mu_i = 1$, i.e. we require each $\bar{C}_{(k,i)}$ to be full rank. $\bar{C}_{(k,i)} = 0$ has rank zero. On the other hand, a scaling factor $S_{ii} = 0$ for diagonal S cannot occur as then the change of basis M is invalid due to $|M| = |V||S| = 0$.