The Incidental Parameter Problem in Network Analysis for Neural Spiking Data

by

Dahlia Nadkarni

B.Tech., Indian Institute of Technology - Bombay; Mumbai, India, 2009

M.Tech., Indian Institute of Technology - Bombay; Mumbai, India, 2009

Sc.M., Brown University; Providence, RI, 2011

A dissertation submitted in partial fulfillment of the

requirements for the degree of Doctor of Philosophy

in The Division of Applied Mathematics at Brown University

PROVIDENCE, RHODE ISLAND

May 2015

This dissertation by Dahlia Nadkarni is accepted in its present form
by The Division of Applied Mathematics as satisfying the
dissertation requirement for the degree of Doctor of Philosophy.

Date_____ _____

Matthew Harrison, Ph.D., Advisor

Recommended to the Graduate Council

Date_____ _____

Stuart Geman, Ph.D., Reader

Date_____ _____

Asohan Amarasingham, Ph.D., Reader

Approved by the Graduate Council

Date_____ _____

Peter M. Weber, Dean of the Graduate School

Acknowledgments

This thesis would not have been possible without the advice, support, and encouragement of many others.

First and foremost, I want to acknowledge my advisor, Matt Harrison. I first got to know Matt as a student in his Recent Applications in Probability and Statistics class. The enthusiasm that he exuded and the lucid explanations that he conveyed as a lecturer stimulated my interest in the topic. Eager to learn more, I began attending his group meetings, and I soon became his student. In an advisory role, I am truly grateful for the time Matt has invested in my growth; whether it was explaining a technical detail, outlining a proof, or suggesting a change in my exposition, Matt was always clear, patient, and supportive. I truly value his close reading of my writing and his willingness to accommodate my schedule. On a more personal level, I appreciate how welcoming Matt was. Going to lunch with Matt on Thayer Street and being invited to get-togethers at his home helped give me a sense of belonging as a member of the Applied Math community at Brown and beyond. Lastly, I would like to thank Matt for his guidance in my transition from PhD student to the next phase of my career.

I would like to thank my readers, Stu Geman and Asohan Amarasingham, for taking the time to closely read through my dissertation. Several structural improvements of the document are due to their valuable feedback.

I am grateful to my mathematical siblings, Jeff and Dan. Their insight, perspective, and kindness helped make our meetings with Matt enjoyable and productive, and pushed me to be a better mathematician myself. I want to specifically thank Dan for his help with the

finer points of Python and network visualizations, and Jeff for his thoughtful and intelligent research advice, particularly with regard to the Bayesian non-parametric part of my work.

My fellow graduate students made the department a fun place to be and were always encouraging as we worked through assignments, discussed research, and talked about our ambitions, hopes, and fears. It has been wonderful to be part of a community that strives for excellence while celebrating each other's successes, and has fun all the while. In particular, I would like to thank Laura, Kelly, Liz and Heyrim for their friendship. I am grateful for the supportive and, at times, joyously chaotic atmosphere created by my office mates Dan, Nat, Joe, and Tasos. They always made our office a fun place to be, which was invaluable in times of duress. I would also like to thank Ivana for her close reading and careful edits of my dissertation. The staff in the department was a joy to work with. They greeted me with open arms when I arrived and maintained a welcoming community with their generous spirit and well-organized departmental activities.

The Indian community at Brown helped make my transition to the US easy and enjoyable. From movie nights to Diwali celebrations to birthday parties, there was always something fun around the corner. I particularly wish to acknowledge Teja, with whom I spent countless hours chatting, laughing, and cooking.

Finally, I am ever so grateful for the support of my family. Throughout my PhD, my parents and grandparents were always there to support me, calm me down, and push me when I needed it. I would not have come to other side of the globe to pursue my PhD if not for their constant encouragement and the belief that they instilled in me as a young child. My brother, Ishan, has been an unwavering source of positive energy; he has challenged me to take risks, remain optimistic, and not always take everything so seriously ;). Roz, Michael, Miriam, and Rachel embraced and supported me throughout and I am lucky to have them as my family. Finally, I owe a debt of gratitude to my husband, Jonah, who has been the stabilizing presence, always by my side, in my life everyday.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The spiking electrical activity of simultaneously recorded neurons is often modeled as a multivariate binary time series [15, 10, 12, 39]. Inference about the fast temporal correlation structure of this multivariate time series, such as zero-lag synchrony and precise lag-lead relationships, is complicated by the time varying response of the neurons to their many unobserved and correlated inputs[1, 11]. For example, Figure 1.1 shows the cross-

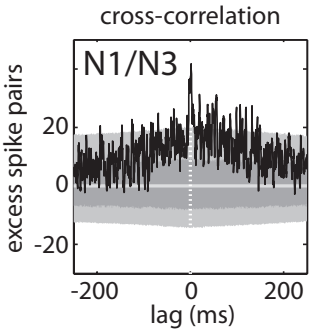


Figure 1.1: Cross-correlation histogram between two spike trains recorded from monkey motor cortex (Figure taken from Amarasingham et al. [1])

correlation histogram between two neuronal spike trains recorded from the motor cortex of a monkey's brain. We see a sharp spike close to 0 ms indicating zero-lag synchrony between the two neurons N1 and N3. In addition, we also see positive correlation between spiking around 0 ms, but at coarser time scales of ± 100 ms. We want to be able to isolate

the zero-lag synchrony from the other correlated effects at coarser time-scales.

The signal of interest is often synchronous firings or precise lag-lead relationships between neurons at times scales that are consistent with mono-synaptic delays of 1-4 ms. We want to separate these fast temporal dynamics from other slower dynamics that likely result from different mechanisms. These background effects can be thought of as noise in the system and modeling this noise often involves making some assumptions about its structure. The background effects are often non-stationary in time and assuming them to be stationary would lead to a misspecified model. Other assumptions, such as piecewise constant background effects, lead to the incidental parameter problems, or the Neyman-Scott [32] problem, in which the number of nuisance parameters grows with the size of the data.

Instead of focusing inference efforts on capturing the noise model accurately, we develop a conditional inference approach to infer the signal of interest (the fast temporal correlation structures such as zero-lag synchrony) while being robust to the dynamics at slower time scales. Previous work using conditional inference was focused on statistical hypothesis testing, often called jitter [1, 22, 26]. Here, we explore the conditional framework on a parametric model. In addition to identifying the significant dynamics, a parametric framework shall allow us to compare different parameters against each other. We demonstrate this conditional approach on log-linear models (maximum entropy models) for zero-lag synchrony [30, 35, 39, 40, 41] but the analysis can easily be extended to incorporate lags and leads, as well as other types of models. We also develop a composite likelihood (or pseudo-likelihood) approximation to the conditional approach that can scale to larger number of neurons and more complex models. In addition to these conditional methods, we also explore a Bayesian nonparametric approach with Dirichlet process mixture (DPM) priors to model the background effects without imposing any parametric assumptions on the noise model. The nonparametric framework helps avoid the incidental parameter problems that arise while estimating the growing number of nuisance parameters.

We begin by exploring the incidental parameter problem and the different approaches that have been proposed in literature to tackle this issue. We then extend these techniques to our setting of neuronal firings to estimate the synchrony parameter between neurons while being robust to non-stationary firing rates of individual neurons.

1.1 Incidental parameter problem and background literature

The incidental parameter problem has been studied extensively in the statistics literature, as well as in econometrics (for panel data models), and the psychometric and educational testing literature (Rasch models).

The incidental parameter problem was first introduced and defined by Neyman and Scott in 1948 [32]. In fact, the incidental parameter problem is often referred to as the *Neyman-Scott problem* in literature. Lancaster [28] provides a brief survey of the incidental parameter problem in statistics and econometric literature starting with the Neyman-Scott [32] paper. The paper defines two types of parameters, incidental and structural. Consider a sequence of independent random variables (X_i) modeled by a set of parameters. The structural parameters $\theta = (\theta_1, \ldots, \theta_d)$ are finite in number and appear in the probability distribution of an infinite number of these random variables (X_i). Incidental parameters[1] $\nu = (\nu_1, \nu_2, \ldots)$ are those that appear in the distribution of only a finite number of random variables, and hence information about each of these incidental parameters depends only on a finite set of observations. As the number of observations increase, we gather more information about the structural parameters θ, but information about each particular ν_i is obtained only from few observations. In the standard setting – with finite parameters and infinitely growing data – the theory of maximum likelihood estimation guarantees consistent estimates for the parameters. However, in the presence of incidental parameters, we

[1]The term nuisance parameters refer to parameters that are not of ultimate interest, but necessary in the model for inferring the parameters of interest. Incidental parameters are nuisance parameters that appear in the distribution of only a finite number of data points and grow with the amount of data.

cannot hope to get consistent estimates for the incidental parameters since they appear in the distribution of only a finite number of observable variables. Furthermore, inference for the structural parameters is also affected by the presence of the incidental parameters and standard methods, such as maximum likelihood estimation, may not give consistent estimates for the structural parameters as well. Neyman-Scott (1948) [32] illustrates mainly two types of problems that arise in the presence of incidental parameters – one where the estimates for structural parameters are not consistent, known as the 'incidental parameter problem', and another where the estimates are consistent but not asymptotically normal. We shall focus on the first type of problem and we analyze one of the examples from the Neyman-Scott paper [32] in detail in Chapter 2.

In Section 1.1.1 we focus on incidental parameter problem in the statistics literature. We explore the two main approaches used to tackle this problem – conditional approach and marginal approach. In Section 1.1.2 we summarize the econometrics literature, particularly the panel data models, and in Section 1.1.3 we survey the Rasch model literature.

1.1.1 Overview of statistics literature

The statistics literature has a rich history of research about dealing with incidental parameters.

Lancaster [28] provides a brief survey of the incidental parameter problem in the statistics and econometric literature. Broadly speaking, there are two different approaches in the statistics literature to tackle the incidental parameter problem - the *conditional frequentist* approach, which involves conditioning on statistics sufficient for the incidental parameters, and the *Bayesian or the marginal approach*, which involves integrating the incidental parameters out using a suitable prior. We introduce the two concepts briefly here, and will explore each concept in more details as we see different examples (panel models in Section 1.1.2, Rasch models in Section 1.1.3, Neyman-Scott Gaussian model from [32] in Chapter 2 and our two-neuron zero-lag synchrony model).

Conditional approach

The main idea in the conditional approach is to factor the likelihood into two parts such that one part is independent of incidental parameters, and then we can use this partial likelihood for inference. Lancaster [28] refers to this as *partial likelihood*[17] procedures since the inference is based only on a part of the likelihood and summarizes the three main categories of models where this idea can be applied. Let θ be the structural parameter, $\nu = (\nu_1, \nu_2, \ldots)$ be the incidental parameters and $S(X)$ be a sufficient statistic for ν.

If the likelihood f factors as follows,

$$f(x|\theta, \nu) = g(S(x)|\nu) \cdot \psi(x|S(x), \theta),$$

θ and ν are said to be *likelihood orthogonal* since $\frac{\partial^2}{\partial\theta\ \partial\nu} \log f = 0$ and we can infer θ based on ψ alone, since ψ has no dependence on ν.

This method works even in cases where the distribution $g(S(x))$ has θ dependence:

$$f(x|\theta, \nu) = g(S(x)|\nu, \theta) \cdot \psi(x|S(x), \theta).$$

Inference can still be based on ψ alone, although we would lose some information about θ in g. We shall see later in Chapter 3 that the zero-lag log-linear model for neurons falls in this category.

For the above two categories, estimation involves using only the conditional distribution ψ and not the full likelihood f and hence these methods are known in the statistics literature as the *conditional likelihood* approach.

The third category of problems mentioned in [28] where the partial likelihood methods can

be applied is when there exists a statistic $L(X)$ such that the likelihood factors as:

$$f(x|\theta,\nu) = \tilde{g}(L|\theta) \cdot \tilde{\psi}(x|L,\nu,\theta),$$

in which case inference can be based on $\tilde{g}$. In this thesis we focus on the first two categories of problems.

Reid and Cox [18, 19, 34] have studied the conditional inference in depth for many years for a variety of different problems. When a separation of likelihood discussed above is not possible, but we have *information orthogonality* (i.e. block diagonality of the Fisher information matrix), [18] suggests approximate conditional inference using conditional likelihood given the maximum-likelihood estimates of the orthogonalized parameters. They have also studied pseudo-likelihood approximations to conditional inference [19] for high dimensional parameters using only pairwise joint distributions. We shall use this idea for our composite likelihood approach in Chapter 5.

Anderson [3] proved the consistency of the conditional maximum likelihood estimates (CMLE) under strong conditions. These conditions are satisfied by the random effects model (one where the incidental parameters ν are treated as random variables with some probability distribution). He also proposed a conditional likelihood ratio test using the conditional likelihood in [2], which can be inverted to obtain confidence intervals.

We explore the conditional inference approach in detail for our two-neuron zero-lag synchrony model in Chapters 3 and 4. Anderson's result from [3] does not directly apply to our two-neuron model, but we adapt his proof to establish consistency of the CMLE in our case. In addition to the point estimates, we also obtain confidence intervals using the conditional likelihood as proposed in [2]. To scale more easily to larger networks, we develop an approximate conditional method based composite likelihood (pseudo-likelihood) as discussed in [19].

Marginal approach

The *marginal or the Bayesian approach* involves treating the incidental parameters $\nu = (\nu_1, \nu_2, \ldots)$ as random variables and integrating them out using some prior. Berger investigated these integrated likelihood methods for different models in [6]. The choice of prior is very important for this approach to work well, much more than in the classical finite-dimensional parameter setting. In general, the Bayesian approach will not help resolve the incidental parameter problem if the prior on ν's is not accurate. We will see in Chapter 2 that a non-informative uniform prior on the incidental parameters ν for the Neyman-Scott Gaussian model gives consistent estimates for the structural parameter θ [28], but using other priors results in inconsistent estimates unless we know the true priors on ν exactly. A hierarchical approach can solve this issue as long as we accurately identify the form of the prior distribution on ν. Berger [6] suggests the use of uninformative flat priors (uniform priors), which works for a range of models (as in the case of Neyman-Scott Gaussian model), but it does not solve the consistency issues for all problems, particularly in panel models with logit or probit distributions [14, 28] (more details in Section 1.1.2). In general, marginalizing may not solve the inconsistency issue for the structural parameter unless the prior used in the model for the incidental parameters depicts the true prior, or the model is flexible enough to capture any type of distribution of ν (for example, using non-parametric priors on ν). Kiefer and Wolfowitz [27] prove the consistency of the maximum likelihood estimates using the marginal likelihood when the distribution on ν is also inferred along with the structural parameter θ (note that there is no parametric assumption on the distribution on ν). This can be achieved by using a non-parametric framework. In Chapter 6, we use a non-parametric Dirichlet process mixture (DPM) prior for the two-neuron zero-lag model, which appears to address the incidental parameter problem.

So far, we have seen the overall approach used to handle the incidental parameters in the statistics literature. We now take a closer look at some of the problems in panel model and Rasch model literature where these methods are used. We shall then explore these ideas

in detail on the Neyman-Scott Gaussian model in Chapter 2, before applying them to our neuron model.

1.1.2 Panel data models

Incidental parameter problems also appear frequently in econometric literature in panel data models with fixed agent-specific effects. Panel data refers to multivariate data for multiple individuals or groups (known in the literature as agents or panels) recorded over time at discrete time points. Panels could represent different individuals, groups of people, economic firms, or even nations. In panel models the parameter of interest θ measures the dependence of the response variable with the covariates, whereas nuisance parameters ν_i's are often introduced to capture the agent specific effects that may be correlated with the covariates. Let us look at some examples of panel data models from [14, 28]. In these examples, we use the notation Y for the response variable and X for the covariates and $i = 1, \ldots, N$ indexes the panels.

Linear regression model with agent specific intercepts:

$$Y_{it} = \theta X_{it} + \nu_i + \epsilon_{it} \ , \qquad \epsilon_{it} \sim \mathcal{N}(0, \sigma^2)$$

Auto-regression model:

$$Y_{it} = \theta Y_{i,t-1} + \nu_i + \epsilon_{it} \ , \qquad \epsilon_{it} \sim \mathcal{N}(0, \sigma^2)$$

Panel logit: If y_{it} is binary,

$$\mathbb{P}(Y_{it} = 1 | \nu_i, \theta, X) = \frac{e^{\nu_i + \theta X_{it}}}{1 + e^{\nu_i + \theta X_{it}}}$$

Panel probit: If y_{it} is binary, and $\Phi(\cdot)$ represents the standard normal cumulative distri-

bution function,

$$\mathbb{P}(Y_{it} = 1|\nu_i, \theta, X) = \Phi(\nu_i + \theta X_{it})$$

There are two approaches to model the likelihood: fixed effects and random effects. In a *fixed effects* model, the ν_i's are treated as fixed constants and the likelihood is written conditional on the ν_i's. On the other hand a *random effects* approach treats the ν_i's as random variables and the likelihood is written marginal on the ν_i's, i.e. we try to integrate these parameters out with respect to some distribution. This distribution could depend on the covariate sequence (X_{it})'s, but is often modeled as independent of the covariates [14, p. 233], [28, p. 396]. For example, a random effects approach for the linear regression model might assume $\nu_i \sim \mathcal{N}(0, \sigma_\mu^2)$. These two approaches correspond to the *conditional* and the *marginal/Bayesian* framework discussed in the previous Section 1.1.1.

Chamberlain [14] explores the method of conditioning on sufficient statistics as well as the random effects approach for different panel data models – linear regression, auto regression, multinomial logit model, etc. – to obtain consistency of structural parameters in presence of incidental parameters. For both panel logit and panel probit, conditioning helps to overcome the incidental parameter problem, but the random effects model with uniform prior (as suggested by [6]) does not give consistent estimates for θ.

1.1.3 Rasch models

The Rasch models, which are widely used for social and psychological data and for item analysis in educational testing, also suffer from the incidental parameter problem when the number of tests or questions are held fixed, but the number of individuals being tested goes to infinity. Consider a setting with n different individuals taking a test with k different questions. Let the data be $(X_{ij} \ : \ i = 1, \dots n, \ j = 1, \dots, k)$ where each X_{ij} indicates if individual i responded to item j correctly. Then the basic Rasch model is given by

$$\mathbb{P}(X_{ij} = 1) = \frac{e^{(\nu_i - \theta_j)}}{1 + e^{(\nu_i - \theta_j)}}.$$

Here ν_i is the ability parameter for the i-th individual and θ_j is the item difficulty parameter for the j-th item. The θ's are the parameter of interest and ν's are considered nuisance parameters. In the setting when the number of items k is fixed, but the number of individuals n goes to infinity, the standard maximum likelihood estimate (MLE) for θ is inconsistent [23].

The Rasch model literature refers to the fixed effects model (where the ν's are assumed to be some fixed constants) as the *functional model* and the random effects model (where ν's are thought of as random variables with some distribution) as the *structural model*.

De Leeuw and Verhelst [20] unify all the different variations of the Rasch model used in literature under one general Rasch model and use this to compare the different techniques for handling incidental parameters, in particular the unconditional maximum likelihood estimate(MLE), the conditional maximum likelihood estimate (CMLE) and the marginal maximum likelihood estimate (MMLE). Let us denote the distribution on ν_i as H_i with density h_i. Based on what assumptions we make on the H_i we can have different models and different types of estimates:

If H_i are assumed to be step functions with a single step, i.e the density h_i is a Dirac-delta function at a value $\hat{\nu}_i$ given by $h_i = \delta(\nu_i = \hat{\nu}_i)$, then this model reduces to the *fixed-effects* or the *functional model* and the estimates are the unconditional maximum likelihood estimate (MLE).

If all H_i are assumed to be the same, $H_i = H \ \forall i$, then this is the same as the *random effects* or the *structural model*. Further, H can be assumed to be exactly known, or to be some parametric family, or even some nonparametric distribution. If H is assumed to be known exactly, the ν's can be integrated out using this H to get the marginal likelihood in terms of θ. We refer to the maximum likelihood estimate of this marginal likelihood as the maximum marginal likelihood estimate (MMLE) in Chapter 2. If instead H is taken to be some parametric family with an unknown hyper-parameter c, the hyper-parameters can be jointly estimated with θ. This is similar to a hierarchical Bayesian approach (see Section

2.3 for more details) and we refer to the joint (θ, c) estimates as the hierarchical maximum likelihood estimate (HMLE). A distribution that is often used for ν's in the Rasch model literature is the finite mixture model [29], where the population is treated as a collection of different groups with group specific parameters and latent variables are introduced to define the group assignment for each ν_i.

If we assume that the H is completely unknown, we can use a non-parametric distribution on the ν's. De Leeuw and Verhelst [20] refer to this type of model as the *extended Rasch model* and treat it separately from the *structural model*. Another way to think about the nonparametric framework is to extend the finite mixture model on ν's to allow for variable number of groups (one can estimate the number of groups) even including the possibility of infinite groups. Lindsay et al. [29] refer to this model as a *semi-parametric model* (since the distribution on incidental parameters ν is non-parametric, whereas the distribution on the structural parameters θ is parametric). Kiefer and Wolfowitz [27] proposed and proved consistency for this type of model (using a continuous distribution H on the ν's and estimating the distribution H simultaneously with θ) and hence these are also known in literature as *Kiefer-Wolfowitz* estimates.

De Leeuw and Verhelst [20] and Lindsay et al. [29] also motivate conditional inference in an alternate way, from a Bayesian perspective, by showing that the conditional MLE and the *Kiefer-Wolfowitz estimates* are closely related, at least in the context of Rasch models. [20] proves that the nonparametric marginal estimates (*Kiefer-Wolfowitz* estimates) are identical to the conditional estimates with probability 1. The unified general Rasch model framework allows for such a comparison to be made. See [29] and [20] for more details.

More recently, Zwinderman [42] suggested the use of composite likelihood (or pseudo-likelihood) methods based on pairs of items taken together at a time. Our conditional composite approach (Section 5) is similar to this idea.

We combine all these different ideas from literature and apply them to our log-linear model for zero-lag synchrony.

1.2 Outline

In Chapter 2 we explore the Neyman-Scott Gaussian model from [32] and various approaches to tackle the incidental parameter problem for this model in detail. We derive the asymptotic properties for the Gaussian model analytically, and then extend the ideas we develop here for our neuron model.

In Chapter 3 we introduce the two-neuron log-linear model for zero-lag synchrony. We explore how the stationary model leads to model misspecification, whereas the non-stationary (piecewise-constant) model leads to incidental parameter problems on simulated data. We introduce the idea of conditional inference for this model and demonstrate in simulation how this leads to consistent estimates.

In Chapter 4 we prove consistency for the conditional maximum likelihood estimates (CMLE) and also obtain confidence intervals using conditional likelihood. We then apply this method to neuronal recordings from the medial prefrontal cortex of a rat running through a maze to infer neuron connectivity.

In Chapter 5 we extend the idea of conditioning to larger networks. As the network size grows, the CMLE computation becomes more intensive. We explore composite likelihood methods (also known as pseudo-likelihood methods) to tackle this computational issue.

In Chapter 6, we explore an alternate non-parametric Bayesian approach using Dirichlet process mixture (DPM) priors on the incidental parameters.

Chapter 2

Neyman-Scott problem for Gaussian distribution

In this chapter we explore the different approaches for estimating parameters of interest in the presence of incidental parameters on a Gaussian model. Although Gaussian models are not our eventual interest, the mathematics is easier in the Gaussian setting and is useful for building intuition.

We first describe the problem setting. Section 2.1 explores the unconditional maximum likelihood estimation, Section 2.2 looks at the conditional approach, and Section 2.3 looks at the marginal approach using three different priors on the incidental parameters.

Let us focus on just one-dimensional data coming from a Gaussian probability density function f. Let $X = (X_1, X_2, \ldots, X_T)$ be independent Gaussian random variables with common variance θ and means $\mu = (\mu_1, \mu_2, \ldots, \mu_T)$, so that

$$\begin{aligned} X_t \mid \mu, \theta &\sim \mathcal{N}(\mu_t, \theta) \qquad \forall\ t = 1, 2, \ldots, T \\ f_{\mu,\theta}(x) &= \prod_{t=1}^{T} \frac{1}{\sqrt{2\pi\theta}} \exp\left(-\frac{(x_t - \mu_t)^2}{2\theta}\right). \end{aligned}$$

Further suppose also that μ is slow varying and can be taken to be piecewise constant in smaller time windows of length D. Partitioning the data into W windows of length D, where w-th window Ω_w is given by $\Omega_w = \{t \in \mathbb{N} : (w-1)D + 1 \leq t \leq wD\}$, we denote the common value of μ_t in the w-th window to be ν_w: $\mu_t = \nu_w \ \forall \ t \in \Omega_w$. For simplicity, let us assume that T is a multiple of D and let $W = T/D$. This re-parameterization gives

$$X_t \mid \nu, \theta \sim \mathcal{N}(\nu_w, \theta) \qquad \forall \ t \ \in \Omega_w, \ \forall \ w = 1, \ldots, W.$$

We re-index the data with respect to which window they are in and denote X_t by X_{wd}, where $w = \lceil t/D \rceil$ and $d = t - (w-1)D$. Hence,

$$X_{wd} \mid \nu, \theta \sim \mathcal{N}(\nu_w, \theta) \qquad \forall \ d = 1, \ldots, D \quad \text{and} \quad w = 1, \ldots, W.$$

This was one of the examples used by Neyman and Scott in [32] to illustrate the incidental parameter problem. Here θ is the parameter of interest and $\nu_1, \nu_2, \ldots, \nu_W$ are the incidental parameters.

We derive all the details in Appendix B and provide only the main results in this chapter.

2.1 Maximum likelihood estimation

The likelihood is given by

$$\begin{aligned} f_{\nu,\theta}(x) &= \prod_{w=1}^{W} \prod_{d=1}^{D} \frac{1}{\sqrt{2\pi\theta}} \exp\left(-\frac{(x_{wd} - \nu_w)^2}{2\theta} \right) \\ \log f_{\nu,\theta}(x) &= -\frac{WD}{2} \log 2\pi - \frac{WD}{2} \log \theta - \sum_{w=1}^{W} \sum_{d=1}^{D} \frac{(x_{wd} - \nu_w)^2}{2\theta}. \end{aligned}$$

The maximum likelihood estimate (we shall denote this as the *MLE* from here onwards) for this is given by

$$\begin{aligned}\hat{\nu}_w^{MLE} &= \frac{1}{D}\sum_{d=1}^{D} x_{wd} \qquad\qquad \forall\ w = 1, \ldots, W \\ \hat{\theta}^{MLE} &= \frac{1}{WD}\sum_{w=1}^{W}\sum_{d=1}^{D} x_{wd}^2 - \frac{1}{WD^2}\sum_{w=1}^{W}\left(\sum_{d=1}^{D} x_{wd}\right)^2 = \frac{DM_2 - M_1}{D},\end{aligned}$$

where

$$M_2 = \frac{1}{WD}\sum_{w=1}^{W}\sum_{d=1}^{D} x_{wd}^2 \qquad \text{and} \qquad M_1 = \frac{1}{WD}\sum_{w=1}^{W}\left(\sum_{d=1}^{D} x_{wd}\right)^2. \tag{2.1}$$

Assume that the data was generated with true parameters $\nu^* = (\nu_w^* : w = 1 \ldots, W)$ and θ^* as follows

$$X_{wd} \mid \nu^*, \theta^* \sim \mathcal{N}(\nu_w^*, \theta^*) \qquad \forall\ d = 1, \ldots, D \quad \text{and} \quad w = 1, \ldots, W. \tag{2.2}$$

We have $\mathbb{E}[X_{wd}^2] = {\nu_w^*}^2 + \theta^* \ \ \forall\ w, d$ and $\mathbb{E}[X_{wd_1}X_{wd_2}] = {\nu_w^*}^2 \ \ \forall\ w,\ d_1 \neq d_2$.

Standard MLE gives inconsistent results (see Appendix B.1 for details). As we gather more data and the number of windows $W \to \infty$, we get

$$\hat{\theta}^{MLE} = \frac{DM_2 - M_1}{D} \longrightarrow \frac{D-1}{D}\theta^*.$$

We are trying to infer large number of incidental parameters, each from a finite amount of data. If we take the window width to be D, each ν_w is inferred from only the corresponding D entries in each window. Since these estimates are used to calculate the MLE for θ, it leads to inconsistent results for $\hat{\theta}^{\text{MLE}}$. This is known as the *incidental parameter problem*. In Chapter 1, we introduced the two main approaches suggested in literature to tackle the incidental parameter problems. We now explore these methods, both conditional and marginal, for this Gaussian model.

2.2 Conditioning

The conditional approach involves finding some statistics of the data such that when you condition on these statistics, the remaining conditional likelihood depends only on the structural parameters θ. Conditioning on statistics sufficient for the incidental parameters ν ensures that the conditional likelihood has no ν dependence. One can then infer the structural parameters from the conditional likelihood.

The sum statistic $S_w(X) \triangleq \sum_d X_{wd}$ is sufficient for the incidental parameters ν_w. We condition on the observed values of the sum statistics $s_w = S_w(x) = \sum_d x_{wd}$. If f is the density function for X, we can write $f(x) = g(S(x)) \cdot \psi(x|S(x))$ where g is the marginal density of $S(X)$ and ψ is the conditional likelihood. The conditional MLE is given by (see Appendix B.2 for details)

$$\begin{aligned} \hat{\theta}^{\text{CMLE}} &= \frac{1}{WD(D-1)} \sum_{w=1}^{W} \left[D \sum_{d=1}^{D} x_{wd}^2 - \left(\sum_{d=1}^{D} x_{wd} \right)^2 \right] \\ &= \frac{DM_2 - M_1}{D-1} , \end{aligned}$$

where M_1 and M_2 are as defined in Equation (2.1). If the data is generated from the true model (2.2), we get (see derivation of Equation (B.8) for details)

$$\hat{\theta}^{\text{CMLE}} = \frac{DM_2 - M_1}{D-1} \longrightarrow \theta^*. \tag{2.3}$$

Conditional Inference gives consistent estimates.

2.3 Marginalization

The marginal approach involves integrating the incidental parameters ν out [6, 28] to get the *marginal likelihood* (also known as *integrated likelihood*) and using this marginal

likelihood for estimating the structural parameter θ. This is also known in literature as the random effects model. Let us explore if this approach resolves the consistency issues for this model.

After assuming a certain prior on ν's and integrating them out, we get the marginal distribution in terms of θ alone. At this stage, we can use the frequentist approach and maximize this integrated/marginal likelihood to get the marginal maximum likelihood estimate for θ (we shall denote this as MMLE from here onwards). Alternately, we can use some prior on θ and look at the posterior mode (MAP estimate). A fully Bayesian framework would involve introducing priors on both θ and ν's: after integrating the ν's out, we are left with only the posterior on θ and we can estimate θ by its posterior mean. As we gather more data and as $W \to \infty$, the effect of the prior will die out and eventually all these three estimates will converge to the same point. For this discussion, we shall focus on the MMLE and look at its asymptotic properties.

In the following sections we explore the marginal approach using three different types of priors on the incidental parameters – uniform prior in Section 2.3.1, and Gaussian distributions in Section 2.3.3 (conjugate setting) and Section 2.3.2 (non-conjugate prior). The uniform prior gives consistent estimates for this particular problem, but the Gaussian priors do not.

For the cases where the MMLE is not consistent (Sections 2.3.3 and 2.3.2), we explore the hierarchical approach. We try to be agnostic about the prior on ν's; we assume that the hyper-parameter c used to parameterize the prior distribution on the ν's is also a random variable that can be estimated. In the hierarchical Bayesian approach we would put a prior on c and get a posterior mean, but for our discussion we stick to a frequentist approach and infer c using a maximum likelihood estimation along with θ (we call these estimates as the hierarchical maximum likelihood estimates or the HMLE).

To be truly agnostic about the form of the prior on ν's, we can use a non-parametric approach. We shall explore the non-parametric Dirichlet process mixture (DPM) prior on

the incidental parameters for our neuron model in Chapter 6. However, for the Neyman-Scott problem in this chapter, we stick to the domain of parametric models.

2.3.1 Using a (non-informative) uniform prior on ν

A prior often suggested in literature for nuisance parameters is a non-informative uniform prior [6]. The likelihood of the data in terms of θ and ν is given by

$$f_{\nu,\theta}(x) = \prod_{w=1}^{W}\prod_{d=1}^{D}\frac{1}{\sqrt{2\pi\theta}}\exp\left(-\frac{(x_{wd}-\nu_w)^2}{2\theta}\right).$$

Integrating each ν_w over $\mathbb{R}$ with respect to an improper uniform prior (see Appendix B.3.1 for details) gives

$$f_{\theta}^{marginal}(x) \;=\; \prod_{w=1}^{W}\frac{1}{(2\pi\theta)^{\frac{D-1}{2}}\sqrt{D}}\exp\left(-\frac{1}{2\theta}\sum_{d}(x_{wd}-s_w)^2\right),$$

where $s_w \triangleq \sum_d x_{wd}$. Since we get this likelihood over θ by integrating the ν's out, this is known as the integrated likelihood[6] or the marginal likelihood. Now we maximize this marginal likelihood to get the marginal maximum likelihood estimate (MMLE) given by Equation (B.10) as

$$\hat{\theta}^{MMLE,unif} = \frac{1}{W(D-1)}\sum_{w=1}^{W}\left(\sum_{d}(x_{wd}-s_w)^2\right) = \frac{DM_2-M_1}{D-1}.$$

Note that this is the same as the CMLE in Equation (2.3) and hence if the data is generated from the true model (2.2), we get

$$\hat{\theta}^{MMLE,unif} = \frac{DM_2-M_1}{D-1} \longrightarrow \theta^*.$$

2.3.2 Using a (non-conjugate) Gaussian prior on ν

Now let us try to integrate out ν with another prior:

$$\nu_w \mid c \sim \mathcal{N}(0,c) \quad \forall\ w.$$

Our original model was $X_{wd} \mid \nu, \theta \sim \mathcal{N}(\nu_w,\theta)\ \forall\ d,\ \forall\ w$. Then we can write

$$X_{wd} = \nu_w + Z_{wd} \quad \forall d,\ \forall\ w,$$

where each Z_{wd} iid $\sim \mathcal{N}(0,\theta)\ \forall w,\ \forall d$ and the Z's are independent from the ν's. Hence $\forall$w

$$\begin{aligned}
\mathbb{E}[X_w] &= \mathbb{E}\nu_w + \mathbb{E}Z_{wd} = 0 \\
\mathbb{E}[X_{wd}^2] &= \mathbb{E}\nu_w^2 + \mathbb{E}Z_{wd}^2 + \mathbb{E}\nu_w Z_{wd} = c + \theta \\
\mathbb{E}[X_{wd_1}X_{wd_2}] &= \mathbb{E}\nu_w^2 + \mathbb{E}\nu_w Z_{wd_1} + \mathbb{E}\nu_w Z_{wd_2} + \mathbb{E}Z_{wd_1}Z_{wd_2} = c \qquad \forall\ d_1 \neq d_2,
\end{aligned}$$

and after integrating the ν's out with the above prior, we get the integrated model

$$X_w \mid (\theta, c) \sim \mathcal{N}\left(\vec{0}, Q\right) \quad \forall\ w = 1, \ldots, W, \tag{2.4}$$

where

$$X_w = \begin{bmatrix} X_{w1} \\ X_{w2} \\ \vdots \\ X_{wD} \end{bmatrix}, \qquad Q = \begin{bmatrix} \theta + c & c & \ldots & c \\ c & \theta + c & & \vdots \\ \vdots & & \ddots & c \\ c & \ldots & c & \theta + c \end{bmatrix}.$$

Log-marginal-likelihood is given by

$$\mathcal{L}_c^{\text{marginal}}(\theta) = -\frac{WD}{2}\log 2\pi - \frac{W}{2}\log|Q| - \frac{1}{2}\sum_{w=1}^{W} x_w^T Q^{-1} x_w,$$

which can be simplified to (see derivation of Equation (B.13) in Appendix B.3.2)

$$-\frac{2}{W}\mathcal{L}_c^{\text{marginal}}(\theta) \quad = \quad D\log 2\pi + (D-1)\log\theta + \log(\theta + cD) + \frac{M_2 D}{\theta} - \frac{M_1 cD}{\theta(\theta + cD)}, \quad (2.5)$$

where M_1 and M_2 are as defined in Equations (2.1) above.

Let the true prior on ν_w be $\nu_w \sim \mathcal{N}(0, c^*) \quad \forall\ w$. The true distribution of the data is then given by

$$X_w \mid (\theta^*, c^*) \ \sim \ \mathcal{N}\left(\vec{0}, Q^*\right) \forall\ w, \quad (2.6)$$

where

$$Q^* \quad = \quad \begin{bmatrix} \theta^* + c^* & c^* & \dots & c^* \\ c^* & \theta^* + c^* & & \vdots \\ \vdots & & \ddots & c^* \\ c^* & \dots & c^* & \theta^* + c^* \end{bmatrix}.$$

This gives $\mathbb{E}[X_w X_w^T] = Q^*$, $\mathbb{E}[X_{wd}^2] = \theta^* + c^* \ \forall\ w, d$ and $\mathbb{E}[X_{wd_1} X_{wd_2}] = c^* \ \forall\ w, d_1 \neq d_2$. Using this we get (see Equation (B.18) for more details) $M_1 \to \theta^* + c^* D$ and $M_2 \to \theta^* + c^*$

MMLE

If we assume c to be a fixed constant, the maximum marginal likelihood estimate (MMLE) is given by maximizing the marginal log-likelihood (2.5) over $\theta \in \mathbb{R}$ for that fixed c. If the data is generated from the true model (2.6), we get $\hat{\theta}^{MMLE} \to \theta^*$ if and only if $c = c^*$, i.e. we know the prior distribution on the ν's exactly (see Appendix B.3.2 for more details). Thus the MMLE need not be consistent in general.

HMLE

Instead of assuming c is a fixed constant, the hierarchical Bayesian approach treats c as random variable with some distribution. Using some priors on c and θ, one can obtain the joint posterior, and the hierarchical Bayesian estimates are given by the posterior mean. However, for this discussion, we use a frequentist approach instead of a Bayesian one – we estimate c jointly along with θ using maximum likelihood estimation and we call this θ estimate the hierarchical maximum likelihood estimate (HMLE). As we get more and more data, the effect of the prior (on θ and c) will eventually die out and the hierarchical Bayesian estimates given by the posterior mean will converge to the HMLE.

Treating the hyper-parameter c as unknown and maximizing Equation (2.5) jointly over c and θ simultaneously gives

$$\hat{c} = \frac{M_1 - M_2}{(D-1)} \qquad \text{and} \qquad \hat{\theta}^{HMLE} = \frac{DM_2 - M_1}{(D-1)}$$

and if the data is generated from the true model (2.6) we get $\hat{\theta}^{HMLE} \to \theta^*$. Hence, $\hat{\theta}^{HMLE}$ is consistent. Note that the HMLE in this case is the same as the CMLE from Equation (2.3).

2.3.3 Using a conjugate prior on (ν, θ)

The conjugate prior setting for (ν, θ) is as follows:

$$\begin{aligned}
\theta \mid (a,b) &\sim \Gamma^{-1}(a,b) \\
\nu_w \mid (\theta, c) &\sim \mathcal{N}(0, \theta c) \quad \forall\ w \\
\Pi(\nu, \theta) &= \frac{b^a}{\Gamma(a)} \left(\frac{1}{\theta}\right)^{a+1} \exp\left(-\frac{b}{\theta}\right) \prod_{w=1}^{W} \left(\frac{1}{\sqrt{2\pi c\theta}} \exp\left(-\frac{\nu_w^2}{2c\theta}\right)\right).
\end{aligned}$$

Using this conjugate prior on ν we get the model similar to equation (2.4)

$$X_w = \ | \ (\theta, c) \sim \mathcal{N}\left(\vec{0},\ \theta R\right) \quad \forall\ w = 1, \ldots, W,$$

where

$$R = \begin{bmatrix} 1+c & c & \ldots & c \\ c & 1+c & & \vdots \\ \vdots & & \ddots & c \\ c & \ldots & c & 1+c \end{bmatrix}.$$

Log-likelihood is given by

$$\mathcal{L}^{\text{marginal}}(\theta, R) \quad = \quad -\left(\frac{WD}{2}\log 2\pi + \frac{W}{2}\log|R| + \frac{WD}{2}\log\theta + \frac{1}{2\theta}\sum_{w=1}^{W} x_w^T R^{-1} x_w\right),$$

which can be simplified to (see derivation of Equation (B.21) in Appendix B.3.3)

$$-\frac{2}{W}\mathcal{L}^{\text{marginal}}(\theta, c) = D\log 2\pi + \log(1+cD) + D\log\theta + \frac{D}{\theta}\left(M_2 - \frac{c}{1+cD}M_1\right), \quad (2.7)$$

where M_1 and M_2 are as defined before in Equations (2.1).

Let the true prior on ν_w's have hyper-parameter c^* and let the true structural parameter be θ^*. Then the true distribution of the data is

$$X_w|(\theta^*, c^*) \ \sim \ \mathcal{N}\left(\vec{0},\ \theta^* R^*\right) \quad \forall\ w, \quad (2.8)$$

where

$$R^* = \begin{bmatrix} 1+c^* & c^* & \ldots & c^* \\ c^* & 1+c^* & & \vdots \\ \vdots & & \ddots & c^* \\ c^* & \ldots & c^* & 1+c^* \end{bmatrix}.$$

This gives $\mathbb{E}[X_w X_w^T] = \theta^* R^*$, $\mathbb{E}[X_{wd}^2] = \theta^*(1+c^*)\ \forall\ w, d$ and $\mathbb{E}[X_{wd_1} X_{wd_2}] = \theta^* c^*\ \forall\ w, d_1 \neq d_2$. Using this we get (see Equation (B.23)) $M_1 \to \theta^*(1 + c^* D)$ and $M_2 \to \quad \theta^*(1 + c^*)$.

MMLE

We can treat c as known and maximize the log marginal likelihood (2.7) over θ to get

$$\hat{\theta}^{MMLE} = M_2 - \frac{c}{1+cD} M_1 \longrightarrow \theta^* \left(1 + \frac{c^* - c}{1 + cD}\right)$$

if the data is generated from the true model (2.8). We will get $\hat{\theta}^{MMLE} \to \theta^*$ if and only if $c = c^*$, i.e. our prior assumption is correct.

HMLE

Treating the hyper-parameter c as unknown and minimizing the negative marginal log-likelihood Equation (2.7) jointly over c and θ simultaneously gives

$$\hat{c} = \frac{M_1 - M_2}{DM_2 - M_1} \qquad \text{and} \qquad \hat{\theta}^{HMLE} = \frac{DM_2 - M_1}{(D-1)}$$

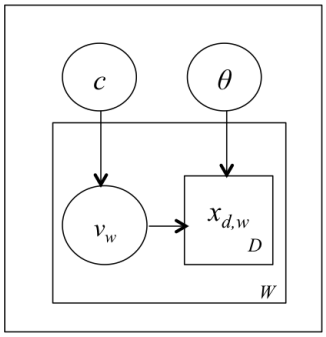
and if the data is generated from the true model (2.8), we get $\hat{\theta}^{HMLE} \to \theta *$. The HMLE with conjugate prior is also the same as the CMLE from Equation (B.8).

In Section 2.2 we assume that the data was sampled from the model (2.2) with true parameters $\nu^* = (\nu_1^*, \ldots, \nu_W^*)$ and θ^*, and the CMLE converges to the true value θ^*. In Sections 2.3.2 and 2.3.3, we assume that the data was generated from models (2.6) and (2.6), with true ν's drawn from Gaussian distributions $\mathcal{N}(0, c^*)$ and $\mathcal{N}(0, c^*\theta^*)$ and respectively. The formulae for HMLE in these sections is the same as that of the CMLE from Section 2.2, and they also converge to the true θ^*. Thus under all three different assumptions (2.2), (2.6) or (2.8), the CMLE (and the HMLE) is consistent.

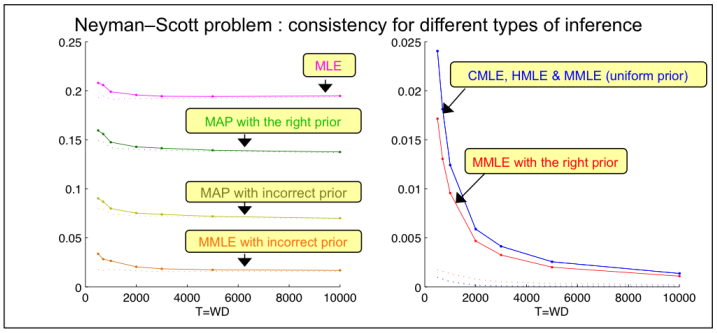
2.4 Summarizing the Neyman-Scott problem for the Gaussian model

We can summarize the results of Chapter 2 as follows

1. MLE is inconsistent due to the incidental parameter problem.
2. Marginal MLE (MMLE) obtained by integrating out ν using a non-conjugate prior $\nu_w \sim N(0, c)\ \forall\ w$ or a conjugate prior $\nu_w \sim N(0, \theta c)\ \forall\ w$ is consistent only if the prior is accurate.
3. CMLE, HMLE (also estimate hyper-parameter c), and MMLE with non-informative uniform prior are all consistent.



(a) Plate notation



(b) Simulation results for the case in Section 2.3.2 using a non-conjugate Gaussian prior on ν's: MSE (solid) and squared bias (dotted)

Figure 2.1: Neyman-Scott problem in a Gaussian model with window width D

Figure 2.1 shows the plate notation for this type of problem on the left. On the right we have results on simulated data with non-stationary ν's. Solid lines represent the mean squared errors (MSE) between estimated $\hat{\theta}$ and true θ^* and dotted lines represent the squared bias. We see that the non-stationary MLE, MAP (posterior mode) estimate, as well as the MMLE with an incorrect prior appear to be inconsistent, whereas the CMLE, HMLE, MMLE with the right prior or non-informative, uniform prior appear to be consistent in simulations.

In the next chapter we introduce the two-neuron log-linear model for zero-lag synchrony. We shall explore how the above techniques for handling incidental parameter problems can be extended to that model. We explore the conditional approach in the next three chapters and the marginal approach in the final chapter.

Chapter 3

Two-neuron log-linear model for zero-lag synchrony

We now introduce the two-neuron log-linear model for zero-lag synchrony. We explore how the stationary assumption can lead to model misspecification (Section 3.1), whereas the non-stationary model can lead to incidental parameter problems (Section 3.2). We then introduce the idea of conditional inference in Section 3.3 for this model and demonstrate, in simulation, how conditioning can overcome the incidental parameter problem and lead to consistent estimates. We also look at the asymptotics of these three estimates for special cases in Sections 3.4 and 3.5. We shall investigate the conditional inference in greater depth in Chapter 4. We shall explore the marginal approach later in Chapter 6.

Basic two-neuron log-linear model for zero-lag synchrony

The data consists of time-discretized spike trains where each entry in the data gives a count of the number of times a neuron cell spikes in the given time bin. If the time scale is fine enough, we can see a maximum of one spike in each time bin. If we have data for T time bins of 1 ms each, we get a $2 \times T$ matrix $X = (X_{it} : i = 1, 2;\ t = 1, \ldots, T) \in \mathcal{X} = \{0, 1\}^{2 \times T}$

consisting of binary random variables X_{it} that indicate whether the neuron i spiked at any time t. Let $x \in \mathcal{X} = \{0,1\}^{2\times T}$ be the observed data. The full log-linear model is given by

$$\mathbb{P}(X = x) = \prod_{t=1}^{T} \frac{e^{(\mu_{1t}x_{1t}+\mu_{2t}x_{2t}+\theta x_{1t}x_{2t})}}{1 + e^{\mu_{1t}} + e^{\mu_{2t}} + e^{(\mu_{1t}+\mu_{2t}+\theta)}}. \tag{3.1}$$

Here θ is the structural parameter we are interested in, which gives time-homogeneous connectivity between the spiking of the two neurons. We are only interested in the parameter θ and want to be robust to the background effects. Parameters $\mu_1 = (\mu_{1t})_{t=1,\ldots,T}$ and $\mu_2 = (\mu_{2t})_{t=1,\ldots,T}$ are the nuisance parameters that account for the variation in individual neuron firing rates due to unobserved background effects.

This model with continuously changing μ's is not identifiable. We need to make additional assumptions on the structure of the nuisance parameters μ's, or use a method that avoids inferring the μ's altogether. We can use domain knowledge to introduce some assumptions about the slowly varying background effects, such as assuming the μ's are stationary (explored in Section 3.1), or piecewise constant (explored in Section 3.2), etc. We can then infer maximum likelihood estimates (MLE) for the nuisance parameters and substitute these in the log-likelihood to estimate the structural parameters. This is the same as trying to get a joint MLE for all the parameters together. The stationary assumption can lead to model misspecification, whereas the piece-wise constant assumption leads to incidental parameter problems. Regularization, i.e. introducing a prior distribution on the nuisance parameters and maximizing the posterior to get the maximum a posteriori (MAP) estimate, does not solve the issue either (Section 3.2).

The two approaches suggested in literature to tackle the incidental parameter problems are conditioning and marginalizing.

We focus mainly on the conditional approach: conditioning on coarsened data so as to be robust to the varying background effects. We shall first study this approach for the two neuron case, where we shall compare the conditional approach to other approaches

and explore the consistency of each method. We shall later extend this method for larger networks and also devise some approximations that can scale with growing number of neurons in Chapter 5.

In the marginal approach, we integrate the nuisance parameters out to get the marginal likelihood in terms of the parameters of interest alone. Once we get the marginal likelihood in terms of θ alone, we can either use the maximum likelihood approach to get the MMLE or put a prior on θ to get the posterior mean. Note that this is different from the MAP estimation, which involves getting the posterior mode for the full set of parameters - the structural and the incidental. As we saw in Chapter 2, we need to accurately identify the true prior on the incidental parameters for the marginalization approach to work. We shall later explore the Bayesian non-parametric approach to achieve this using the Dirichlet process mixture (DPM) prior in Chapter 6.

3.1 Stationary model: model misspecification

The stationary model assumes that the firing rates are constant over time. Let $\mu_{1t} = \nu_1,\ \mu_{2t} = \nu_2\ \forall\ t = 1, \ldots, T$. The stationary model is then given by

$$p(x|\theta, \nu_1, \nu_2) = \prod_{t=1}^{T} \frac{e^{(\nu_1 x_{1t} + \nu_2 x_{2t} + \theta x_{1t} x_{2t})}}{1 + e^{\nu_1} + e^{\nu_2} + e^{(\nu_1+\nu_2+\theta)}}. \tag{3.2}$$

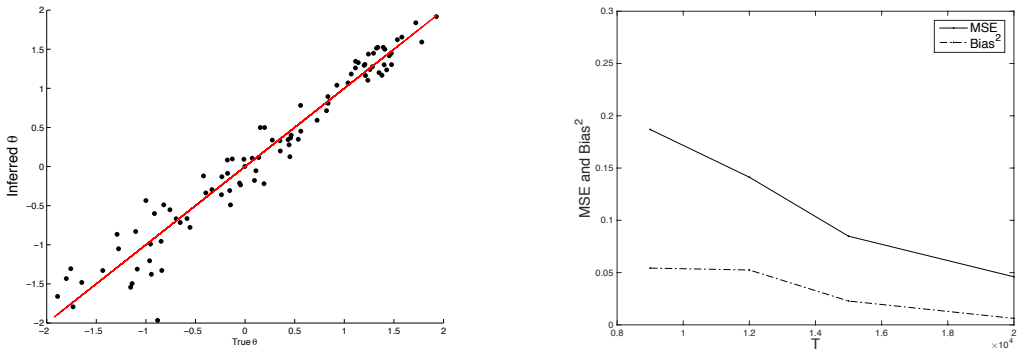
The log-likelihood of the observed data is given by

$$\begin{aligned}
\mathcal{L}(\theta, \nu_1, \nu_2) &= \log p(x|\theta, \nu_1, \nu_2) \\
&= \nu_1 \sum_{t=1}^{T} x_{1t} + \nu_2 \sum_{t=1}^{T} x_{2t} + \theta \sum_{t=1}^{T} x_{1t} x_{2t} - T \log\left(1 + e^{\nu_1} + e^{\nu_2} + e^{(\nu_1+\nu_2+\theta)}\right) \\
&= \nu_1 S_1(x) + \nu_2 S_2(x) + \theta L(x) - T \log\left(1 + e^{\nu_1} + e^{\nu_2} + e^{(\nu_1+\nu_2+\theta)}\right) \\
&= \nu_1 s_1 + \nu_2 s_2 + \theta l - T \log\left(1 + e^{\nu_1} + e^{\nu_2} + e^{(\nu_1+\nu_2+\theta)}\right),
\end{aligned}$$

where the sufficient statistics for the data are $S_1(x) = \sum x_{1t} = s_1$, $S_2(x) = \sum x_{2t} = s_2$ and $L(x) = \sum x_{1t}x_{2t} = l$ and the maximum likelihood estimator (MLE) for this model is given by

$$\begin{aligned} \hat{\nu}_1 &= \log \frac{s_1 - l}{T - s_1 - s_2 + l} \\ \hat{\nu}_2 &= \log \frac{s_2 - l}{T - s_1 - s_2 + l} \\ \hat{\theta} &= \log \frac{l\,(T - s_1 - s_2 + l)}{(s_1 - l)\,(s_2 - l)}. \end{aligned} \tag{3.3}$$

If the data were to come from an experiment where the firing rates are held constant, then the model gives accurate and consistent results, as seen in Figure 3.1. Here we simulated data from the stationary model (3.2) with true parameters $\nu_1^* = -2.0842$ and $\nu_2^* = -2.1952$ and $\theta^* \sim \text{Unif}(-2, 2)$. Using MLE for the stationary model gives consistent results since we use the accurate model.



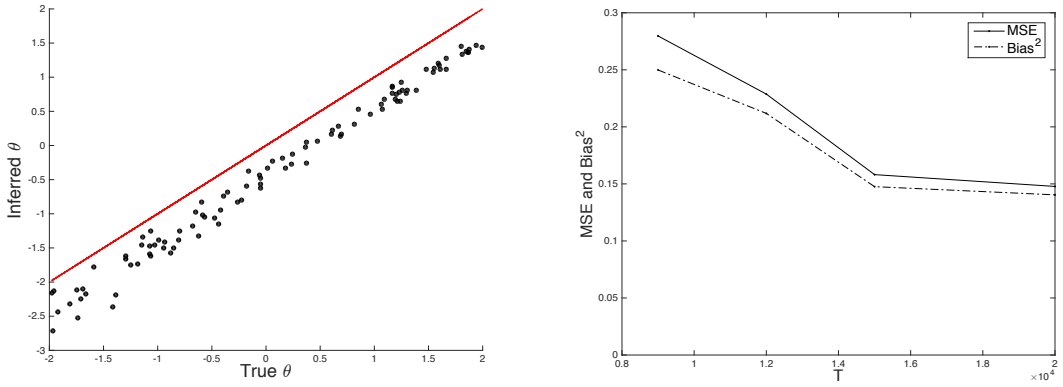
(a) Scatter plot of $\hat{\theta}$ versus θ^* **(b)** MSE (solid) and squared bias (dotted)

Figure 3.1: Stationary MLE on stationary data: On the left we have inferred $\hat{\theta}$ (stationary MLE) versus true θ for various simulated trials with $T = 12000$ ms. On the right we have mean square error and squared bias plots averaged over multiple simulations for varying values of T. Both the MSE and bias^2 tend to 0 as T increases and the MLE is consistent. (Data generated from the stationary model (3.2) with true parameters $\nu_1^* = -2.0842$ and $\nu_2^* = -2.1952$ and $\theta^* \sim \text{Unif}(-2, 2)$)

However, the assumption that the background effects are stationary is unrealistic. In any real experiment, the firing rates vary over time. If the data were to come from a non-stationary model with time-varying background effects, and we infer θ using a stationary model (i.e. assuming that the 'ν's are constant over time), we have a problem of model

misspecification, and we cannot expect to achieve consistency.

We show this on simulated data: we generate data from a non-stationary model (3.1) with slowly varying (sinusoidal) ν_{1t} and ν_{2t} and $\theta^* \sim \text{Unif}(-2, 2)$. Figure 3.2 shows the results for stationary-MLE on this non-stationary data. For this particular dataset, the stationary MLE is lower than the true θ^* for all simulations, and the average mean squared errors and squared bias appears to saturate at a positive value and (does not tend to 0 as T increases).



(a) Scatter plot of $\hat{\theta}$ versus θ^* **(b)** MSE (solid) and squared bias (dotted)

Figure 3.2: Stationary MLE on non-stationary data: On the left we have inferred (stationary MLE) versus true θ for various simulated trials with $T = 12000$ ms. The data is generated from a non-stationary distribution (3.1) with smoothly varying ν's but the inference uses MLE for the model 3.2. On the right we have mean square error (MSE) and bias^2 plots averaged over multiple simulations for varying values of T. The stationary-MLE appears to be inconsistent (because of model misspecification).

3.2 Non-stationary model: incidental parameter problem

As seen earlier, the stationary model (3.2) is misspecified and the full log-linear model (3.1) with time-varying firing rates is not identifiable. One way to get around the identifiability issue is to introduce some assumptions on the structure of the firing rates, based on some scientific knowledge in this domain. The firing rates vary slowly over the timescales we are interested in. We can assume μ's in Equation (3.1) to be piecewise constant in small windows of width 4 or 5 ms, which will make the model identifiable.

Let us take the window width to be $D = 4$ ms. For simplicity we assume that T is exactly divisible by D and the number of windows is given by $W = T/D$. Let Ω_w represent the w-th bin: $\Omega_w = \{t : (w-1)D < t \leq wD\}$ and $\mu_{1t} = \nu_{1w},\ \mu_{2t} = \nu_{2w}\ \forall\ t \in \Omega_w$ for each $w = 1, 2, \ldots, W$. The nuisance parameters for this piecewise constant model are $\vec{\nu}_1 = (\nu_{1w} : w = 1, \ldots, W)$ and $\vec{\nu}_2 = (\nu_{2w} : w = 1, \ldots, W)$. The non-stationary piecewise constant model is then given by

$$p(x|\theta, \vec{\nu}_1, \vec{\nu}_2) = \prod_{w=1}^{W} \prod_{t \in \Omega_w} \frac{\exp(\nu_{1w} x_{1t} + \nu_{2w} x_{2t} + \theta x_{1t} x_{2t})}{\left(1 + e^{\nu_{1w}} + e^{\nu_{2w}} + e^{(\nu_{1w}+\nu_{2w}+\theta)}\right)^D}. \tag{3.4}$$

The sufficient statistics for the data are $S_{1w}(X) = \sum_{t \in \Omega_w} X_{1t},\ S_{2w}(X) = \sum_{t \in \Omega_w} X_{2t}$ in each window, and $L(X) = \sum_w L_w(X)$ with $L_w(X) = \sum_{t \in \Omega_w} X_{1t} X_{2t}\ \forall\ w$. Let the observed statistics be $s_{1w} = S_{1w}(x),\ s_{2w} = S_{2w}(x),\ l_w = L_w(x)\ \forall\ w$ and $l = L(x) = \sum l_w$.

The likelihood can be then written as:

$$\begin{aligned} p(x|\theta, \vec{\nu}_1, \vec{\nu}_2) &= \prod_{w=1}^{W} \frac{\exp\left(\nu_{1w} S_{1w}(x) + \nu_{2w} S_{2w}(x) + \theta L_w(x)\right)}{\left(1 + e^{\nu_{1w}} + e^{\nu_{2w}} + e^{(\nu_{1w}+\nu_{2w}+\theta)}\right)^D} \\ &= \prod_{w=1}^{W} \frac{\exp\left(\nu_{1w} s_{1w} + \nu_{2w} s_{2w} + \theta l_w\right)}{\left(1 + e^{\nu_{1w}} + e^{\nu_{2w}} + e^{(\nu_{1w}+\nu_{2w}+\theta)}\right)^D}. \end{aligned}$$

The log-likelihood is:

$$\begin{aligned} \mathcal{L}(\theta, \vec{\nu}_1, \vec{\nu}_2) &= \sum_{w=1}^{W} \left(\nu_{1w} s_{1w} + \nu_{2w} s_{2w} + \theta l_w - D \log\left(1 + e^{\nu_{1w}} + e^{\nu_{2w}} + e^{(\nu_{1w}+\nu_{2w}+\theta)}\right)\right) \\ &= \sum_{w=1}^{W} \mathcal{L}_w(\theta, \nu_{1w}, \nu_{2w}), \end{aligned}$$

where

$$\mathcal{L}_w(\theta, \nu_{1w}, \nu_{2w}) = \nu_{1w} s_{1w} + \nu_{2w} s_{2w} + \theta l_w - D \log\left(1 + e^{\nu_{1w}} + e^{\nu_{2w}} + e^{(\nu_{1w}+\nu_{2w}+\theta)}\right). \tag{3.5}$$

The non-stationary MLE for θ is given by maximizing this log-likelihood

$$\begin{aligned}\hat{\theta}^{\text{MLE}} &= \arg\max_{\theta\in\mathbb{R}} \mathcal{L}(\theta, \vec{\nu}_1, \vec{\nu}_2) \\ &= \arg\max_{\theta\in\mathbb{R}} \sum_{w}^{W} \sup_{\nu_{1w},\nu_{2w}} \mathcal{L}_w(\theta, \nu_{1w}, \nu_{2w}). \end{aligned} \tag{3.6}$$

Proposition 1. *The windows where either $s_{1w} = 0$, $s_{2w} = 0$, $s_{1w} = D$ or $s_{2w} = D$ can be ignored for the MLE calculation in (3.6)*

Proof. Suppose $s_w = (s_{1w}, s_{2w}) = (0, 0)$, which will give $l_w = 0$. Then

$$\begin{aligned}\sup_{\nu_{1w},\nu_{2w}} \mathcal{L}_w(\theta, \nu_{1w}, \nu_{2w}) &= \sup_{\nu_{1w},\nu_{2w}} \left[-D\log\left(1 + e^{\nu_{1w}} + e^{\nu_{2w}} + e^{(\nu_{1w}+\nu_{2w}+\theta)}\right)\right] \\ &= -D\log\left(\inf_{\nu_{1w},\nu_{2w}} \left[1 + e^{\nu_{1w}} + e^{\nu_{2w}} + e^{(\nu_{1w}+\nu_{2w}+\theta)}\right]\right) \\ &= -D\log(1) = 0.\end{aligned}$$

Similarly if $s_w = (D, 0)$ or $(0, D)$ or (D, D), we get $\sup_{\nu_{1w},\nu_{2w}} \mathcal{L}_w(\theta, \nu_{1w}, \nu_{2w}) = 0$.

Suppose $s_w = (0, cD)$ where $0 < c < 1$, then

$$\begin{aligned}\sup_{\nu_{1w},\nu_{2w}} \mathcal{L}_w(\theta, \nu_{1w}, \nu_{2w}) &= \sup_{\nu_{1w},\nu_{2w}} \left[cD\nu_{2w} - D\log\left(1 + e^{\nu_{1w}} + e^{\nu_{2w}} + e^{(\nu_{1w}+\nu_{2w}+\theta)}\right)\right] \\ &= D\sup_{\nu_{2w}} \left[c\nu_{2w} - \log\left(\inf_{\nu_{1w}} \left[1 + e^{\nu_{1w}} + e^{\nu_{2w}} + e^{(\nu_{1w}+\nu_{2w}+\theta)}\right]\right)\right] \\ &= D\sup_{\nu_{2w}} \left[c\nu_{2w} - \log\left(1 + e^{\nu_{2w}}\right)\right] \\ &= D(c\log c + (1-c)\log(1-c)).\end{aligned}$$

Note that there is no θ dependence. Similarly if $s_w = (cD, 0)$ or (D, cD) or (cD, D) for $0 < c < 1$, $\sup_{\nu_{1w},\nu_{2w}} \mathcal{L}_w(\theta, \nu_{1w}, \nu_{2w}) = c\log c + (1-c)\log(1-c)$ is a constant that does not depend on θ.

Thus whenever either of s_{1w} or s_{2w} is 0 or D, $\sup_{\nu_{1w},\nu_{2w}} \mathcal{L}_w(\theta, \nu_{1w}, \nu_{2w})$ is either 0 or a constant in θ and can be ignored in Equation (3.6). □

Let us define $\mathbb{S}_{good}$ to be the set of all possible window sums $n = (n_1, n_2)'$ where the windows are not degenerate, i.e.

$$\begin{aligned}\mathbb{S}_{good} &= \left\{ n = \begin{pmatrix} n_1 \\ n_2 \end{pmatrix} \; : \; n_1 \neq 0, \; n_2 \neq 0, \; n_1 \neq D, \; n_2 \neq D \right\} \\ &= \left\{ n = \begin{pmatrix} n_1 \\ n_2 \end{pmatrix} \; : \; n_1, n_2 \in \{1, 2, \ldots, (D-1)\} \right\}. \end{aligned} \tag{3.7}$$

Hence, we can ignore all the windows where $s_w \notin \mathbb{S}_{good}$. For all other non-degenerate windows, $\sup_{\nu_{1w}, \nu_{2w}} L_w(\theta, \nu_{1w}, \nu_{2w})$ is attained and is equal to the maximum. So we can prune the degenerate windows and the $\hat{\theta}^{MLE}$ is given by maximizing

$$\mathcal{L}_{good}(\theta, \vec{\nu}_1, \vec{\nu}_2) \triangleq \sum_{w: s_w \in \mathbb{S}_{good}} \mathcal{L}_w(x|\theta, \nu_1, \nu_2) \tag{3.8}$$

$$\begin{aligned}\hat{\theta}^{\text{ MLE}} &= \arg\max_{\theta \in \mathbb{R}} \max_{\vec{\nu}_1, \vec{\nu}_2} \mathcal{L}_{good}(\theta, \vec{\nu}_1, \vec{\nu}_2) \\ &= \arg\max_{\theta \in \mathbb{R}} \sum_{w: s_w \in \mathbb{S}_{good}} \max_{\nu_{1w}, \nu_{2w}} \mathcal{L}_w(\theta, \nu_{1w}, \nu_{2w}). \end{aligned} \tag{3.9}$$

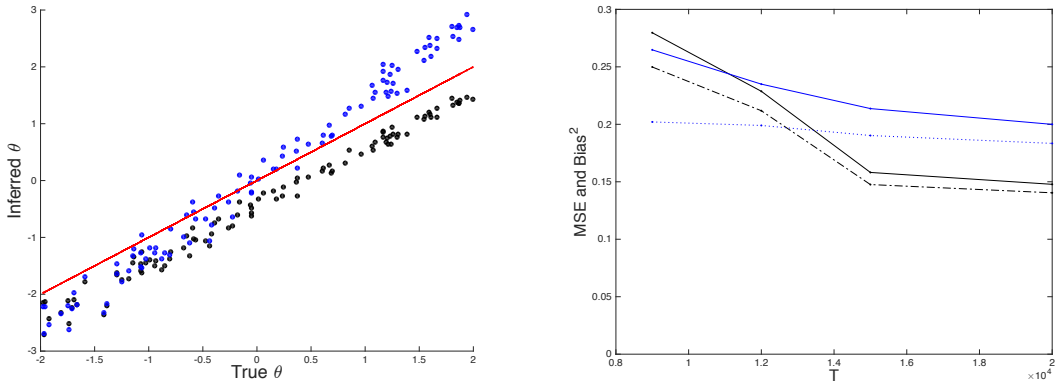
We equate the partial derivatives of $\mathcal{L}_{good}(x|\theta, \vec{\nu}_1, \vec{\nu}_2) \triangleq \sum_{w: s_w \in \mathbb{S}_{good}} \mathcal{L}_w(x|\theta, \nu_1, \nu_2)$ to zero:

$$\frac{s_{1w}}{D} - \frac{e^{\hat{\nu}_{1w}} + e^{\hat{\nu}_{1w} + \hat{\nu}_{2w} + \hat{\theta}}}{1 + e^{\hat{\nu}_{1w}} + e^{\hat{\nu}_{2w}} + e^{\hat{\nu}_{1w} + \hat{\nu}_{2w} + \hat{\theta}}} = 0 \qquad \forall w : s_w \in \mathbb{S}_{good} \tag{3.10}$$

$$\frac{s_{2w}}{D} - \frac{e^{\hat{\nu}_{2w}} + e^{\hat{\nu}_{1w} + \hat{\nu}_{2w} + \hat{\theta}}}{1 + e^{\hat{\nu}_{1w}} + e^{\hat{\nu}_{2w}} + e^{\hat{\nu}_{1w} + \hat{\nu}_{2w} + \hat{\theta}}} = 0 \qquad \forall w : s_w \in \mathbb{S}_{good} \tag{3.11}$$

$$\sum_{w: s_w \in \mathbb{S}_{good}} \left(\frac{l_w}{D} - \frac{e^{\hat{\nu}_{1w} + \hat{\nu}_{2w} + \hat{\theta}}}{1 + e^{\hat{\nu}_{1w}} + e^{\hat{\nu}_{2w}} + e^{\hat{\nu}_{1w} + \hat{\nu}_{2w} + \hat{\theta}}} \right) = 0 \tag{3.12}$$

Solving Equations (3.10) and (3.11) simultaneously will give the MLE for $\hat{\nu}_{1w}$ and $\hat{\nu}_{2w}$ in each window, as a function of θ. These can then be substituted in the likelihood to get the profile likelihood and maximizing the profile likelihood over θ using (3.12) will give $\hat{\theta}^{\text{ MLE}}$. In simulations, $\hat{\theta}^{\text{ MLE}}$ can obtained numerically using a standard optimization solver.



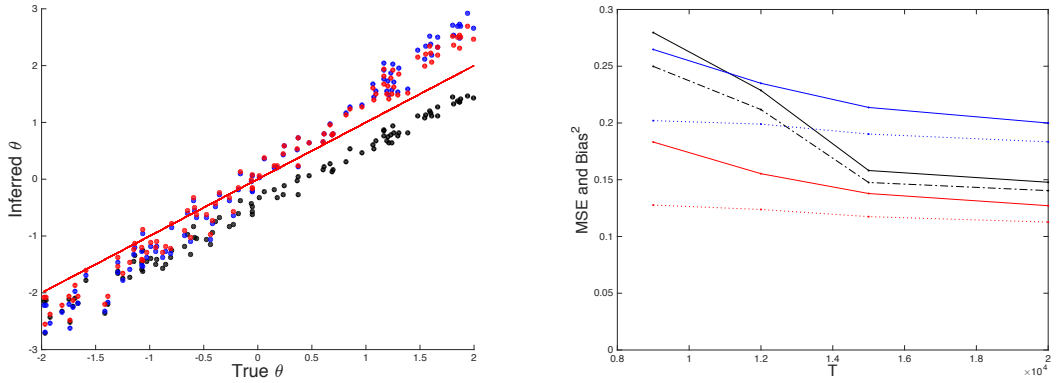
(a) Scatter plot of $\hat{\theta}$ versus θ^* **(b)** MSE (solid) and squared bias (dotted)

Figure 3.3: Non-stationary MLE (blue) and stationary MLE (black) on non-stationary data: On the left we have inferred $\hat{\theta}$ (stationary MLE and non-stationary MLE) versus true θ for various simulated trials with $T = 12000$ ms. On the right we have mean square error and squared bias plots averaged over multiple simulations for varying values of T. Both stationary MLE and the non-stationary MLE appear to be inconsistent

In Figure 3.3 we see how the non-stationary MLE compares with stationary MLE on the same simulated non-stationary data described before. Both stationary and non-stationary MLE results appear to be inconsistent. Stationary MLE gives incorrect results because of model misspecification, whereas the non-stationary MLE suffers from the incidental parameter problem. Size of ν grows as the number of windows grow, hence they are incidental parameters. Each (ν_{1w}, ν_{2w}) pair is inferred from only the corresponding D entries in each window x_w. Since these estimates are used to calculate the MLE for θ, it leads to inconsistent results for $\hat{\theta}^{\text{MLE}}$.

Instead of the Maximum Likelihood Estimation (MLE), we could also use a a suitable prior on the parameters and maximizing the posterior to get the MAP estimate. Figure 3.4 shows the simulation results with the MAP estimate included. The MAP estimates closely resemble the non-stationary MLE and appears to be inconsistent as well. Regularization doesn't help overcome the incidental parameter problem.

We could also integrate out the incidental parameters to get the marginal likelihood as a function of θ alone. We can then either use maximum likelihood estimation on this marginal likelihood to get the MMLE, or use a full Bayesian approach with a prior on θ and sample from the posterior. As we saw in Chapter 2, this approach works if we are able



(a) Scatter plot of $\hat{\theta}$ versus θ^* **(b)** MSE (solid) and squared bias (dotted)

Figure 3.4: Map estimate (red), non-stationary MLE (blue) and stationary MLE (black) on non-stationary data: On the left we have inferred $\hat{\theta}$ versus true θ for various simulated trials with $T = 12000$ ms. On the right we have mean square error and squared bias plots averaged over multiple simulations for varying values of T.

to accurately identify the true prior on the ν's. We can achieve this by using a Bayesian non-parametric approach with Dirichlet process mixture (DPM) prior on the ν's, which we shall explore in Chapter 6. Let us first explore the conditional approach to overcome the incidental parameter problem.

3.3 Conditioning on coarse temporal structure

We are interested in the zero-lag connectivity of the neuron network, and this inference is complicated by the non-stationary background effects. We want to develop an inference procedure that will be robust to these coarse temporal effects, and at the same time be able to capture the network connectivity of finer time scales.

The original data $X \in \mathcal{X} = \{0, 1\}^{2\times T}$ has a time resolution of 1 ms, so that each time-bin of 1 ms has either 1 or 0 spikes. X contains information about fine temporal synchrony, as well as, the coarse temporal background effects.

Let us now take a look at the same data coarsened over slightly larger time-windows. Let S represent the coarsened data which keeps a count of the number of spikes in each coarsened

time-window of width D. Theoretically, each bin can have up to D spikes. For simplicity we assume that T is exactly divisible by D and the total number of windows in S is given by $W = T/D$. If $\Omega_w = \{t : (w-1)D < t \leq wD\}$ represents the w-th window, the coarse data is given by the statistic $S(X) = (S_{iw}(X) : i = 1, 2;\ w = 1, \ldots, W) \in \{0, \ldots, D\}^{2\times W}$ where $S_{iw}(X) = \sum_{t\in\Omega_w} X_{it}$ as defined in the previous section (see Figure 3.5). $S(X)$ encodes some of the time-varying effects, however it contains little information about the fine-temporal connectivities since now the data is aggregated over wider windows of D ms.

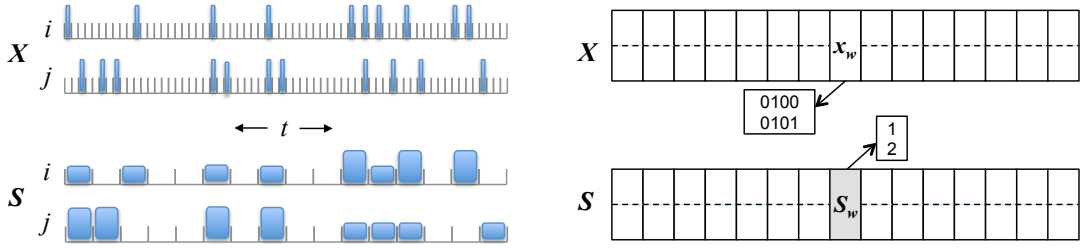


Figure 3.5: X is the original spiking data and S is its temporal coarsening using larger bins. We model the conditional distribution of the precise spike timing, given the coarse data in S.

So far we have been modeling the probability of the entire data $\mathbb{P}(X = x)$, which can be split as follows:

$$\begin{aligned} \mathbb{P}(X = x) &= \mathbb{P}\left(S(X) = S(x)\right) \cdot \mathbb{P}\left(X = x | S(X) = S(x)\right) \\ p(x) &= g(S(x)) \cdot \psi\left(x | S(x)\right) \end{aligned} \tag{3.13}$$

Instead of modeling all X, including S, we propose to focus inference on the conditional distribution $\psi(x|S(x))$ of precise spike timing given coarse spike timing. This conditional distribution contains the signal of interest about synchrony, while being completely robust to the dynamics on coarser time scales.

For ease of notation, we also introduce some additional notation. Let $X_w = (X_{it} : i = 1, 2;\ t \in \Omega_w) \in \{0, 1\}^{2\times D}$ represent the data in the window Ω_w. Hence the full data is $X = (X_1, \ldots, X_w)$. Let us define statistics on $X_w \in \{0, 1\}^{2\times D}$: $S_1(X_w) \triangleq S_{1w}(X) = \sum_{t\in\Omega_w} X_{1t}$, $S_2(X_w) \triangleq S_{2w}(X) = \sum_{t\in\Omega_w} X_{2t}$. Let $S(X_w)$ be the sum statistics for X_w given by the two-dimensional vector $S(X_w) \triangleq (S_1(X_w), S_2(X_w))' = (S_{1w}(X), S_{2w}(X))'$. We use

a shorthand notations S_{1w} and S_{2w} to represent $S_1(X_w)$ and $S_2(X_w)$ respectively, and $S_w = (S_{1w}, S_{2w})'$ to represent $S(X_w) = (S_1(X_w), S_2(X_w))'$. We can then write the coarse data statistic $S(X)$ by putting all the S_w's together as $S(X) = (S_1, \ldots, S_w, \ldots, S_W)$.

Furthermore, we define observed statistics by their corresponding small letters as follows: If $x \in \mathcal{X} = \{0,1\}^{2\times T}$ is observed data, then s represents the observed sum statistics are $s \triangleq S(x)$. We use $x_w = (x_{it} : i = 1,2;\ t \in \Omega_w) \in \{0,1\}^{2\times D}$ to denote the data in the *w-th* window of x and $s_w = (s_{1w}, s_{2w})' \triangleq (S_1(x_w), S_2(x_w))'$, which gives $s = (s_1, \ldots, , s_w, \ldots, s_W)$.

Let us consider the stationary distribution:

$$\mathbb{P}(X = x) = \frac{1}{Z(\theta, \nu_1, \nu_2)} \prod_{t=1}^{T} e^{\nu_1 x_{1t} + \nu_2 x_{2t} + \theta x_{1t} x_{2t}},$$

where the normalizing constant is $Z(\theta, \nu_1, \nu_2) = (1 + e^{\nu_1} + e^{\nu_2} + e^{(\nu_1+\nu_2+\theta)})^T$. Then,

$$\begin{aligned} \mathbb{P}(X = x) &= \frac{1}{Z(\theta, \nu_1, \nu_2)} \prod_{w=1}^{W} \prod_{t\in\Omega_w} e^{\nu_1 x_{1t} + \nu_2 x_{2t} + \theta x_{1t} x_{2t}} \\ &= \frac{1}{Z(\theta, \nu_1, \nu_2)} \prod_{w=1}^{W} e^{\nu_1 S_1(x_w) + \nu_2 S_2(x_w) + \theta L(x_w)}, \end{aligned}$$

where $S_1(x_w) = \sum_{t\in\Omega_w} x_{1t}$, $S_2(x_w) = \sum_{t\in\Omega_w} x_{2t}$, as defined above, and $L(x_w) = \sum_{t\in\Omega_w} x_{1t}x_{2t}$ counts the number of synchronies in x_w.

Note that each window is independent and that the joint probability splits into the product of individual window probabilities:

$$\mathbb{P}(X = x) = \prod_{w=1}^{W} \mathbb{P}(X_w = x_w) = \prod_{w=1}^{W} \frac{e^{\nu_1 S_1(x_w) + \nu_2 S_2(x_w) + \theta L(x_w)}}{(1 + e^{\nu_1} + e^{\nu_2} + e^{(\nu_1+\nu_2+\theta)})^D}.$$

Hence the probability of coarse data being the same as the observed coarse data $\mathbb{P}(S(X) = S(x))$ also splits into the product of $\mathbb{P}(S(X_w) = S(x_w))$. For each window, $\mathbb{P}(S(X_w) = S(x_w))$ is given by summing over all possible $y \in \{0,1\}^{2\times D}$ which have $S(y) = S(x_w)$,

i.e.$S_1(y) = S_1(x_w)$ and $S_2(y) = S_2(x_w)$. Hence

$$\begin{aligned}
\mathbb{P}(S(X) = S(x)) &= \prod_w \mathbb{P}(S(X_w) = S(x_w)) \\
&= \prod_w \sum_y \mathbb{1}(S(y) = S(x_w)) \cdot \frac{e^{\nu_1 S_1(y) + \nu_2 S_2(y) + \theta L(y)}}{(1 + e^{\nu_1} + e^{\nu_2} + e^{(\nu_1+\nu_2+\theta)})^D} \\
&= \prod_w \sum_y \mathbb{1}(S(y) = S(x_w)) \cdot \frac{e^{\nu_1 S_1(x_w) + \nu_2 S_2(x_w) + \theta L(y)}}{(1 + e^{\nu_1} + e^{\nu_2} + e^{(\nu_1+\nu_2+\theta)})^D} \\
&= \prod_w \frac{e^{\nu_1 S_1(x_w) + \nu_2 S_2(x_w)}}{(1 + e^{\nu_1} + e^{\nu_2} + e^{(\nu_1+\nu_2+\theta)})^D} \sum_y \mathbb{1}(S(y) = S(x_w)) e^{\theta L(y)} \\
&= \frac{1}{Z(\theta, \nu_1, \nu_2)} \prod_w e^{\nu_1 S_1(x_w) + \nu_2 S_2(x_w)} \sum_{y: S(y) = S(x_w)} e^{\theta L(y)}.
\end{aligned}$$

In the conditional distribution, the terms corresponding to the nuisance parameters ν's cancel out to give

$$\begin{aligned}
\psi\left(x | S(x), \theta\right) &= \mathbb{P}(X = x | S(X) = S(x)) = \frac{\mathbb{P}(X = x)}{\mathbb{P}(S(X) = S(x))} \\
&= \prod_{w=1}^{W} \frac{e^{\theta L(x_w)}}{\sum_{y: S(y) = S(x_w)} e^{\theta L(y)}}.
\end{aligned} \tag{3.14}$$

Instead of maximizing the full likelihood $\mathbb{P}(X = x)$ as done in standard maximum likelihood methods, the conditional maximum likelihood estimate (CMLE) $\hat{\theta}^{\text{CMLE}}$ is given by maximizing this conditional likelihood:

$$\hat{\theta}^{\text{CMLE}} = \arg\max_{\theta} \ \psi\left(x | S(x), \theta\right).$$

Conditional inference focuses on the part of the likelihood that governs the fine-temporal interactions that we are interested in, and avoids modeling the coarser dynamics $g(S(x) = s)$, in effect reducing the errors caused due to inaccurate modeling and inference of the background effects. Thus this inference technique applies to a wider class of models, with different models for $g(s)$ which still result in the same $\psi(x|s)$. For example, the conditional distribution obtained from the non-stationary model (3.4) will also give rise to the same conditional distribution (3.14) which was derived from the stationary model (3.2).

While computing the CMLE, instead of calculating the denominator in (3.14) by summing over all possible $y \in \{0,1\}^{2\times D}$, we can pre-calculate the coefficients:

$$\begin{aligned}\sum_{y:S_w(y)=s_w} e^{\theta L(y)} &= \sum_{k=0}^{D} e^{\theta k} \cdot \#\left\{y \in \{0,1\}^{2\times D} : S(y) = S(x_w)\ \&\ L(y) = k\right\} \\ &= \sum_{l=0}^{D} e^{\theta k} \cdot h_{S(x_w)}(k),\end{aligned}$$

where the coefficient $h_{S(x_w)}(k)$ is the hypergeometric coefficient which counts the number of $2 \times D$ binary matrices with k number of synchronies and row sums given by $S(x_w) = (S_1(X_w), S_2(x_w))'$. If $n = (n_1, n_2)'$, where n_1 and n_2 are integers, the hypergeometric coefficient is given by

$$h_n(k) = \begin{pmatrix} D \\ n_1 \end{pmatrix} \begin{pmatrix} n_1 \\ k \end{pmatrix} \begin{pmatrix} D - n_1 \\ n_2 - k \end{pmatrix}, \tag{3.15}$$

where $\begin{pmatrix} a \\ b \end{pmatrix} = \dfrac{a!}{b!(a-b)!}$ if $0 \le b \le a$, and 0 otherwise.

Using these coefficients we get

$$\psi\left(x|S(x),\theta\right) = \prod_w \frac{e^{\theta L(x_w)}}{\sum_{y:S_w(y)=S(x_w)} e^{\theta L(y)}} = \prod_w \frac{e^{\theta L(x_w)}}{\sum_{k=0}^{D} e^{\theta k} h_{s_w}(k)} = \prod_w p_w(x_w|\theta), \tag{3.16}$$

where

$$p_w(x_w|\theta) = \frac{e^{\theta L(x_w)}}{\sum_{y:S_w(y)=S(x_w)} e^{\theta L(y)}} = \frac{e^{\theta L(x_w)}}{\sum_{k=0}^{D} e^{\theta k} \cdot h_{S(x_w)}(k)} \tag{3.17}$$

and the CMLE is

$$\hat{\theta}^{\text{CMLE}} = \arg\max_{\theta}\ \psi\left(x|S(x)=s\right) = \arg\min_{\theta} \sum_w \phi_w(X_w|\theta), \tag{3.18}$$

where $\phi_w(x_w|\theta) \triangleq -\log p_w(x_w|\theta)\ \forall\ w$ is the negative log-likelihood for each window. Since (3.16) is an exponential family, the function to be minimized is convex (see proof of Theorem 1 for more details).

Proposition 2. *Let $\mathbb{S}_{good}$ be as defined in (3.7). We can ignore degenerate windows where $s_w \notin \mathbb{S}_{good}$ in the CMLE computation and the CMLE can be defined as*

$$\hat{\theta}^{CMLE} = \arg\min_{\theta} \sum_{w:s_w \in \mathbb{S}_{good}} \phi_w(x_w|\theta). \tag{3.19}$$

Proof. If either $s_{1w} = 0$ or $s_{2w} = 0$, we would have $l_w = 0$ and also $h_{s_w}(k) = 0 \ \forall \ k \neq 0$. Hence $p_w(\cdot|\theta)$ would be a constant function in θ (in fact, $p_w(\cdot|\theta) = 1$). Hence for all practical purposes, we can ignore these degenerate windows from our calculations. The same applies to windows where at least one of s_{1w} or s_{2w} is equal to D. Without loss of generality, let $s_{1w} = D$ and $0 < s_{2w} = c < D$. Then we have $l_w = c$ and also $h_{s_w}(k) = 0 \ \forall \ k \neq c$. So $p_w(\cdot|\theta)$ has a term $e^{\theta c}$ in both the numerator and denominator, which cancels out to give a constant term. So we can also ignore windows where either of the spike count sums is 0 or D. Ignoring the degenerate windows, we get

$$\psi\left(x|S(x) = s\right) \ \propto \ \psi_{\text{good}}\left(x|S(x) = s\right) \ = \prod_{w:S(x_w) \in \mathbb{S}_{good}} p_w(x_w|\theta). \tag{3.20}$$

We can now maximize this modified conditional likelihood $\psi_{\text{good}}\left(x|S(x) = s\right)$ to get the conditional maximum likelihood estimate (CMLE):

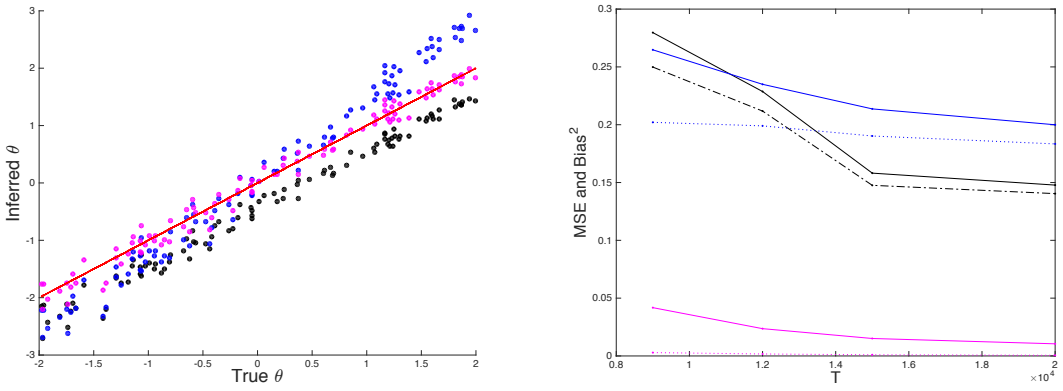
$$\hat{\theta}^{\text{CMLE}} = \arg\max_{\theta} \ \psi_{good}\left(x|S(x) = s\right) \ = \arg\min_{\theta} \sum_{w:s_w \in \mathbb{S}_{good}} \phi_w(x_w|\theta), \tag{3.21}$$

where $\phi_w(x_w|\theta) \triangleq -\log p_w(x_w|\theta) \ \forall \ w$ □

Each different D will give a slightly different θ. The choice of D depends on what time scales we want to be robust to. If the background rates are suspected to change rapidly, we will choose a small D, say 4-5 ms, but if we think the background effects vary slowly, we will choose a larger D, say 10 ms, or 50 ms, etc. We look at some of these effects of changing D in simulated data in Section 4.2.

Figure 3.6 shows the CMLE results superimposed on standard MLE results (stationary

MLE and non-stationary MLE) on the same simulated non-stationary data from before. We can see how the CMLE captures the true θ^* more accurately as compared to the other MLE estimates, and that the mean squared error and the squared bias appear to tend to 0 as T increases.



(a) Scatter plot of $\hat{\theta}$ versus θ^* **(b)** MSE (solid) and squared bias (dotted)

Figure 3.6: Conditional inference on non-stationary data: stationary MLE (black), non-stationary MLE (blue) and CMLE (pink). On the left we have inferred $\hat{\theta}$ versus true θ for various simulated trials with $T = 12000$ ms. On the right we have mean square error and squared bias plots averaged over multiple simulations for varying values of T. Both stationary MLE and non-stationary MLE appear to be inconsistent, while CMLE appears to be consistent.

In the next section we get analytic expressions for these different estimates (stationary MLE, non-stationary MLE and conditional MLE) with $D = 2$ or $D = 3$ and explore their asymptotic behavior. If the true data is generated from the stationary model or the piecewise constant non-stationary model, we see that only the CMLE converges to the true θ^*, whereas both stationary MLE and non-stationary MLE are inconsistent. We shall see a more general proof for consistency of the CMLE in Chapter 4 and also obtain confidence intervals.

Computing the normalizing constant in (3.17) involves summing over all possible combinations of spikes with the same coarse temporal structure. This can get computationally intensive for larger networks. As we begin to look at larger networks, we need to look at different types of approximations for the conditional likelihood. We explore one such approximation – the composite likelihood (also knows as the pseudo-likelihood) in Section 5.2.

3.4 Asymptotic results for window width $D = 2$

We look at some asymptotic results, both consistency and asymptotic normality for the three estimators we have seen so far - stationary MLE, non-stationary MLE and CMLE for $D = 2$. We look at results on stationary data in Section 3.4.1, and non-stationary data in Section 3.4.2. Detailed calculations are provided in Appendix C.1.

3.4.1 Results on stationary data

Let us assume that the data comes from a stationary model with true parameters ν_1^*, ν_2^*, and θ^*. Hence the true data probability is

$$\mathbb{P}(X = x) = \prod_{t=1}^{T} \frac{e^{\left(\nu_1^* x_{1t} + \nu_2^* x_{1t} + \theta^* x_{1t} x_{2t}\right)}}{z^*},$$

where the normalizing constant is given by $z^* = (1 + e^{\nu_1^*} + e^{\nu_2^*} + e^{(\nu_1^* + \nu_2^* + \theta^*)})$. At each time t, $x_t \in \{(0,0)', (0,1)', (1,0)', (1,1)'\}$ has a categorical distribution with probabilities $(p_{00}, p_{01}, p_{10}, p_{11})$ given by

$$p_{00} = \frac{1}{z^*}, \qquad p_{10} = \frac{e^{\nu_1^*}}{z^*}, \qquad p_{01} = \frac{e^{\nu_2^*}}{z^*}, \qquad p_{11} = \frac{e^{\nu_1^* + \nu_2^* + \theta^*}}{z^*}. \tag{3.22}$$

We can use Law of Large Numbers, Central Limit theorem and the Multivariate Delta method (result A.3) to get the limiting values and the asymptotic distributions for the three estimates. Refer to Appendix C.1.2 for details.

The limiting values are given by Equations (C.7), (C.12) and (C.15)

$$\hat{\theta}^{\text{stat}} \longrightarrow \log \frac{e^{(\nu_1^* + \nu_2^* + \theta^*)}}{e^{\nu_1^*} \cdot e^{\nu_2^*}} = \theta^* \tag{3.23}$$

$$\hat{\theta}^{\text{ nonstat}} \longrightarrow 2 \log \frac{2e^{(\nu_1{}^* + \nu_2^* + \theta^*)}}{2e^{(\nu_1{}^* + \nu_2^*)}} = 2\theta^* \tag{3.24}$$

$$\hat{\theta}^{CMLE} \longrightarrow \log \frac{2e^{(\nu_1{}^* + \nu_2^* + \theta^*)}}{2e^{(\nu_1{}^* + \nu_2^*)}} = \theta^* \tag{3.25}$$

which shows that the stationary MLE and the CMLE (with $D = 2$) are consistent for stationary data, but the non-stationary MLE (with $D = 2$) is inconsistent.

The asymptotic distributions are given by Equations (C.9), (C.13) and (C.16):

$$\begin{aligned}
\sqrt{T}\left(\hat{\theta}^{\text{ stat}} - \theta^*\right) &\longrightarrow \mathcal{N}\left(0, \frac{1}{p_{00}} + \frac{1}{p_{10}} + \frac{1}{p_{01}} + \frac{1}{p_{11}}\right) \\
\sqrt{T}\left(\hat{\theta}^{\text{ MLE}} - 2\theta^*\right) &\longrightarrow \mathcal{N}\left(0, \frac{8}{p_{111}} + \frac{8}{p_{110}}\right) \\
\sqrt{T}\left(\hat{\theta}^{\text{ CMLE}} - \theta^*\right) &\longrightarrow \mathcal{N}\left(0, \frac{2}{p_{111}} + \frac{2}{p_{110}}\right)
\end{aligned}$$

where $p_{110} = 2 \cdot p_{10} \cdot p_{01}$, $p_{111} = 2 \cdot p_{11} \cdot p_{00}$ and $p_{00},\ p_{10},\ p_{01},\ p_{11}$ are as defined in (3.22).

Simulation results for stationary data

Let us look at some simulated results. We took the true parameters to be $\nu_1^* = \log(4)$, $\nu_2^* = \log(2)$, $\theta^* = 1$. This gives $\frac{1}{p_{00}} + \frac{1}{p_{10}} + \frac{1}{p_{01}} + \frac{1}{p_{11}} = 51.63$, $\frac{2}{p_{111}} + \frac{2}{p_{110}} = 141.29$ and $\frac{8}{p_{111}} + \frac{8}{p_{110}} = 565.17$ and hence we get

$$\begin{aligned}
\sqrt{T}\left(\hat{\theta}^{\text{ stat}} - \theta^*\right) &\longrightarrow \mathcal{N}(0, 51.63) \\
\sqrt{T}\left(\hat{\theta}^{\text{ MLE}} - 2\theta^*\right) &\longrightarrow \mathcal{N}(0, 565.17) \\
\sqrt{T}\left(\hat{\theta}^{\text{ CMLE}} - \theta^*\right) &\longrightarrow \mathcal{N}(0, 141.29)
\end{aligned} \tag{3.26}$$

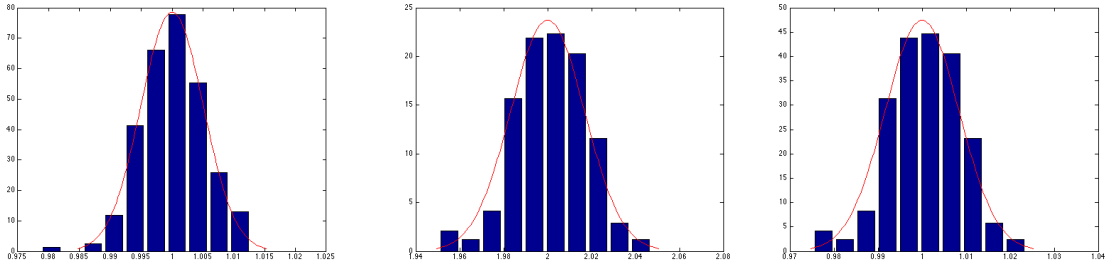
We simulated 250 samples from this true distribution at $T = 2 \times 10^6$ and calculated the three estimators for each sample. In simulations, the sample variances for the normalized estimators (from Equation (3.26)) were 62.3, 580 and 145 respectively.

The asymptotic theory predicts, in this case, that the distributions of the estimators are

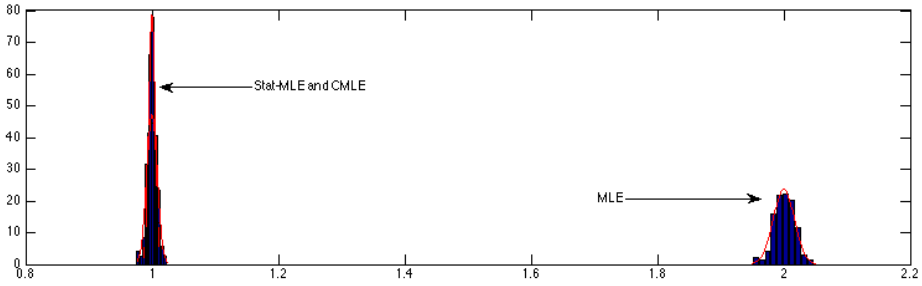
approximately

$$\begin{aligned}
\hat{\theta}^{\text{ stat}} &\overset{d}{\approx} \mathcal{N}\left(\theta^*, 51.63/T\right) = \mathcal{N}\left(1, 2.58 \times 10^{-5}\right) \\
\hat{\theta}^{\text{ MLE}} &\overset{d}{\approx} \mathcal{N}\left(2\theta^*, 565.17/T\right) = \mathcal{N}\left(2, 2.83 \times 10^{-4}\right) \\
\hat{\theta}^{\text{ CMLE}} &\overset{d}{\approx} \mathcal{N}\left(\theta^*, 141.29/T\right) = \mathcal{N}\left(1, 7.06 \times 10^{-5}\right)
\end{aligned} \tag{3.27}$$

In simulations, at $T = 2 \times 10^6$ms, the sample variances for the three estimators were 3.11×10^{-5}, 2.90×10^{-4} and 7.25×10^{-5} respectively. The histograms for the three estimates and their corresponding asymptotic distributions are shown in Figure 3.7.



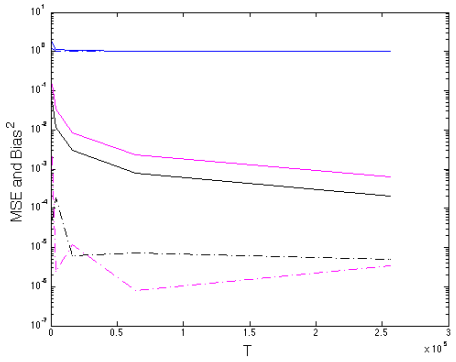
(a) stat-MLE, $\mathcal{N}\left(1, 2.58 \times 10^{-5}\right)$ **(b)** MLE, $\mathcal{N}\left(2, 2.83 \times 10^{-4}\right)$ **(c)** CMLE, $\mathcal{N}\left(1, 7.06 \times 10^{-5}\right)$



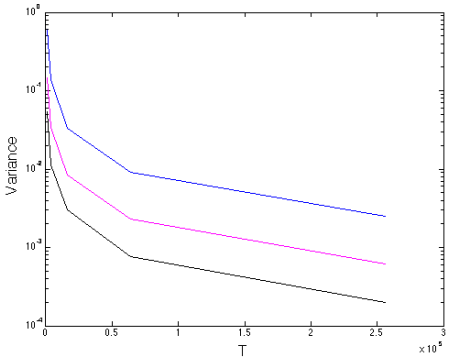
(d) all together

Figure 3.7: Distribution of the estimators (stat-MLE, MLE with $D = 2$ and CMLE with $D = 2$) for 250 samples with $T = 2 \times 10^6$ on stationary data with true θ^*=1: Blue bar graphs are the histograms of the estimators and red curves are the asymptotic distributions from Equations (3.27).

Figure 3.8 shows the mean squared errors (MSE), squared bias and variances (in logarithmic scale) of the three estimates on simulated stationary data (same true parameters $\nu_1^* = \log(4)$, $\nu_2^* = \log(2)$, $\theta^* = 1$ as above) with increasing T values.



(a) MSE (solid line) and squared bias (dotted line)



(b) Variance

Figure 3.8: MSE, squared bias and variance for the three estimators: stationary MLE (black), non-stationary MLE (blue) and CMLE (pink), as T increases using 250 samples from stationary distribution (with true parameters $\nu_1^* = \log(4)$, $\nu_2^* = \log(2)$, $\theta^* = 1$). Note that the y-axis has a logarithmic scale to be able to distinguish between the stat-MLE and CMLE curves.

Asymptotic relative efficiency of CMLE to stationary-MLE

From (C.9) and (C.16), the Asymptotic Relative Efficiency (ARE) [13, Def (10.1.16)] of the CMLE with respect to the stationary MLE is

$$ARE(\hat{\theta}^{\text{CMLE}}, \hat{\theta}^{\text{ stat}}) = \left(\frac{1}{p_{00}} + \frac{1}{p_{10}} + \frac{1}{p_{01}} + \frac{1}{p_{11}}\right) / \left(\frac{2}{p_{111}} + \frac{2}{p_{110}}\right),$$

where $p_{00},\ p_{10},\ p_{01},\ p_{11}$ are as defined in (3.22) and $p_{110} = 2 \cdot p_{10} \cdot p_{01}$ and $p_{111} = 2 \cdot p_{11} \cdot p_{00}$. This gives

$$\begin{aligned} ARE(\hat{\theta}^{\text{CMLE}}, \hat{\theta}^{\text{ stat}}) &= \frac{e^{\theta^*}(p_{10} + p_{01}) + (p_{11} + p_{00})}{e^{\theta^*} + 1} \\ &= \frac{e^{\theta^*}(p_{10} + p_{01}) + (p_{11} + p_{00})}{e^{\theta^*} + 1} \\ &\in \left[\frac{1}{e^{\theta^*} + 1}, \frac{e^{\theta^*}}{e^{\theta^*} + 1}\right] \\ &\leq 1. \end{aligned}$$

We see that the CMLE is less efficient than the stationary MLE, which is not surprising since the data is generated from a stationary model. While the conditional inference does give consistent results, we lose some information by ignoring $g(S(x))$ and using only $\psi(x|S(x))$ from (3.13). We can see the variance of simulated estimates in Figure 3.8(b).

3.4.2 Results on non-stationary data

Now let us explore how these estimates behave asymptotically if the data is generated from a non-stationary model. The simplest case would be if the data came from one set of parameters for the first half and a different set of parameters for the second half. Let us assume that T is divisible by 4 (we want each half to break into windows of width 2). Let the parameters be $(\nu_1^*, \nu_2^*, \theta^*)$ for $t \leq T/2$, and $(\lambda_1^*, \lambda_2^*, \theta^*)$ for $t > T/2$. Hence the true data probability is

$$\mathbb{P}(X = x) = \prod_{t=1}^{T/2} \frac{\exp\left(\nu_1^* x_{1t} + \nu_2^* x_{1t} + \theta^* x_{1t} x_{2t}\right)}{z_\nu^*} \prod_{t=\frac{T}{2}+1}^{T} \frac{\exp\left(\lambda_1^* x_{1t} + \lambda_2^* x_{1t} + \theta^* x_{1t} x_{2t}\right)}{z_\lambda^*},$$

where the normalizing constants are given by $z_\nu^* = (1 + e^{\nu_1^*} + e^{\nu_2^*} + e^{(\nu_1^* + \nu_2^* + \theta^*)})$ and $z_\lambda^* = (1 + e^{\lambda_1^*} + e^{\lambda_2^*} + e^{\lambda_1^* + \lambda_2^* + \theta^*})$.

Detailed calculations are in Appendix C.1.3. The limiting values are given by (C.24), (C.26) and (C.29) as

$$\begin{aligned}
\hat{\theta}^{\text{stat}} &\longrightarrow \theta_0^{\text{stat}} = \theta^* + \log \frac{\left(z_\lambda^* + z_\nu^*\right)\left(z_\lambda^* e^{\nu_1^* + \nu_2^*} + z_\nu^* e^{\lambda_1^* + \lambda_2^*}\right)}{\left(z_\lambda^* e^{\nu_1^*} + z_\nu^* e^{\lambda_1^*}\right)\left(z_\lambda^* e^{\nu_2^*} + z_\nu^* e^{\lambda_2^*}\right)} \\
\hat{\theta}^{\text{ nonstat}} &\longrightarrow 2\theta^* \\
\hat{\theta}^{CMLE} &\longrightarrow \theta^*
\end{aligned}$$

For this non-stationary data, only the conditional MLE is consistent. $\hat{\theta}^{\text{stat}}$ is consistent if and only if $\nu_1^* = \lambda_1^*$ or $\nu_2^* = \lambda_2^*$ (details in Appendix C.1.3). The asymptotic distributions are given by (C.25), (C.28) and (C.30).

Simulation results for non-stationary data

Let us look at some simulated results. We took the true parameters to be $\nu_1^* = \log 4$, $\nu_2^* = \log 2$ for $t \leq T/2$, $\lambda_1^* = \log 4$, $\lambda_2^* = \log 3$ for $t > T/2$, and $\theta^* = 1$ for all t.

This gives $z_\nu^* = z_\lambda^* = 7 + 8e^{\theta^*}$ and the categorical probabilities given by (C.18) and (C.19) evaluate to

$$p_{00}^1 = p_{00}^2 = \frac{1}{7+8e^{\theta^*}}, \quad p_{10}^1 = p_{01}^2 = \frac{4}{7+8e^{\theta^*}}, \quad p_{01}^1 = p_{10}^2 = \frac{2}{7+8e^{\theta^*}}, \quad p_{11}^1 = p_{11}^2 = \frac{8e^{\theta^*}}{7+8e^{\theta^*}}.$$

Using these parameter values and Equations (C.24), (C.26) and (C.29), the limiting values of the three estimators are given by

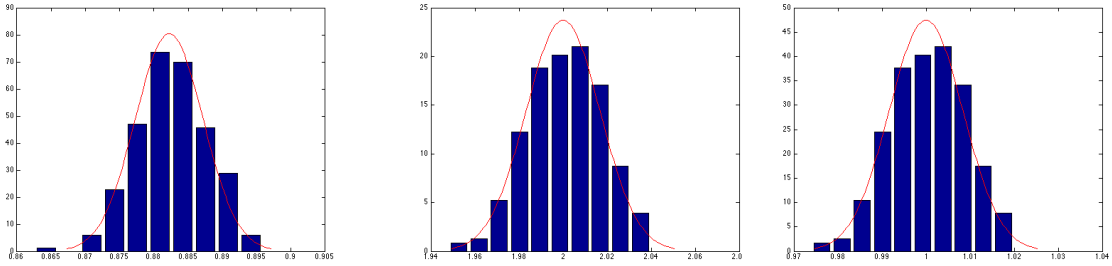
$$\hat{\theta}^{\text{stat}} \longrightarrow \theta^* + \log\frac{8}{9}, \qquad \hat{\theta}^{\text{ nonstat}} \longrightarrow 2\theta^*, \qquad \hat{\theta}^{CMLE} \longrightarrow \theta^*.$$

For this non-stationary data, only the conditional MLE is consistent.

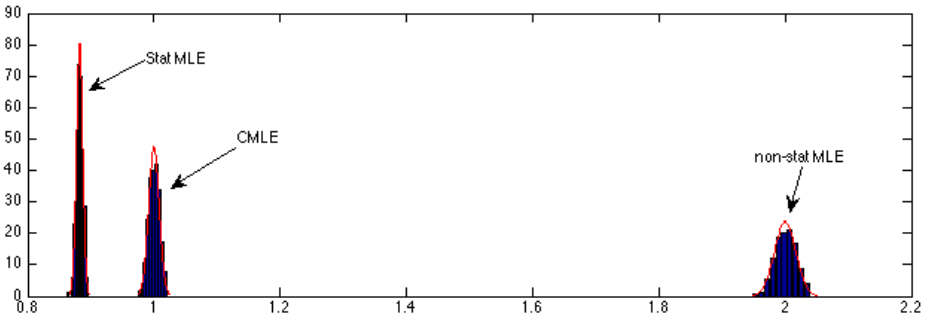
Substituting these values in (C.25), (C.28), (C.30) gives

$$\begin{aligned} \sqrt{T}\left(\hat{\theta}^{\text{ stat}} - \left(\theta^* + \log\frac{8}{9}\right)\right) &= \sqrt{T}\left(\hat{\theta}^{\text{ stat}} - 0.88\right) \longrightarrow \mathcal{N}(0, 49.23) \\ \sqrt{T}\left(\hat{\theta}^{\text{ nonstat}} - 2\theta^*\right) &= \sqrt{T}\left(\hat{\theta}^{\text{ nonstat}} - 2\right) \longrightarrow \mathcal{N}(0, 565.17) \\ \sqrt{T}\left(\hat{\theta}^{\text{ CMLE}} - \theta^*\right) &= \sqrt{T}\left(\hat{\theta}^{\text{ CMLE}} - 1\right) \longrightarrow \mathcal{N}(0, 141.29) \end{aligned} \tag{3.28}$$

We simulated 250 samples for $T = 2 \times 10^6$ ms from this non-stationary (piecewise constant) distribution The sample variances for these normalized estimators (from Equation (3.28)) were 54.34, 579.99 and 145.00 respectively. The asymptotic theory (from (3.28)) predicts



(a) stat-MLE, $\mathcal{N}\left(.88, 2.46 \times 10^{-5}\right)$ **(b)** MLE, $\mathcal{N}\left(2, 2.83 \times 10^{-4}\right)$ **(c)** CMLE, $\mathcal{N}\left(1, 7.06 \times 10^{-5}\right)$



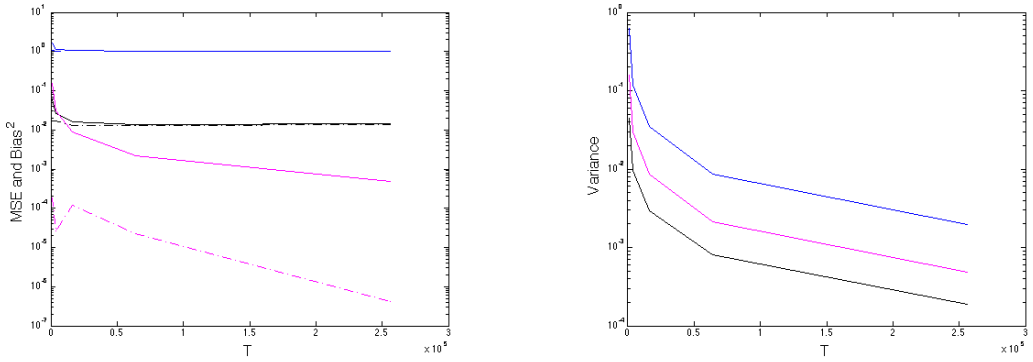
(d) all together

Figure 3.9: Distribution of the estimators (stat-MLE, MLE with $D = 2$ and CMLE with $D = 2$) for 250 samples with $T = 2 \times 10^6$ on non-stationary data with true θ^*=1: Blue bar graphs are the histograms of three estimators and red curves are the asymptotic distributions from equations (3.29).

that the distributions of the estimators are approximately

$$\begin{aligned}
\hat{\theta}^{\text{ stat}} &\overset{d}{\approx} \mathcal{N}\left(0.88, \frac{49.23}{T}\right) = \mathcal{N}\left(0.88, 2.46 \times 10^{-5}\right) \\
\hat{\theta}^{\text{ MLE}} &\overset{d}{\approx} \mathcal{N}\left(2, \frac{565.17}{T}\right) = \mathcal{N}\left(2, 2.83 \times 10^{-4}\right) \\
\hat{\theta}^{\text{ CMLE}} &\overset{d}{\approx} \mathcal{N}\left(1, \frac{141.29}{T}\right) = \mathcal{N}\left(1, 7.06 \times 10^{-5}\right)
\end{aligned} \tag{3.29}$$

In simulations, the variances at $T = 2{\times}10^6$ ms for $\hat{\theta}^{\text{ stat}}$, $\hat{\theta}^{\text{ nonstat}}$ and $\hat{\theta}^{\text{ CMLE}}$ are $2.72{\times}10^{-5}$, 2.90×10^{-4} and 7.25×10^{-5} respectively (see Figure 3.9). The Mean Squared Error (MSE), squared bias and variance for the three estimators using these 250 samples at each T for increasing values of T is shown in Figure 3.10. Comparing with Figure 3.8, we see that the stationary MLE is also inconsistent, and only the CMLE gives results that appear to be consistent.



(a) MSE (solid line) and squared bias (dotted line) **(b)** Variance

Figure 3.10: MSE, squared bias and variance for the three estimators: stationary MLE (black), non-stationary MLE (blue) and CMLE (pink), as T increases using 250 samples from non-stationary distribution (with parameters $\theta^* = 1$, $\nu^*_{1t} = \log 4,\ \nu^*_{2t} = \log 2\ \forall\ t \leq T/2$ and $\nu^*_{1t} = \log 2,\ \nu^*_{2t} = \log 4\ \forall\ t \geq T/2$). The y-axis has a logarithmic scale.

3.5 Asymptotic results for window width $D = 3$

We now explore the asymptotics for the case $D = 3$. We look at the asymptotic performance of the non-stationary MLE and CMLE on stationary data in Section 3.5.1. In Section 3.5.2, we look at simulation results on stationary data as well as non-stationary data.

With window width $D = 3$, possible spike count sums in each window for each neuron are 0, 1, 2 or 3. For MLE and CMLE calculations, we ignore the degenerate windows with sums 0 or 3. Hence the only non-degenerate windows have sums $s_w = (s_{1w}, s_{2w})' \in \mathbb{S}_{good} = \{(1,1)',\ (1,2)',\ (1,1)', (2,2)'\}$. Let us denote the number of these non-degenerate windows as W_{11}, W_{12}, W_{21} and W_{22} respectively, and in general we use W_{ijk} to count the number of windows with $(s_{1w}, s_{2w}, l_w) = (i, j, k)$.

The possible synchrony counts for windows with sums $(1,1)$, $(1,2)$ and $(2,1)$ are 0 and 1 and for windows with sums $(2,2)$, the possible synchronies are 1 and 2. Hence $W_{11} = W_{110} + W_{111}$, $W_{12} = W_{120} + W_{121}$, $W_{21} = W_{210} + W_{211}$ and $W_{22} = W_{221} + W_{222}$.

Non-stationary MLE with $D = 3$

The MLE is given by solving (3.10) and (3.11) to get $\hat{\nu}_{1w}$ and $\hat{\nu}_{2w}$ as a function of θ for each window with $s_w \in \mathbb{S}_{good}$ and then substituting all these estimates in (3.12) to get $\hat{\theta}^{MLE}$. Detailed calculations are included in Appendix C.2.1.

The non-stationary MLE is given by $\hat{\theta}^{\text{MLE}} = \log \hat{\rho}$, where $\hat{\rho}$ is a root of $h(\rho)$ given by Equation (C.35)

$$h(\rho) \triangleq \frac{W_{111} + W_{121} + W_{211} + W_{221} + 2W_{222} - 3\ (W_{11}\gamma_{11} + W_{12}\gamma_{12} + W_{21}\gamma_{21} + W_{22}\gamma_{22})}{W},$$

where $\gamma_{11},\ \gamma_{12}, \gamma_{21}, \gamma_{22}$ are functions of ρ.

CMLE with $D = 3$

Using (3.17) and (3.20), the conditional distribution to be optimized is given by

$$\psi_{good}(x|S = s) \quad = \quad \prod_{w:s_w \in \mathbb{S}_{good}} \frac{\exp(\theta l_w)}{\sum_{k=0}^{D} \exp(\theta k) \cdot h_{s_{1w}, s_{2w}}(k)}.$$

Then the CMLE that maximizes the above conditional likelihood is given by Equation (C.41) (see Appendix C.2.2 for details) as

$$\hat{\theta}^{CMLE} = \log \frac{(2A + B - 2C - D) + \sqrt{(2A + B - 2C - D)^2 + 4(A + 2B)(C + 2D)}}{2(C + 2D)}, \tag{3.30}$$

where $A = 2(W_{111} + W_{222})$, $B = W_{121} + W_{211}$, $C = 2(W_{120} + W_{210})$ and $D = W_{110} + W_{221}$.

3.5.1 Asymptotic results on stationary data

Let us assume that the data comes from a stationary model with true parameters ν_1^*, ν_2^*, and θ^*. Hence the true data probability is

$$\mathbb{P}(X = x) = \prod_{t=1}^{T} \frac{e^{\left(\nu_1^* x_{1t} + \nu_2^* x_{2t} + \theta^* x_{1t} x_{2t}\right)}}{z^*} = \prod_{t=1}^{T} \frac{a^{x_{1t}} \cdot b^{x_{2t}} \cdot c^{x_{1t} x_{2t}}}{z^*},$$

where $a = e^{\nu_1^*}$, $b = e^{\nu_2^*}$ and $c = e^{\theta^*}$ and the normalizing constant is given by $z^* = 1 + e^{\nu_1^*} + e^{\nu_2^*} + e^{\nu_1^* + \nu_2^* + \theta^*} = 1 + a + b + abc$.

Asymptotic behavior of non-stationary MLE on stationary data

Equations (C.37) and (C.38) from Appendix C.2.1 give $h(\rho) \to h^*(\rho)$ where

$$\begin{aligned} h^*(\rho) \;=\; & \frac{3ab}{z^{*3}} \left(\left[c - (c+2) \frac{2\rho + 1 - \sqrt{8\rho+1}}{2(\rho-1)} \right] + (a+b) \left[2c - (2c+1) \frac{3\rho - \sqrt{\rho(8+\rho)}}{2(\rho-1)} \right] \right. \\ & \left. + \; abc \left[2(c+1) - (c+2) \frac{4\rho - 1 - \sqrt{8\rho+1}}{2(\rho-1)} \right] \right). \end{aligned}$$

If $\hat{\rho}$ is a root of h and ρ_0 is a root of the limiting function h^*, then $\hat{\rho} \to \rho_0$, since the roots of a polynomial vary continuously with the co-efficients [24]. This gives $\hat{\theta}^{\mathrm{MLE}} \to \theta_0 = \log \rho_0$. Thus for the stationary distribution, θ_0 (the limiting value of $\hat{\theta}^{\mathrm{MLE}}$) depends on the true values of the incidental parameters as well. Hence we cannot use any sort of a clever correction to correct the MLE (as we can do for the Gaussian case in the Neyman-Scott problem). For varying $\nu_1{}^*$ and $\nu_2{}^*$, with $\theta^* = 1$, the mesh plot of the values of $\theta_0 = \log \rho_0$ is shown in Figure 3.11. For different values of $\nu_1{}^*$ and $\nu_2{}^*$, we get different limiting θ_0, and hence $\hat{\theta}^{\mathrm{MLE}}$ is not consistent.

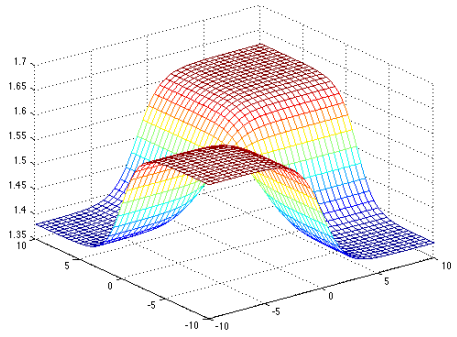


Figure 3.11: Limiting values of $\hat{\theta}^{MLE}$ as a function of varying ${\nu_1}^*$ and ${\nu_2}^*$ with $\theta^* = 1$: $\hat{\theta}^{MLE} \nrightarrow \theta^* = 1$

.

Asymptotic behavior of CMLE on stationary data

For a stationary data, the CMLE with $D = 3$ given by (C.41) converges to the true θ^*: $\hat{\theta}^{\text{CMLE}} \to \theta^*$ (see Appendix C.2.2 for details).

3.5.2 Simulation results

Simulation results for stationary data

For our simulations we use stationary data with parameters $\nu_1^* = \log 4$, $\nu_2^* = \log 2$ and $\theta^* = 1$. Substituting these values in (C.38), the root of the limiting function h^* is given by $\rho_0 = 4.815$ and hence $\hat{\theta}^{\text{MLE}} \to \log \rho_0 = 1.5717$. On the other hand, $\hat{\theta}^{\text{CMLE}} \to \theta^* = 1$. Figure 3.12 shows simulated results for the MLE and CMLE with $D = 3$ in addition to the ones shown in Figure 3.7,

Simulation results for non-stationary data

Figure 3.13 shows simulated results for the MLE and CMLE with $D = 3$ in our previous simulations (non-stationary data with structural parameter $\theta^* = 1 \ \forall \ t$ and incidental

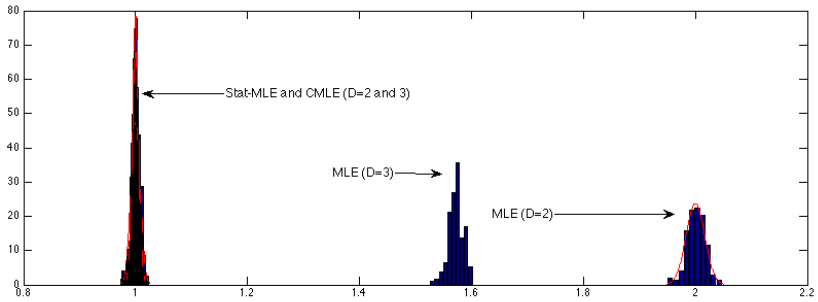


Figure 3.12: Distribution of the estimators (stat-MLE, MLE with $D = 2$ and CMLE with $D = 2$, MLE with $D = 3$ and CMLE with $D = 3$) for 250 samples with $T = 2 \times 10^6$ on stationary data with true θ^*=1: Blue bar graphs are the histograms of the estimators and the red curves are the asymptotic distributions.

parameters $\nu_1^* = \log 4$ and $\nu_2^* = \log 2$ for $t \leq T/2$, and $\lambda_1^* = \log 2$ and $\lambda_2^* = \log 4$ for $t > T/2$) in addition to those shown in Figure 3.9

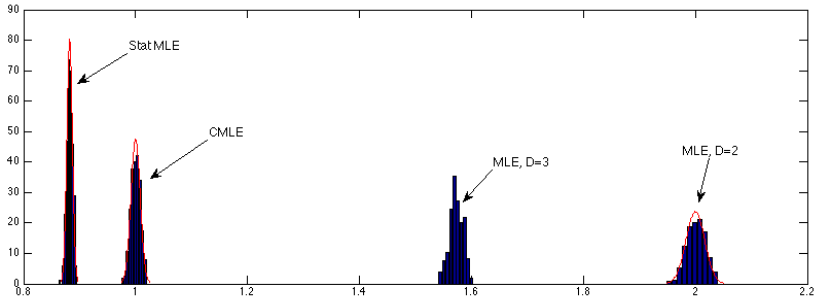


Figure 3.13: Distribution of the estimators (stat-MLE, MLE with $D = 2$ and CMLE with $D = 2$, MLE with $D = 3$ and CMLE with $D = 3$) for 250 samples with $T = 2 \times 10^6$ on non-stationary data with true θ^*=1: Blue bar graphs are the histograms of the estimators and the red curves are the asymptotic distributions.

So far we have seen that the stationary assumption can lead to model misspecification, whereas the non-stationary model can lead to incidental parameter problems. Conditioning appears to overcome the incidental parameter problem and lead to consistent estimates. We shall investigate the conditional inference in greater depth in the next chapter. We prove the consistency of the CMLE for this model and also obtain confidence intervals. We also extend the conditional inference framework to networks with more neurons and develop an approximate composite likelihood approach to scale to larger networks.

Chapter 4

Conditional Inference

So far we have seen that the stationary MLE and the non-stationary MLE were inconsistent for simulated data, whereas the CMLE appeared to be consistent. In this chapter we take a closer look at conditional inference.

We first prove the consistency of the CMLE for our two-neuron zero-lag synchrony model in Section 4.1. We then explore the effect of varying the window width that we choose for conditioning in Section 4.2: larger window widths can lead to model misspecification, whereas smaller window widths can lead to over-conditioning. Next we see how to obtain confidence intervals around the CMLE in Section 4.3. Finally, in Section 4.4, we demonstrate our methods on data recorded from the medial prefrontal cortex of a rat running through a maze to infer neuronal connectivity. We apply our methods to infer synchrony between pairs of neurons, as well as lags and leads.

4.1 Proof of consistency

In this section, we prove the consistency of the CMLE for data generated from some true conditional distribution. Anderson [3] proved consistency of the CMLE for a general setting which corresponds to our non-stationary model with piecewise constant incidental parameters. Our two-neuron model does not satisfy some of the assumptions used in his proof, but we adapt his proof to our case under a more general set of assumptions. We do not make any assumptions about the structure of incidental parameters; instead we prove consistency directly for the conditional model.

The full conditional likelihood is given by Equations (3.16) and (3.17):

$$\psi\left(x|S(x),\theta\right) = \prod_{w} p_w(x_w|\theta)$$

where

$$p_w(x_w|\theta) = \frac{e^{\theta L(x_w)}}{\sum_{y:S(y)=S(x_w)} e^{\theta L(y)}} \tag{4.1}$$

with each $x_w \in \{0,1\}^{2\times D}$. The CMLE is given by Equation (3.18):

$$\hat{\theta}^{\text{CMLE}} = \arg\max_{\theta} \psi\left(x|S(x),\theta\right) = \arg\min_{\theta} \frac{1}{W}\sum_{w} \phi_w(X_w|\theta)$$

where $\phi_w(x_w|\theta) \triangleq -\log p_w(x_w|\theta)\ \forall\ w$.

We saw earlier, in Section 3.3, that we can ignore degenerate windows that do not contribute to the conditional distribution and use only the windows with window sums $s_w \in S_{good}$ where $\mathbb{S}_{good} = \left\{ n = \begin{pmatrix} n_1 \\ n_2 \end{pmatrix} : n_1, n_2 \in \{1,\ldots,D-1\} \right\}$. Ignoring the degenerate windows, (3.20) gives

$$\psi\left(x|S(x),\theta\right) \propto \psi_{good}\left(x|S(x)=s,\theta\right) = \prod_{w:s_w\in\mathbb{S}_{good}} p_w(x_w|\theta)$$

Let $W_{good} = \#\{w : s_w \in \mathbb{S}_{good}\}$. Then the CMLE becomes

$$\hat{\theta}^{\text{CMLE}} = \arg\max_{\theta} \psi_{good}\left(x|S(x) = s, \theta\right) \; = \arg\min_{\theta} \; \frac{1}{W_{good}} \sum_{w:s_w \in \mathbb{S}_{good}} \phi_w(x_w|\theta) \tag{4.2}$$

Theorem 1. *Let $X_w \in \{0,1\}^{2\times D}$, for $D \geq 2$. Let the true parameter be θ^* and let the data $X_1, X_2, \ldots$ be drawn independently from the true conditional distribution*

$$X_w \;\sim\; p_w(\cdot|\theta^*) \qquad \textit{independently} \;\; \forall \; w$$

where p_w is as defined in (4.1). If $S(X_w) = s_w \in \mathbb{S}_{good}$ infinitely often, then the CMLE $\hat{\theta}^{CMLE}$, as defined by (4.2), is strongly consistent, i.e. $\hat{\theta}^{CMLE} \xrightarrow{a.s.} \theta^*$

Proof. First let us assume that $s_w \in S_{good}$ for all $w = 1, \ldots, W$. We have

$$\phi_w(x_w|\theta) = \log\left(\sum_{y:S(y)=S(x_w)} e^{\theta L(y)}\right) - \theta L(x_w) = \log \sum_{y:S(y)=S(x_w)} e^{\theta(L(y)-L(x_w))} \; . \tag{4.3}$$

Note that $\phi_w(\cdot|\theta)$ is differentiable in θ for all w with

$$\frac{\partial}{\partial\theta}\phi_w(x_w|\theta) = \frac{\sum_{y:S(y)=S(x_w)} L(y)e^{\theta L(y)}}{\sum_{y:S(y)=S(x_w)} e^{\theta L(y)}} - L(x_w) \; .$$

Since (4.1) is an exponential family, $\phi_w(\cdot|\theta)$ is a convex function in θ. The second derivative of $\phi_w(\cdot|\theta)$ with respect to θ is given by

$$\begin{aligned}
\frac{\partial^2}{\partial\theta^2}\phi_w(x_w|\theta) &= \frac{\sum_{y:S(y)=S(x_w)} L(y)^2 e^{\theta L(y)}}{\sum_{y:S(y)=S(x_w)} e^{\theta L(y)}} - \left(\frac{\sum_{y:S(y)=S(x_w)} L(y)e^{\theta L(y)}}{\sum_{y:S(y)=S(x_w)} e^{\theta L(y)}}\right)^2 \\
&= \mathbb{E}_\theta\left[L(X)^2|S(X) = S(x_w)\right] - \mathbb{E}_\theta\left[L(X)|S(X) = S(x_w)\right]^2 \\
&= \mathbb{E}_\theta\left[\; \left(L(X) - \mu_L\right)^2 \;\middle|\; S(X) = S(x_w) \;\right] \\
&= \mathbb{V}_\theta\left[\; L(X) \;\middle|\; S(X) = S(x_w) \;\right] \geq 0 \; ,
\end{aligned} \tag{4.4}$$

where $\mu_L = \mathbb{E}_\theta\left[L(X)|S(X) = S(x_w)\right]$. Equality holds in Equation (4.4) if and only if $L(y) = \mu_L \;\forall\; y \in \{y' : S(y') = S(x_w)\}$. Since μ_L is a constant, equality holds only

for those combinations of $s_w = (s_{1w}, s_{2w})$ where all $y \in \{y' : S(y') = s_w\}$ give a single value of $L(y)$. For our case, this happens only for $s_w \in \{(0,0),(0,D),(D,0),(D,D)\}$, i.e. $s_w \notin \mathbb{S}_{good} : L(y) = 0$ for all y with $S(y) \in \{(0,0),(0,D),(D,0)\}$ and $L(y) = D$ for all y with $S(y) = (D,D)$. For all $s_w \in \mathbb{S}_{good}$, $\partial^2\phi_w(x_w|\theta)/\partial\theta^2 > 0$ for all θ, and $\phi_w(x_w|\theta)$ is a strictly convex function in θ. Since we assumed $s_w \in \mathbb{S}_{good}$ for all w, $\phi_w(x_w|\theta)$ is a strictly convex function for all w and hence $\frac{1}{W}\sum_{w=1}^{W}\phi_w(x_w|\theta)$ is strictly convex in θ.

Now we know that for all $y \in \{0,1\}^{2\times D}$ we have $0 \le L(y) \le D$ and $0 \le L(x_w) \le D$ where D is the window width and the sample space had 4^D points. This implies $-D \le L(y) - L(x_w) \le D$ and hence from (4.3) we get

$$\begin{aligned}
\phi_w(x_w|\theta) &\le \log\left(\sum_{y:S(y)=S(x_w)} e^{|\theta|D}\right) \\
&\le \log\left(4^D e^{|\theta|D}\right) \\
&\le D(|\theta| + \log 4) \ .
\end{aligned}$$

Let us define $\mu_w(\theta)$ and $\sigma_w^2(\theta)$ as the mean and variance of $\phi_w(X_w|\theta)$ under the true probability distribution $p_w(\cdot|\theta^*)$ as follows:

$$\begin{aligned}
\mu_w(\theta) &\triangleq \mathbb{E}_{\theta^*}\left[\phi_w(X_w|\theta)\right] \\
\sigma_w^2(\theta) &\triangleq \mathbb{V}_{\theta^*}\left[\phi_w(X_w|\theta)\right] \le \mathbb{E}_{\theta^*}\left[\phi_w(x_w|\theta)^2\right] \le D^2(|\theta| + \log 4)^2 \ .
\end{aligned}$$

Note that

$$\sum_{w=1}^{\infty}\frac{\sigma_w^2(\theta)}{w^2} \le D^2(|\theta| + \log 4)^2\sum_{w=1}^{\infty}\frac{1}{w^2} < \infty \qquad \forall\ \theta \in \mathbb{R}.$$

Now we can use the Strong Law of Large Numbers for independent random variables (A.1) to get

$$\frac{1}{W}\sum_{w=1}^{W}\left(\phi_w(X_w|\theta) - \mu_w(\theta)\right) \xrightarrow{a.s.} 0 \qquad \forall\ \theta \in \mathbb{R}. \tag{4.5}$$

Applying (4.5) simultaneously to θ^* and $(\theta^* + \delta)$ for some small $\delta > 0$, we get

$$\frac{1}{W}\sum_{w=1}^{W}\{\phi_w(X_w|\theta^* + \delta) - \phi_w(X_w|\theta^*)\} - B_W \xrightarrow{a.s.} 0\ , \tag{4.6}$$

where

$$\begin{aligned}
B_W &= \frac{1}{W}\sum_{w=1}^{W}\{\mu_w(\theta^* + \delta) - \mu_w(\theta^*)\} \\
&= \frac{1}{W}\sum_{w=1}^{W}\{\mathbb{E}_{\theta^*}[\phi_w(X_w|\theta^* + \delta)] - \mathbb{E}_{\theta^*}[\phi_w(X_w|\theta^*)]\} \\
&= \frac{1}{W}\sum_{w=1}^{W}\mathbb{E}_{\theta^*}\left[\log\frac{p_w(X_w|\theta^*)}{p_w(X_w|\theta^* + \delta)}\right] \\
&= \frac{1}{W}\sum_{w=1}^{W}\mathcal{D}\left(\ p_w(\cdot|\theta^*)\ ||\ p_w(\cdot|\theta^* + \delta)\ \right) \\
&\geq \min_w\ \mathcal{D}\left(\ p_w(\cdot|\theta^*)\ ||\ p_w(\cdot|\theta^* + \delta)\ \right),
\end{aligned}$$

where $\mathcal{D}(P||Q) \triangleq \mathbb{E}_P[\log(P(X)/Q(X))]$ represents the relative entropy [16, p. 18] or the Kullback-Leibler divergence between probability distributions P and Q. The relative entropy is always non-negative and 0 if and only if $P = Q$ almost everywhere. Let us define distribution $p^s(\cdot|\theta)$ for any $x \in \{0,1\}^{2\times D}$ as follows:

$$p^s(x|\theta) = \frac{\mathbb{1}(S(x) = s)\cdot e^{\theta L(x_w)}}{\sum_{y:S(y)=s} e^{\theta L(y)}} \qquad \theta \in \mathbb{R}.$$

The distribution p^s is similar to p_w in (4.1) before, but for a given fixed statistic s. Let us define κ to be

$$\kappa(\theta^*, \delta) = \min_{s\in\mathbb{S}_{good}}\mathcal{D}(\ p^s(\cdot|\theta^*)\ ||\ p^s(\cdot|\theta^* + \delta)\ ).$$

In this case, $p^s(\cdot|\theta^*+\delta)$ and $p^s(\cdot|\theta^*)$ will be equal almost everywhere only for the degenerate windows $s \notin \mathbb{S}_{good}$ where $p^s(.|\theta)$ is a constant function in θ. Hence $\kappa(\theta^*, \delta) > 0$. Since $s_w \in \mathbb{S}_{good}$ for all windows w, we get

$$B_W \geq \min_w\ \mathcal{D}\left(\ p_w(\cdot|\theta^*)\ ||\ p_w(\cdot|\theta^* + \delta)\ \right) \geq \kappa(\theta^*, \delta) > 0\ , \tag{4.7}$$

where $\kappa(\theta^*, \delta)$ is a strictly positive lower bound on B_W that does not depend on W.

From (4.6) and (4.7), we get

$$\frac{1}{W}\sum_{w=1}^{W}\phi_w(X_w|\theta^*+\delta) > \frac{1}{W}\sum_{w=1}^{W}\phi_w(X_w|\theta^*) \tag{4.8}$$

eventually with probability 1. Repeating this for $(\theta^* - \delta)$, we get

$$\frac{1}{W}\sum_{w=1}^{W}\phi_w(X_w|\theta^*-\delta) > \frac{1}{W}\sum_{w=1}^{W}\phi_w(X_w|\theta^*) \tag{4.9}$$

eventually with probability 1.

(4.8) and (4.9) combined with the fact that $\frac{1}{W}\sum_{w=1}^{W}\phi_w(X_w|\theta)$ is differentiable in θ implies the existence of a critical point in the interval $(\theta^* - \delta, \theta^* + \delta)$ eventually with probability 1. Also since $\frac{1}{W}\sum_{w=1}^{W}\phi_w(X_w|\theta)$ is strictly convex, the critical point is a global minimum. Since $\delta > 0$ can be arbitrarily small, we have $\hat{\theta}^{CMLE} \to \theta^*$ with probability 1.

Now let us allow degenerate windows as well. Since $s_w \in \mathbb{S}_{good}$ infinitely often, we see that $W_{good} \to \infty$. Using (4.2), the proof immediately follows by considering only $w \in \mathbb{S}_{good}$.

□

4.2 Variation of window width D: over-conditioning versus misspecification

The choice of window width D plays an important role in the inference procedure. If we let $D = T$, this represents the stationary model from Section 3.1. In practice, the choice of D will be influenced by how fast we think the firing rates vary. If we expect the non-stationarities in the data to be minimal and the firing rates to vary very slowly with time, we can choose a large D. On the other hand if we expect fast-changing firing rates,

we would choose a smaller D.

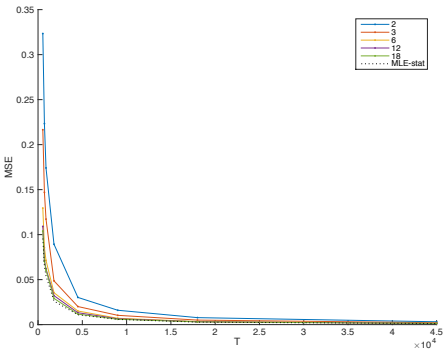
We explore the effects of changing values of D on inference on simulated data - both stationary and non-stationary data. We generate the non-stationary data from a piecewise constant ν^* with $D^* = 6$, $\nu^*_{1w} \in \{-1, -2\}$, $\nu^*_{2w} \in \{-3, -2\}$. To generate the stationary data, we use $\nu^*_1 = -1.5$ and $\nu^*_2 = -2.5$. The true $\theta^* = 1$ for both. We compare the CMLE obtained by using different values of $D = 2,\ 3,\ 6,\ 12,\ 18$ in inference.

Figure 4.1 shows the mean squared error of the CMLE versus T in linear as well as logarithmic scale for both stationary and non-stationary data. On stationary data, the stationary-MLE as well as the CMLE with all different D's appear to be consistent. As D increases, the MSE gets smaller and smaller, and we get the smallest errors for the stationary-MLE. For non-stationary data, we initially see the same phenomenon, i.e. as D increases, the MSE is smaller. But once $D > D^*$, we no longer get consistent estimates since the model is now misspecified.

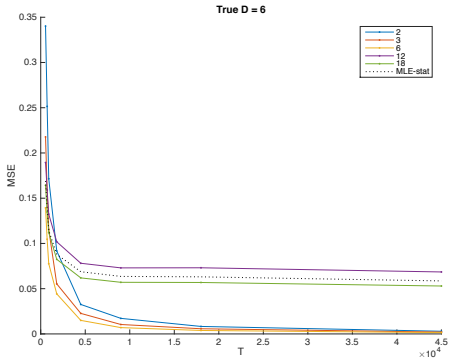
We need to choose D small enough to capture most of the variation in the firing rates, since larger values of D could lead to model-misspecification. On the other hand, taking D to be much smaller than needed leads to over-conditioning. As D gets smaller, the number of statistics to be conditioned on increases. The more we condition, the more information we lose. Even so, the estimates would still be consistent despite the slow convergence. It is better to err on the side of smaller D's and lose some efficiency than choose a large value for D and lose consistency. Ultimately, the choice of D depends on the scientific questions being asked and on what time scales we want to be robust to.

4.3 Confidence intervals

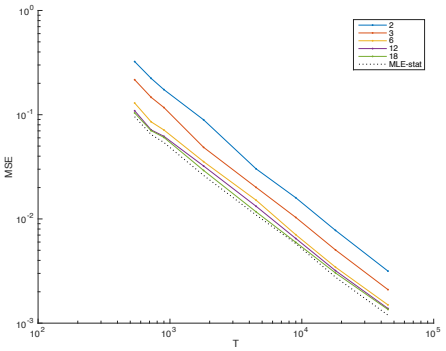
So far we have seen how to get the point estimates using conditional likelihood. We now obtain confidence intervals around these point estimates. We can do so by one of



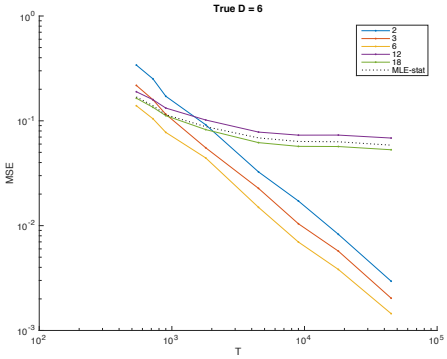
(a) Stationary data (linear scale)



(b) Nonstationary data with true window width D=6 (linear scale)



(c) Stationary data (log scale)



(d) Nonstationary data with true window width D=6 (log scale)

Figure 4.1: Effect of changing window widths on CMLE: over-conditioning versus misspecification

two methods – either using the idea of significance testing by inverting p-values under the conditional distribution)(Section 4.3.1), or by using conditional likelihood ratio tests (Section 4.3.2)

4.3.1 Confidence intervals: significance testing

We can obtain confidence intervals by a two-sided hypothesis test under the null hypothesis $H_0 : \theta = \theta_0$. The confidence interval at level $(1 - \alpha)$ is given by

$$C_\alpha = \{\theta_0 : p^{\theta_0}(x) > \alpha\},$$

where $p^{\theta_0}(x)$ is the p-value i.e. the probability of obtaining a test statistic at least as extreme as the one observed assuming that the null hypothesis is true. Suppose the observed statistics are $L(x) = l_0$ and $S(x) = s = (s_1, \ldots, s_w, \ldots, s_W)$. The two-sided p-value, under the conditional distribution $\psi_\theta(x|S(x))$, is given by

$$\begin{aligned} p^\theta(x) &= 2\min\{\ \mathbb{P}_\theta(L(X) \geq l_0 \mid S(X) = s)\ ,\ \mathbb{P}_\theta(L(X) \leq l_0 \mid S(X) = s)\ \} \\ &= 2\min\left\{\sum_{l=0}^{l_0} \mathbb{P}_\theta(L(X) = l \mid S(X) = s), \sum_{l=l_0}^{T} \mathbb{P}_\theta(L(X) = l \mid S(X) = s)\right\} \\ &= 2\min\left\{\sum_{l=0}^{l_0} \beta_\theta(l), \sum_{l=l_0}^{T} \beta_\theta(l)\right\}, \end{aligned}$$

where

$$\begin{aligned} \beta_\theta(l) \triangleq \mathbb{P}_{\psi_\theta}\ (L(X) = l \mid S(X) = s) &= \sum_{x:L(x)=l} \prod_w p_w(x_w|\theta) \\ &= \sum_{l_w:\sum k_w = l} \prod_w \sum_{x_w:L(x_w)=k_w} p_w(x_w|\theta) \\ &= \sum_{l_w:\sum k_w = l} \prod_w q_w(k_w). \end{aligned}$$

Here $p_w(\cdot|\theta)$ is as defined in Equation (3.17) and $q_w(k)$ is the probability of observing the statistic $L(x_w) = k_w$ in the w-th window given by

$$\begin{aligned} q_w(k) &= \sum_{x_w:L(x_w)=k} p_w(x_w|\theta) \\ &= \sum_{x_w:L(x_w)=k} \frac{\mathbb{1}(S(x_w) = s_w) \cdot e^{\theta L(x_w)}}{\sum_{m=0}^{D} e^{\theta m} \cdot h_{s_w}(m)} \\ &= \sum_{x_w:L(x_w)=k} \frac{\mathbb{1}(S(x_w) = s_w) \cdot e^{\theta k}}{\sum_{m=0}^{D} e^{\theta m} \cdot h_{s_w}(m)} \\ &= \frac{e^{\theta k}}{\sum_{m=0}^{D} e^{\theta m} \cdot h_{s_w}(m)} \cdot \#\{x_w : L(x_w) = k, S(x_w) = s_w\} \\ &= \frac{e^{\theta k} \cdot h_{s_w}(k)}{\sum_{m=0}^{D} e^{\theta m} \cdot h_{s_w}(m)}. \end{aligned}$$

Since each window is independent, we can compute $\beta_\theta(l)$ as a W-fold convolution of $q_w(k_w)$: $\beta_\theta(l) = (q_1 * q_2 * \cdots * q_w * \cdots * q_W)(l)$.

For our p-value calculations, we need to calculate $\beta_\theta(l)$ at all possible values of l, which will require us to calculate $q_w(k)$ at all values of k. We pre-compute these and build a table of all possible $q_w(\cdot)$, convolve all combinations, and sum them up to get the p-value. Inverting the p-value at a given level gives the confidence interval at that level. To find the two end-points we can find the two roots of the θ equation $p^\theta(x) - \alpha = 0$.

In Section 4.3.3 we show results for coverage probability on simulated data. In Section 4.4 we show how we can use the confidence intervals to make inferences about neuronal connectivity. Before that, let us explore another method of getting confidence intervals.

4.3.2 Confidence intervals: likelihood ratio tests

We can also obtain a confidence interval by inverting a likelihood ratio test [13, p. 423], [38, p. 384,447]. Anderson [2] shows that using the full likelihood for the likelihood ratio test fails in the presence of incidental parameters, but suggests using the conditional likelihood ratio (the likelihood ratio obtained from the conditional likelihood) instead. If x is the observed data and $\hat{\theta}^{\text{CMLE}}$ is the conditional MLE, the conditional likelihood ratio is given by

$$\lambda_\theta(x) = \frac{\mathbb{P}\left(X = x \mid \theta, S(X) = S(x)\right)}{\mathbb{P}\left(X = x \mid \hat{\theta}_{\text{CMLE}}, S(X) = S(x)\right)} = \frac{\psi_\theta(x|S(x))}{\psi_{\hat{\theta}^{CMLE}}(x|S(x))}.$$

Suppose we want to get the confidence interval C_α at level $(1-\alpha)$, i.e. we need to find the region C_α such that

$$\mathbb{P}_\theta(\theta \in C_\alpha) < 1 - \alpha.$$

Asymptotically $\Lambda_\theta(X) = -2\log\lambda_\theta(X)$ is a χ^2 distributed random variable [2]. If $h(\cdot)$ is the cumulative distribution function of a χ^2 distribution, then

$$\mathbb{P}_\theta(\Lambda_\theta(X) < h^{-1}(1-\alpha)) < 1 - \alpha.$$

Let us denote $c_\alpha \triangleq h^{-1}(1-\alpha)$. Then

$$\mathbb{P}_\theta\left(\Lambda_\theta(X) < c_\alpha\right) = \mathbb{P}_\theta\left(-2\log\lambda_\theta(X) < c_\alpha\right) = \mathbb{P}_\theta\left(\lambda_\theta(X) > \exp\left(-\frac{c_\alpha}{2}\right)\right) < 1-\alpha.$$

The confidence interval is then given by

$$C_\alpha(x) = \left\{\theta \in \mathbb{R} : \lambda_\theta(x) > \exp\left(-\frac{c_\alpha}{2}\right)\right\}.$$

We can form a fine grid around $\hat{\theta}^{CMLE}$ to find the region where $\lambda_\theta > \exp(-\frac{c_\alpha}{2})$, or we can get the boundary points of C_α by finding the two roots of the function $\lambda_\theta - \exp(-\frac{c_\alpha}{2}) = 0$.

4.3.3 Results for coverage probability of intervals

We simulate data from some true θ^* and then obtain 95% confidence intervals using our two methods: inverting p-values and using likelihood ratio tests (LRT). We then check whether the intervals contain the true θ^*. We repeat this procedure with 2500 different datasets with different values of θ^*. The simulation results are shown in Table 4.1. The true parameters for the data were as follows: the ν^*'s are piecewise constant with true window width $D^* = 50$, $\nu^*_{1w} \sim \text{Unif}(-5.2, -2.2)$, $\nu^*_{2w} \sim \text{Unif}(0,5)\ \forall$ w and $\theta^* \sim \mathcal{N}(0, 0.8)$. For inference we use $D = 5$ and increasing lengths of data: $T = 1250,\ 2500,\ 5000,\ 10000,\ 20000,\ 40000$ ms and we repeat the procedure for 2500 different datasets to get the percentage of times confidence intervals covered the true θ^*. We observe that the confidence intervals obtained by inverting the p-values were wider than the ones obtained from the likelihood ratio tests, and hence had higher coverage probabilities. We also see that as the data grows larger, the confidence intervals appear to approach 95%.

Table 4.1: Simulation results for coverage probability of confidence intervals (percentage of simulations where true θ is in the 95 % confidence intervals) with 2500 iterations.

T(ms)	1250	2500	5000	10000	20000	40000
p-value	97.47 %	96.44 %	96.40 %	96.44	96.40 %	95.92 %
LRT	94.98 %	94.67 %	95.32 %	95.28	95.92 %	95.52 %

4.4 Results on neuronal recordings

We now apply this conditional inference technique to analyze neuronal recordings from the medial prefrontal cortex of a rat running in a maze. This data was generously provided to us by Fujisawa and Buzsaki and is described in more detail in [22]. We apply our methods to infer lag-lead relationships (Section 4.4.1), as well as for synchrony between pairs of neurons (Section 4.4.2).

4.4.1 Mono-synaptic connections (lag-lead relationships)

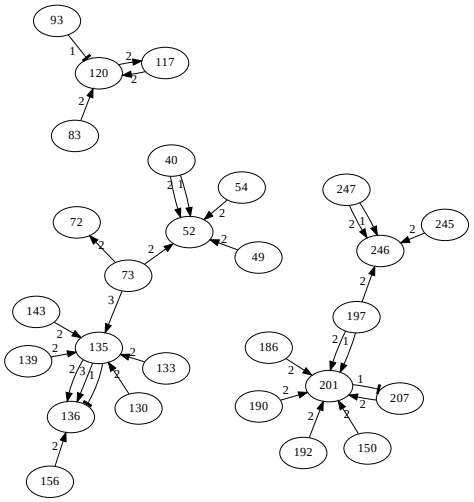
So far we have only looked at the two-neuron zero-synchrony model. We can easily extend this to infer lags and leads between neuron cells by looking at shifted spike trains. If we are interested in exploring the connection from neuron cell i to j at a time lag of l ms, we can shift the spike train of neuron cell j by l ms to the left and then apply conditional inference for the zero-lag model to get point estimates and confidence intervals.

Let us look at the dataset described in [22] and [25] for measuring mono-synaptic connections. The dataset includes spike train recordings with time bins of 1 ms (hence each bin has at most 1 spike), recorded over a period of around 47 minutes ($T = 2811890$ ms) for N=117 different neuron units. The observed data $x = (x_{i,t} : i = 1, \ldots, N,\ t = 1, \ldots, T)$ is a $N \times T$ binary matrix. We analyze each pair for lagged connections up to 4 ms. Hence θ is a $117 \times 116 \times 4$-dimensional. We take the window width D to be 10 ms, which corresponds to the length of the jitter-interval used in [22] and [25]. To infer θ_{ijk} using the two neuron zero-lag synchrony model, we use the spike trains $x_{i,1:T-l}$ and $x_{j,l+1:T}$.

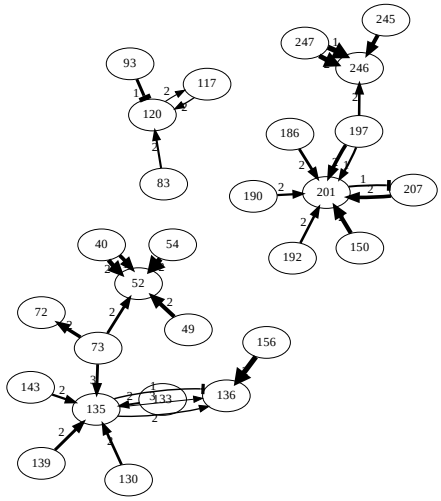
Since we are interested in $m = 117 \times 116 \times 4 = 54288$ different simultaneous intervals, we use the Bonferroni correction [38, p. 468-470]. If we wish to have an overall confidence level of $(1 - \alpha)$, we use the level $(1 - \alpha/m)$ for each individual confidence interval. To identify a network of excitatory and inhibitory connections between neurons, we can then

prune all the edges that have 0 in their confidence intervals, and keep only the edges that are entirely positive (excitatory) or entirely negative (inhibitory).

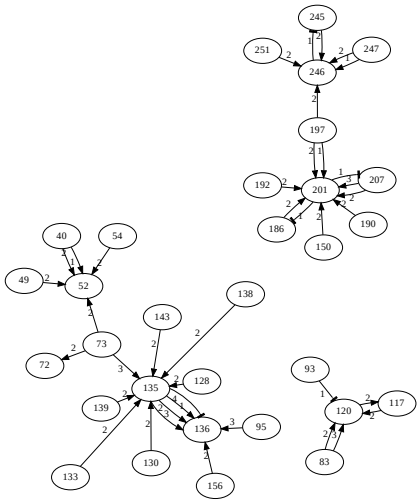
An alternate way of obtaining the same pruning, is to use a multiple testing framework by computing the p-value of $\theta_{ijl} \neq 0$ and adjusting for multiple tests. If we use Bonferroni correction, we get the same pruning as mentioned before. We can also use other corrections such as the Benjamini-Hochberg-Yekutieli (BHY) [5] or the Benjamini-Hochberg (BH) corrections [4]. (Appendix A.1 of [25] contains more details of each type of correction.)



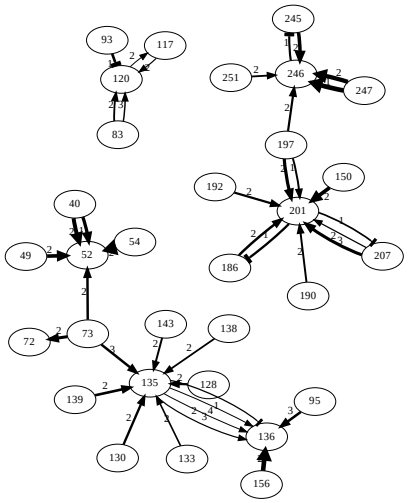
(a) Bonferroni pruning



(b) Bonferroni: Edges weighted by the CMLE value



(c) BHY pruning

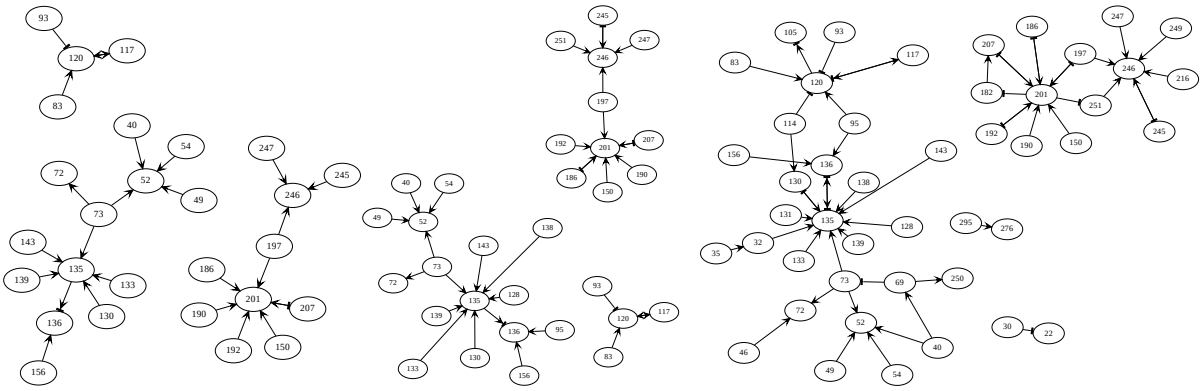


(d) BHY: Edges weighted by the CMLE value

Figure 4.2: Network visualization for lag-lead relationships using Bonferroni and BHY corrections: A standard arrow (with arrow head) represents an excitatory connection ($\theta_{ijl} > 0$) and an arrow with a flat head represents an inhibitory connection ($\theta_{ijl} < 0$). The number next to the arrow represents the lag. On the right, we weight the thickness of the arrow using the absolute value of the CMLE.

Using $\alpha = 0.05$, we obtained 31, 40 and 73 significant edges using the Bonferroni, the BHY and the BH corrections respectively. Figure 4.2 shows a network visualization of our results. The numbers inside the nodes represent the neuron unit number, as numbered in the dataset. A standard arrow (with arrow head) represents an excitatory connection ($\theta_{ijl} > 0$) and an arrow with a flat head represents an inhibitory connection ($\theta_{ijl} < 0$). The number next to the arrow represents the lag in milliseconds. For the figures on the right, the thickness of the arrows have been weighted by the value of the CMLE.

In Figure 4.3, we show condensed networks (using all three types of pruning) with only a single connection shown between two neuron units when at least one of the 4 lags is observed to be significant. This graph closely resembles the network visualization from [25, p. 442, Fig. 4] and also [22, p. 826, Fig. 3d]. In this collapsed network, we get 26, 32 and 54 significant edges using the Bonferroni, the BHY and the BH corrections, respectively. [22] reported 78 significant edges, whereas [25] obtained 26, 27, and 46 rejections for this dataset.



(a) Bonferroni correction **(b)** BHY **(c)** BH

Figure 4.3: Condensed visualization for lag-lead relationships using Bonferroni, BHY and BH corrections: A standard arrow (with arrow head) represents an excitatory connection ($\theta_{ijl} > 0$) and an arrow with a flat head represents an inhibitory connection ($\theta_{ijl} < 0$).

4.4.2 Zero-lag synchrony

We also analyzed the data for synchronous firing. The data was recorded using 8 different shanks; we ignore the pairs of neuron units that belong to the same shank. Using a confidence level of 99% (α=0.01) for each neuron pair, we get 41 significant edges as shown in Figure 4.4.

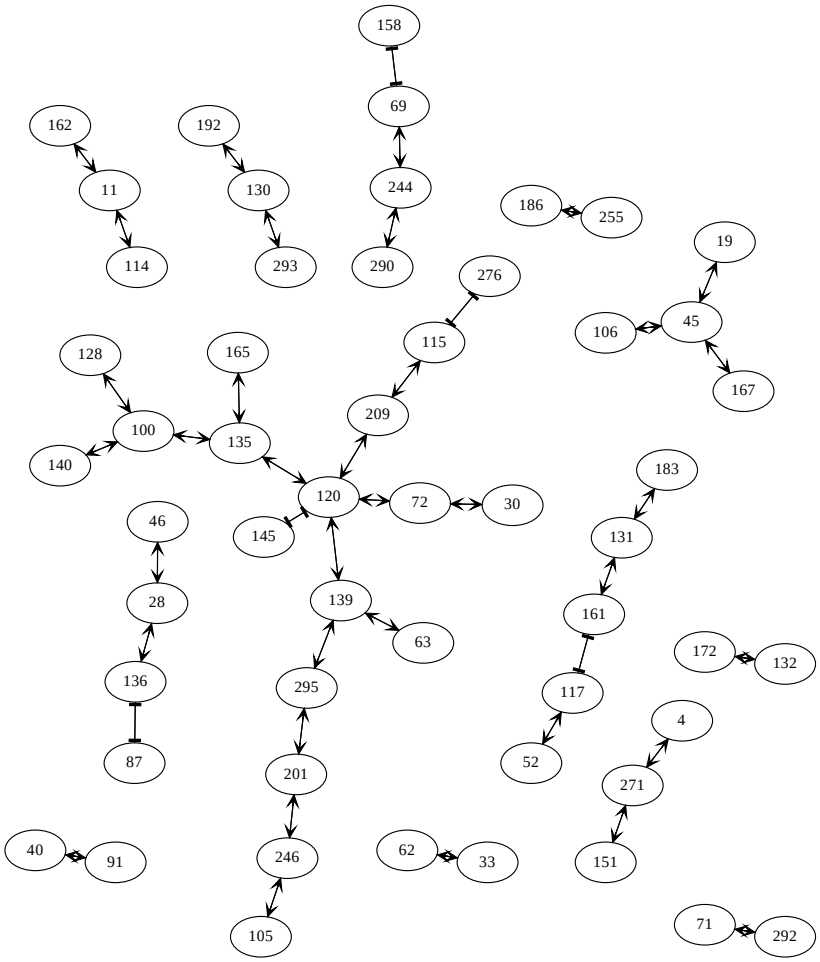
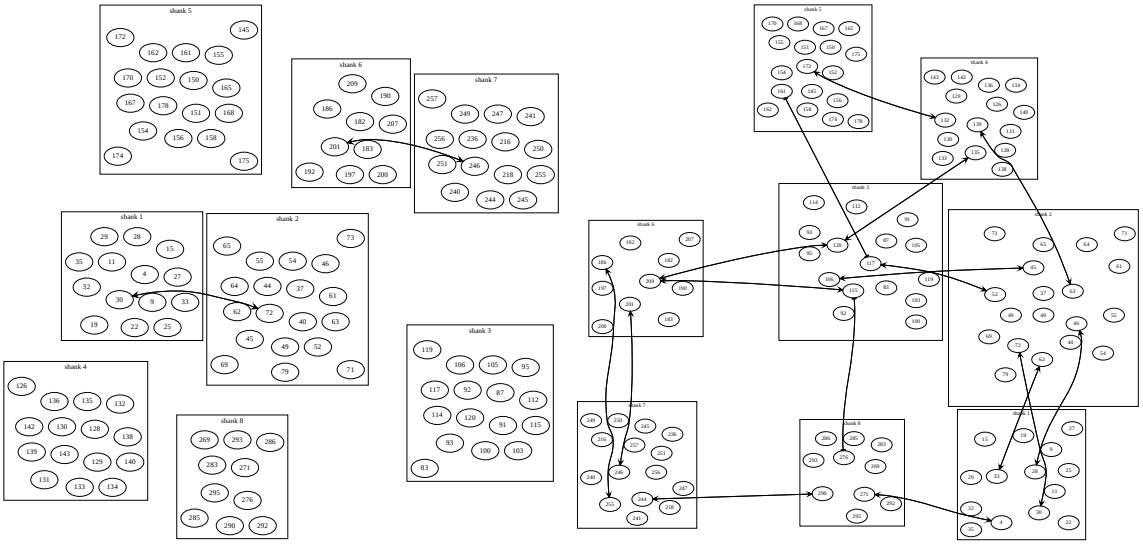


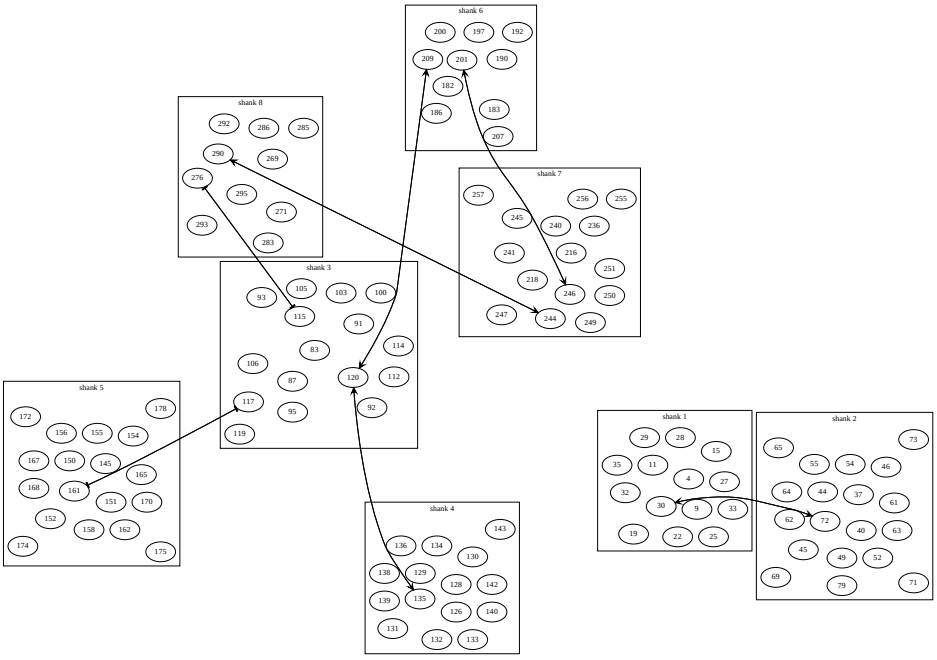
Figure 4.4: Synchronous pairs that are significant at p-value of 0.01.

We can also use a multiple testing correction for the network. Using the BHY and the BH corrections at overall level 90% (α=0.1) gives 2 and 16 significant different-shank synchronies respectively, whereas using the BH corrections at an overall level of 95% (α=0.05) gives 7 synchronies as shown in Figure 4.5.



(a) BHY correction with overall α=0.1

(b) BH correction with overall α=0.1



(c) BH correction with overall α=0.05

Figure 4.5: Synchronous pairs that are significant using Bonferroni, BHY and BH corrections.

Chapter 5

Composite likelihood approach

So far we explored the conditional inference for a two-neuron model. We even used our two-neuron model to infer larger networks by estimating each edge independently. In this chapter, we extend this technique to a model for larger networks that can estimate all the different θ's, corresponding to different neuron pairs, together. As the network size grows, the CMLE computation becomes more intensive. We explore composite likelihood methods (also known as pseudo-likelihood methods) to tackle this computational issue in Section 5.2.

5.1 Log linear model for N neurons

Consider a network of N neurons. Let the pairwise coefficient for the interaction of unit i and unit j be θ_{ij}. The data is a $N \times T$ binary matrix given by $x = (x_{it} : i = 1, \ldots, N \; t = 1, \ldots, T) \in \mathcal{X} = \{0, 1\}^{N \times T}$. The stationary log-linear model is of the form

$$p(x) = \frac{1}{Z(\nu, \theta)} \exp\left(\sum_{t=1}^{T}\left(\sum_{i=1}^{N} \nu_i x_{it} + \sum_{i=1}^{N}\sum_{j>i} \theta_{ij} x_{it} x_{jt}\right)\right),$$

where the normalizing constant is given by

$$Z(\nu,\theta) = \sum_{y\in\mathcal{X}} \exp\left(\sum_{t=1}^{T}\left(\sum_{i=1}^{N}\nu_i y_{it} + \sum_{i=1}^{N}\sum_{j>i}\theta_{ij}y_{it}y_{jt}\right)\right).$$

The conditional distribution is given by $\psi(x|S=s) = \prod_w \psi_w(x_w|S(x_w)=s_w)$, where

$$\begin{aligned}
\psi_w(x_w|S(x_w)=s_w) &= \mathbb{P}(X_w = x_w|S(x_w)=s_w) \\
&\propto \exp\left(\sum_{i=1}^{N}\sum_{j>i}\left(\theta_{ij}\sum_{t\in\Omega_w} x_{it}x_{jt}\right)\right)\mathbb{1}(S(x_w)=s_w) \\
&\propto \exp\left(\sum_{i=1}^{N}\sum_{j>i}\theta_{ij}L_w^{ij}(x_w)\right)\mathbb{1}(S(x_w)=s_w) \\
&\propto \exp\left(\boldsymbol{\theta}\cdot\boldsymbol{L}_w(x_w)\right)\cdot\mathbb{1}(S(x_w)=s_w),
\end{aligned}$$

where $\boldsymbol{\theta}$ is a vector obtained by concatenating all the θ_{ij}'s together, $L_w^{ij}(x_w) = \sum_{t\in\Omega_w} x_{it}x_{jt}$, and $\boldsymbol{L}^w(x_w)$ is a vector by concatenating L_w^{ij}s.

The normalizing constant for each window is then given by summing over all possible combinations $y_w \in \{0,1\}^{N\times D}$, where the sums of the spike counts in that window are the same as those of the data:

$$Z_w = \sum_{y_w}\exp(\boldsymbol{\theta}\cdot\boldsymbol{L}_w(y_w))\mathbb{1}(S(x_w)(y_w)=s_w).$$

Hence we obtain

$$\begin{aligned}
\psi(x|S=s) &= \prod_{w=1}^{W}\frac{\exp(\boldsymbol{\theta}\cdot\boldsymbol{L}_w(x_w))\cdot\mathbb{1}(S(x_w)=s_w)}{\sum_{y_w}\exp(\boldsymbol{\theta}\cdot\boldsymbol{L}_w(y_w))\cdot\mathbb{1}(S(x_w)(y_w)=s_w)} \\
&= \prod_{w=1}^{W}\frac{\exp(\sum_i\sum_{j>i}\theta_{ij}L_w^{ij}(x_w))\cdot\mathbb{1}(S(x_w)=s_w)}{\sum_{y_w}\exp(\sum_i\sum_{j>i}\theta_{ij}L_w^{ij}(y_w))\cdot\mathbb{1}(S(x_w)(y_w)=s_w)}.
\end{aligned} \tag{5.1}$$

In principle, we can use the same gradient descent methods to find the CMLE in this multidimensional setting.

Figures 5.1 and 5.2 show results for a 7-neuron network on simulated non-stationary data.

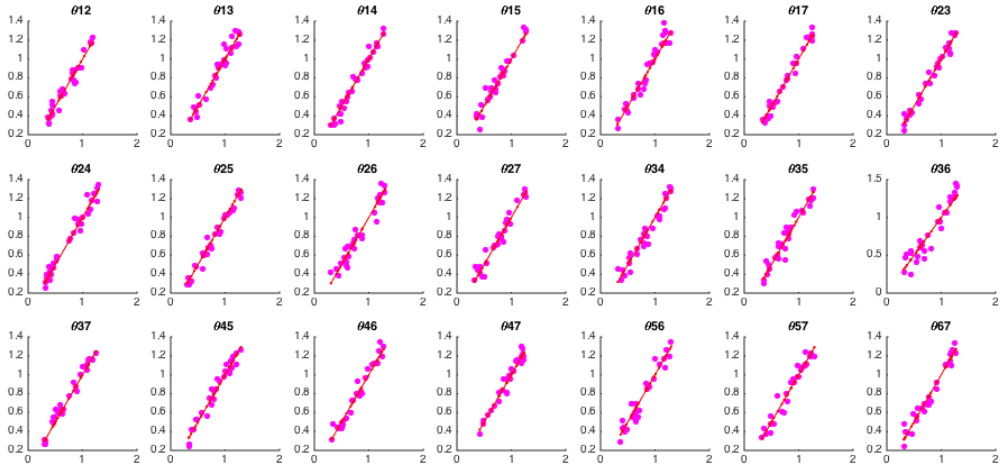
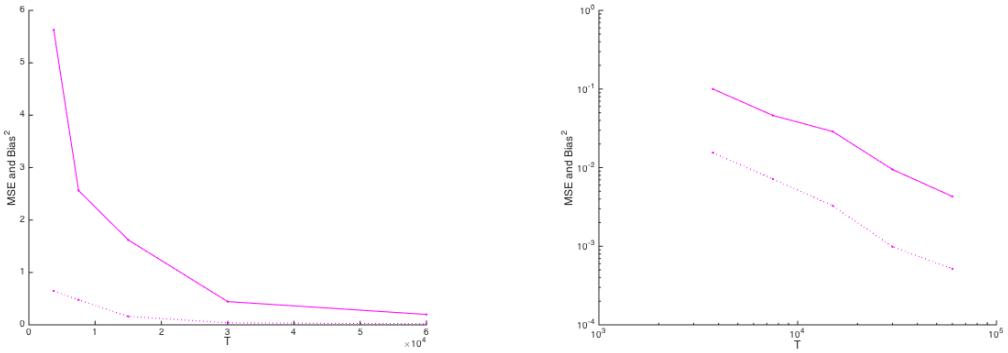


Figure 5.1: Scatter plot of $\hat{\theta}_{ij}^{CMLE}$ versus θ_{ij}^* for different neuron pairs on simulated trials at $T = 60$ s.



(a) Linear scale **(b)** Log scale

Figure 5.2: MSE (solid) and squared bias (dotted) for joint CMLE as T increases, for all θ_{ij}'s together.

We show scatter plots of $\hat{\theta}_{ij}^{CMLE}$ versus true θ_{ij}^* for each of the 21 different pairs in Figure 5.1. Figure 5.2 shows the mean squared error and squared bias averaged over various simulated trials for all the θ_{ij}'s combined. We see that the joint CMLE for this 21-dimensional $\boldsymbol{\theta}$ also appears to be consistent.

The normalizing constant involves summing over all possible $y_w \in \{0,1\}^{N\times D}$ in a window where $S(y_w) = s_w$. As the network size grows this becomes computationally burdensome. In the next section we explore an approximate method that can scale better with growing number of neurons.

5.2 Composite likelihood method

Exact computation of the conditional distribution involves summing over all possible spike trains with the same coarser dynamics and this can be computationally intensive as the network size increases. [19] and [42] suggested using an approximate conditional approach using composite likelihood (or pseudo-likelihood) for such cases with higher dimensional θ. In this section we explore a similar composite likelihood framework for our model.

The exact conditional distribution is given by (5.1)

$$\begin{aligned} \psi(x|S(x)=s,\theta) &= \prod_{w=1}^{W} \psi_w(x_w|S(x_w)=s_w,\theta) \\ \psi_w(x_w|S(x_w)=s_w,\theta) &= \frac{\exp(\boldsymbol{\theta}\cdot \boldsymbol{L}^w(x))\cdot \mathbb{1}(S(x_w)=s_w)}{\sum_{y_w}\exp(\boldsymbol{\theta}\cdot \boldsymbol{L}^w(y_w))\cdot \mathbb{1}(S(y_w)=s_w)}, \end{aligned}$$

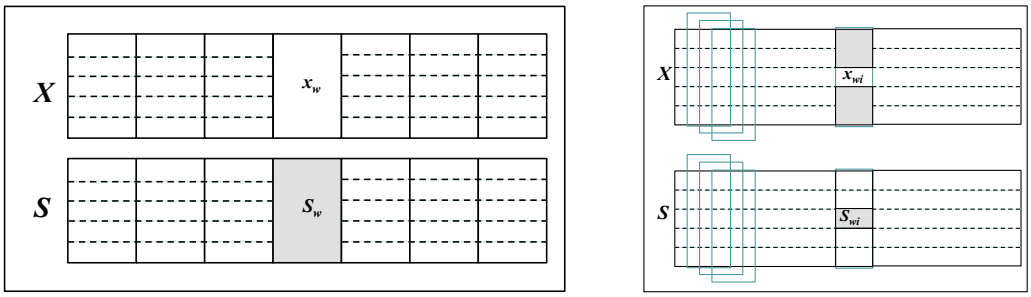
where the normalizing constant for ψ is

$$Z=\prod_w Z_w=\prod_{w=1}^{W}\left(\sum_{y_w}\mathbb{1}(S(y_w)=s_w)\cdot\exp(\boldsymbol{\theta}\cdot\boldsymbol{L}^w(y_w))\right).$$

Each window is independent, however within each window the normalizing constant calculation involves summing over all possible y_w that have the same window sums as the given data x_w, which is a computationally burdensome as the number of neurons in the network increases.

We shall explore a composite likelihood method. We approximate the conditional probability of the full window X_w given its sum $S(X_w)$ by a composite likelihood consisting of the conditional distribution of a particular neuron X_{iw} given its sum $S(X_{iw})$ and the remaining data $(X_{jw} : j \neq i)$, and then we take the product over all neurons in the network. In each window, this composite likelihood looks like

$$\prod_{i=1}^{N}\mathbb{P}\left(X_{iw}=x_{iw}\mid S(x_{iw})=s_{iw},\ X_{wj}\ \forall j\neq i\right).$$



(a) Conditional inference binning **(b)** Composite conditional inference: Sliding windows

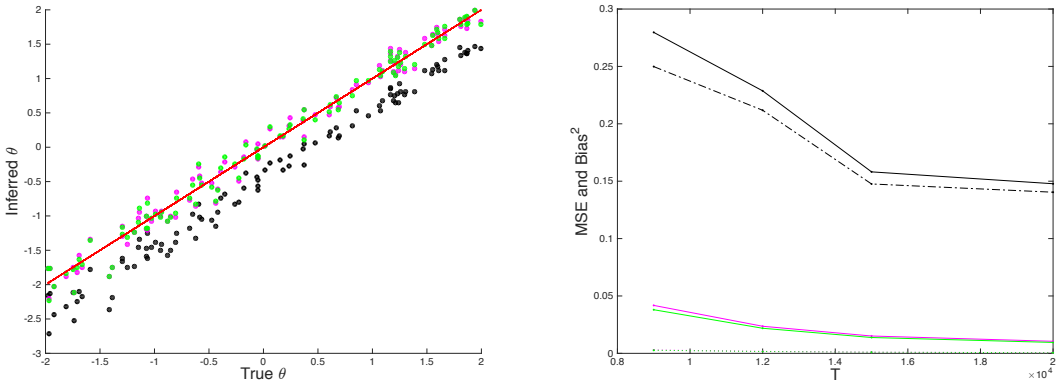
Figure 5.3: Schematic diagram depicting the difference between conditional likelihood and the approximate composite conditional likelihood for the log-linear model.

In addition, we average over sliding windows instead of looking at disjoint windows. There are no discrete windows inherently present in the data – the concept of coarse binning was introduced by us in the model. In an effort to average out this effect, we use sliding bins and the composite likelihood is obtained as a product over all different sliding bins. So the composite likelihood will be

$$\prod_{w=1}^{T-D+1}\left(\prod_{i=1}^{N}\mathbb{P}\left(X_{iw}=x_{iw}\mid S(x_{iw})=s_{iw},\ x_{wj}\forall j\neq i\right)\right).$$

In simulations we see that our composite conditional approach gives very similar results as the conditional inference. Figure 5.4 shows the results using this composite likelihood approach on the simulated data from Chapter 3, superimposed on the CMLE and stationary MLE results. The composite CMLE appears to a good approximation to the CMLE and shows very similar performance as the CMLE on our simulations.

We also include some simulation results for a three neuron network. In this case, there are three parameters of interest, θ_{12}, θ_{23} and θ_{13}, each corresponding to the three pairwise interactions. For each θ_{ij} we see that composite inference closely resembles the exact conditional inference and performs better than the stationary model.



(a) Scatter plot of $\hat{\theta}$ versus θ^* **(b)** MSE (solid) and squared bias (dotted)

Figure 5.4: CMLE (pink) and composite CMLE (green) compared with stationary-MLE (black) on non-stationary data from Chapter 3: On the left we have inferred $\hat{\theta}$ versus true θ for various simulated trials with $T = 12000$ ms. On the right we have mean square error and squared bias plots averaged over multiple simulations for varying values of T. Composite CMLE appears to be a good approximation to the CMLE.

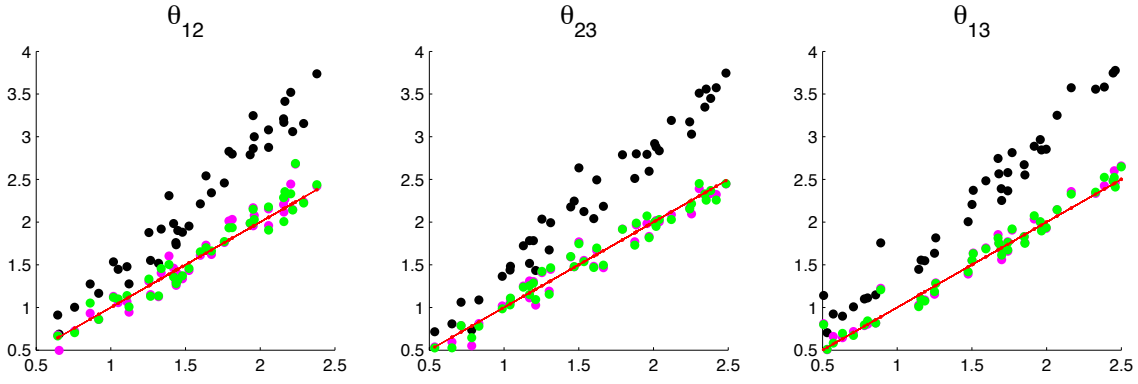


Figure 5.5: Scatter plot of inferred $\hat{\theta}_{ij}$ (on y-axis) versus true θ^*_{ij} (on x-axis) at $T = 12000$ ms for a 3 neuron network: Black - stationary MLE, pink - CMLE, green - composite conditional MLE.

Chapter 6

Non-parametric Bayesian approach with Dirichlet process mixture prior

In Section 3.2 we introduced the non-stationary model with piecewise constant ν's given by Equation (3.4)

$$p(x|\vec{\nu_1}, \vec{\nu_2}, \theta) = \frac{1}{Z(\vec{\nu_1}, \vec{\nu_2}, \theta)} \prod_{w=1}^{W} \prod_{t \in \Omega_w} \exp(\nu_{1w} x_{1t} + \nu_{2w} x_{2t} + \theta x_{1t} x_{2t}).$$

We saw that the maximum likelihood estimate gave inconsistent results for θ because of the incidental parameter problem. In Section 3.3 we saw that the conditional inference approach gives us consistent results. We now explore the alternate marginal (Bayesian) approach.

The marginal approach involves integrating out the incidental parameters using a prior distribution. As we saw in Chapter 2, for a Bayesian approach to work we need to identify this prior accurately. We need to be agnostic about the form of the prior and allow for

our method to capture the true prior without imposing parametric restrictions. Kiefer and Wolfowitz [27] proved consistency for this type of model (using a continuous distribution H on the ν's and estimating the distribution H simultaneously with θ). In practice we can model the continuous distribution H by using a non-parametric prior on the ν's.

In Section 6.1, we introduce the non-parametric Dirichlet process mixture (DPM) priors and describe a Gibbs sampling algorithm to sample from a DPM distribution [31]. We then use this DPM distribution as a prior on the incidental parameters ν_1 and ν_2 in our two-neuron zero-lag synchrony model from Chapter 3. In Section 6.2 we use independent DPM priors on ν_1 and ν_2, whereas in Section 6.3 we use a two-dimensional joint DPM prior on $\nu = (\nu_1, \nu_2)'$.

6.1 Dirichlet process mixture (DPM) priors

Let $\nu_1, \nu_2, \ldots, \nu_W$ be distributed according to a Dirichlet process mixture (DPM) prior as as described in [31]:

$$\left.\begin{aligned} G &\sim DP(G_0, \alpha) \\ \beta_w &\sim G \\ \nu_w \mid \beta_w &\sim F(\cdot|\beta_w) \end{aligned}\right\} \tag{6.1}$$

Here, $F(\cdot|\beta)$ is a family of distributions parameterized by β and with corresponding densities $f(\cdot|\beta)$ with respect to some common measure. G is a random discrete measure over the parameter space. Hence, given G, the ν's are iid observations from an (infinite) mixture model based on mixing $F(\cdot|\beta)$ over different β's. The prior on G is the well-known Dirichlet Process, $DP(G_0, \alpha)$ with base measure G_0 and concentration parameter α. The resulting marginal distribution on the ν's is called a Dirichlet process mixture prior [9, 21, 31, 36, 37], which we denote as

$$\nu = (\nu_1, \ldots, \nu_W) \sim DPM(F, G_0, \alpha) \tag{6.2}$$

Each ν_w can be multivariate with continuous or discrete components.

One way to think about the nonparametric framework is to use a finite mixture model on the ν's, and to allow for variable number of groups, even including the possibility of infinite groups. Consider a finite mixture model with K clusters:

- Sample cluster parameters: $\phi_k \sim$ iid $G_0 \quad \forall\, k = 1, \ldots K$
- Sample cluster weights: $\boldsymbol{p} = (p_1, \ldots, p_K) \sim$ Dirichlet $\left(\frac{\alpha}{K}, \ldots, \frac{\alpha}{K}\right)$
- Draw $\nu_w \sim$ iid $F_0 \quad \forall\, w$ where F_0 is a mixture distribution over $F(\cdot|\phi_k)$'s with corresponding densities given by

$$f_0 = \sum_{k=1}^{K} p_k \; f(\cdot|\phi_k)$$

Letting $K \to \infty$ gives the DPM model described above in Equation (6.1).

The infinite mixture model representation of a DPM can be described directly by sampling p from a stick-breaking prior [36] as follows:

- Sample cluster parameters: $\phi_1, \phi_2, \ldots \sim$ iid G_0
- Generate $p_1, p_2, \ldots$ s.t. $\sum_{k=1}^{\infty} p_k = 1$ from the stick breaking prior with concentration parameter α:

$$\begin{aligned} q_1, q_2, \ldots &\sim \text{iid Beta}(1, \alpha) \\ p_k &= q_k \prod_{j=1}^{k-1} (1 - q_j) \quad \forall\, k = 1, 2, \ldots \end{aligned} \tag{6.3}$$

- Sample $\nu_w \sim$ iid $F_0 \quad \forall\, w$ where F_0 is a mixture distribution over all $F(\cdot|\phi_k)$'s with corresponding densities given by

$$f_0 = \sum_{k=1}^{\infty} p_k \; f(\cdot|\phi_k)$$

To sample the ν_w's from F_0 we can use latent variables z_w's that represent the cluster assignments[1] for each ν_w:

- Sample $z_w \sim$ iid Discrete $(p_1, p_2 \ldots) \quad \forall\ w$
- Draw $\nu_w \sim F(\cdot|\phi_{z_w})\ \forall\ w$

This can be represented in a plate notation as shown in Figure 6.1.

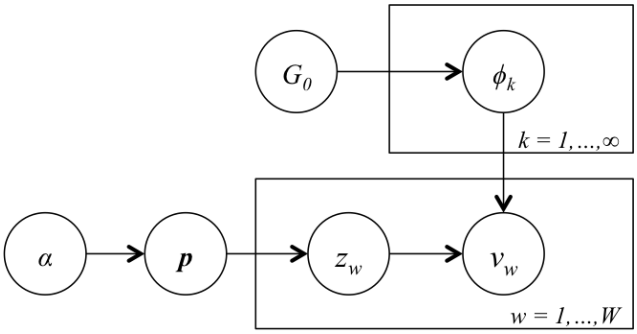


Figure 6.1: Plate notation for DPM sampling using latent cluster assignments

We use Gibbs sampling Algorithm 2 from [31], described below in Section 6.1.1, to sample the z's and ϕ's given the ν's. Let us use $z_{-w} = (z_v : 1 \leq v \leq W, v \neq w)$ to denote all the z's except for z_w and let $n_k = \#\{w = 1, \ldots, W : z_w = k\}$ represents number of elements in cluster k for each unique $k \in \{z_1, \ldots, z_W\}$. The conditional probabilities of z_w given all the other z_{-w} are [9, 31]

$$\mathbb{P}(z_w = k \mid z_{-w}) = \begin{cases} n_k/b & \text{for existing } k \in z_{-w} \\ \alpha/b & \text{for new } k \notin z_{-w} \end{cases}$$

where b is some normalizing constant.

This is the prior distribution for z_w used in the Gibbs sampling algorithm described below. z_w gets assigned to one of the existing clusters with probability proportional to the cluster

[1] We are brushing some details under the rug here; the z_w's do not actually refer to the cluster labels, but are rather just a convenient shorthand for talking about the clusters themselves.

size, or gets assigned to a new cluster with probability proportional to α. This is analogous to the Chinese restaurant process [9]. We can think of z_w as a new customer walking into a restaurant with infinite number of tables, each of which can seat infinite number of people. z_w then decides to join one of the existing tables based on their popularity (i.e. with probability proportional to the number of people already sitting at the tables) or decides to sit at a new tables with probability proportional to the concentration parameter α.

Notation:

α	:	Concentration parameter
G_0	:	Base distribution [e.g: Normal-Inverse Gamma $\mathcal{N}\Gamma^{-1}(a, b, c, \mu_0)$]
$\phi_1, \phi_2, \ldots$	:	Cluster parameters [e.g: $\phi_k = (\mu_k, \sigma_k^2) \sim G_0 = \mathcal{N}\Gamma^{-1}(a, b, c, \mu_0)$]
F	:	Distribution used to sample each ν_w [e.g: Gaussian $\mathcal{N}(\mu_k, \sigma_k^2)$]
$p_1, p_2, \ldots$	:	Cluster weights sampled from a Stick-breaking prior [36]
$z_1, z_2, \ldots$	:	Latent cluster assignment corresponding to each ν_w
z_{-w}	:	all the z's except for a particular z_w, $z_{-w} = (z_v : 1 \leq v \leq W, v \neq w)$
n_k	:	number of elements in cluster k

6.1.1 Gibbs Sampling for conjugate setting (Algorithm 2 from [31])

If there are W windows, i.e. $\nu = (\nu_1, \ldots, \nu_W)$, we can have a mixture of at most W distinct clusters. We use C to denote the number of clusters at any stage of the sampling, where $1 \leq C \leq W$. Corresponding to each cluster $k \in \{1, \ldots, C\}$ we have cluster parameters $\phi_k \in \boldsymbol{\phi} = (\phi_1, \ldots, \phi_C)$. The cluster assignment for each data ν_w is given by the vector $\boldsymbol{z} = (z_1, \ldots, z_w, \ldots, z_W)$, where each $z_w \in \{1, \ldots C\}$. Each ν_w is then drawn from the cluster z_w with parameter ϕ_{z_w}. Let the current state of the Markov chain be $\boldsymbol{z} = (z_1, \ldots, z_W)$, $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_C)$. In addition we keep track of number of observations in each distinct cluster given by $\boldsymbol{n} = (n_1, \ldots, n_C)$.

We use the notation $f(\nu_w|\phi_k)$ to denote the probability density of observation ν_w under

some distribution $F(\cdot|\phi_k)$ and we use $h(\cdot)$ to denote the posterior density under prior distribution G_0. For this algorithm, we need G_0 to be the conjugate prior corresponding to the distribution F.

- Sampling $\boldsymbol{z}$: For each $w = 1, \ldots, W$,
  - Set $n_{z_w} = n_{z_w} - 1$.
  - If z_w was the only observation in its cluster (i.e. $z_w \neq z_v \; \forall \; v \neq w$ and $n_{z_w} = 0$), then remove the corresponding ϕ_{z_w} from $\boldsymbol{\phi}$.
  - Draw a new value of z_w from the posterior given by the following distribution:

$$\mathbb{P}(z_w = k \mid z_{-w},\ \nu_w,\ \boldsymbol{\phi}) \propto \begin{cases} n_k \cdot f(\nu_w|\phi_k) & \text{for existing } k \in z_{-w} \\ \alpha \cdot \tilde{f}_{G_0}(\nu_w) & \text{for new } k \notin z_{-w} \end{cases}$$

    where $\tilde{f}_{G_0} = \int f(\nu_w|\phi) dG_0(\phi)$ is the prior-predictive density under the prior G_0.
  - If this sampled z_w is a new cluster assignment not associated with the other z_{-w}, i.e. $z_w \notin \boldsymbol{\phi}$, draw ϕ_{z_w} from the posterior distribution h under the prior G_0 and a single observation ν_w, and add it to $\boldsymbol{\phi}$.
- Sampling $\boldsymbol{\phi}$: For each $k = 1, \ldots, C$,
  - Draw a new value for ϕ_k from the posterior distribution h under the prior G_0 and all the ν_w's associated with cluster k (i.e all ν_w where $z_w = k$).

Using $\boldsymbol{\phi}$ and $\boldsymbol{z}$, the prior for each ν_w is given by $F(\cdot|\phi_{z_w})$.

Let us now see how we can use this DPM prior for our two-neuron zero-lag synchrony model.

6.2 Independent DPM priors on ν_1s and ν_2s

We use Metropolis-Hastings algorithm [7, p. 537-542] to sample ν_1, ν_2 and θ.

First, we assume ν_1 and ν_2 are drawn independently from the prior $DPM\left(F, G_0, \alpha\right)$ described above. For our DPM model, we take F to be a normal distribution with cluster parameters $\phi_k = (\mu_k, \sigma_k^2)$ and G_0 to be the conjugate prior for normal distribution, which is the normal-inverse-gamma $\mathcal{N}\Gamma^{-1}$ distribution with parameters (a, b, c, μ_0).[2] Thus the generative DPM model for $\nu \sim DPM\left(\mathcal{N},\ \mathcal{N}\Gamma^{-1}(a, b, c, \mu_0),\ \alpha\right)$ is given by:

$$\begin{aligned} \phi_1 = (\mu_1, \sigma_1^2),\ \phi_2 = (\mu_1, \sigma_1^2),\ \ldots &\sim \mathcal{N}\Gamma^{-1}(a, b, c, \mu_0) \\ p_1,\ p_2,\ \ldots &\sim \text{Stick-breaking prior}(\alpha) \\ z_1, \ldots, z_W &\sim \text{Discrete}(p_1, p_2, \ldots) \\ \nu_w &\sim \mathcal{N}\left(\mu_{z_w}, \sigma_{z_w}^2\right) \quad \forall\ w \end{aligned} \tag{6.4}$$

We use this distribution as the prior in the Metropolis-Hastings algorithm for both ν_1 and ν_2 (sampled independently):

$$\begin{aligned} \nu_1 = (\nu_{11}, \ldots, \nu_{1W}) &\sim DPM\left(\mathcal{N},\ \mathcal{N}\Gamma^{-1}(a, b, c, \mu_0),\ \alpha\right) \\ \nu_2 = (\nu_{21}, \ldots, \nu_{2W}) &\sim DPM\left(\mathcal{N},\ \mathcal{N}\Gamma^{-1}(a, b, c, \mu_0),\ \alpha\right). \end{aligned}$$

We use algorithm 2 from [31] (described above in Section 6.1.1) to sample cluster assignments $\boldsymbol{z^{(1)}}$ and $\boldsymbol{z^{(2)}}$ and cluster parameters $\boldsymbol{\phi^{(1)}}$ and $\boldsymbol{\phi^{(2)}}$ corresponding to ν_1 and ν_2 respectively. In our case where F is normal and G_0 is the conjugate $\mathcal{N}\Gamma^{-1}$ distribution, the posterior h is a $\mathcal{N}\Gamma^{-1}$ distribution as well, and the prior-predictive $\tilde{f}_{G_0}$ is a Student's t-distribution.

For θ we use the prior $\theta \sim \mathcal{N}(\mu_\theta, \sigma_\theta^2)$.

[2] A sample (μ, σ^2) from a normal-inverse-Gamma distribution $\mathcal{N}\Gamma^{-1}(a, b, c, \mu_0)$ can be obtained by first drawing σ^2 from an inverse-Gamma distribution $\sigma^2 \sim \Gamma^{-1}(a, b)$ and then μ from a normal distribution $\mu \sim \mathcal{N}\left(\mu_0, \frac{\sigma^2}{c}\right)$.

This model can be represented by the plate notation shown in Figure 6.2.

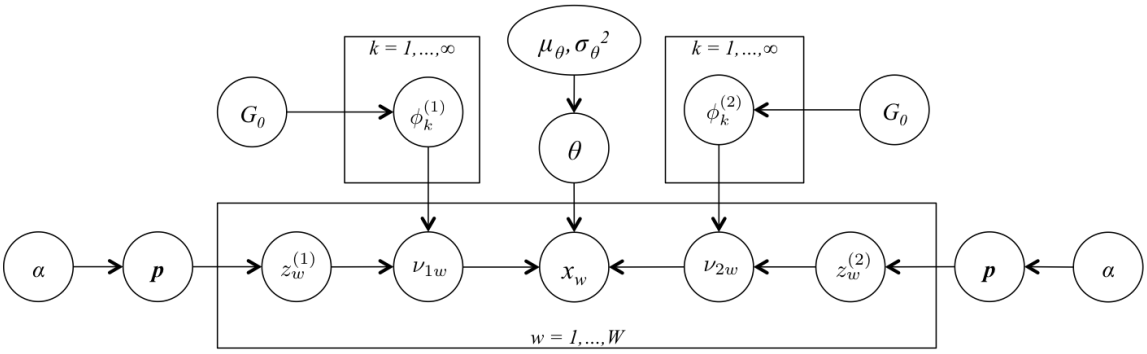


Figure 6.2: Plate notation for model with independent DPM priors

The overall algorithm steps are as follows:

- Sample $\boldsymbol{z^{(1)}}$ and $\boldsymbol{\phi^{(1)}}$ given the current ν_1 (Gibbs sampling from Section 6.1.1)
- Sample $\boldsymbol{z^{(2)}}$ and $\boldsymbol{\phi^{(2)}}$ given the current ν_2 (Gibbs sampling from Section 6.1.1)
- Sample ν_1's from the posterior distribution given the data x, and current sample values for θ, $\boldsymbol{z^{(1)}}$ and $\boldsymbol{\phi^{(1)}}$ (Metropolis-Hastings)
- Sample ν_2's from the posterior distribution given the data x, and current sample values for θ, $\boldsymbol{z^{(2)}}$ and $\boldsymbol{\phi^{(2)}}$ (Metropolis-Hastings)
- Sample θ from the posterior given the data x, and the current sample values of ν_1 and ν_2 using the $\mathcal{N}(\mu_\theta, \sigma_\theta^2)$ prior on θ (Metropolis-Hastings)

6.2.1 DPM results on data with uncorrelated inputs

For our simulations, we generated data using the true parameter $\theta^* = 2$. The ν_{1w}^*'s were set to be -1 or -2 randomly with equal probability, and the ν_{2w}^*'s were set to be -2 or -3 randomly with equal probability. Note that ν_1^* and ν_2^* are independent and not correlated with each other. The prior model parameters were taken to be $\mu_\theta = 0$, $\sigma_\theta = 2.5$, $a = 1$, $b = 0.1$, $c = 0.001$, $\mu_0 = 0$ and $\alpha = 2$. We took the window width to be $D = 4$ and varied

the recording time from $T = 500$ms to $T = 16000$ms. We generated 30 data samples for each value of T using the above mentioned true parameters, and inferred θ estimates using all the different techniques seen so far, namely stationary-MLE, non-stationary-MLE, CMLE and the Bayesian estimate using the nonparametric DPM prior. The relative mean squared error and squared bias are as shown in Figure 6.3.

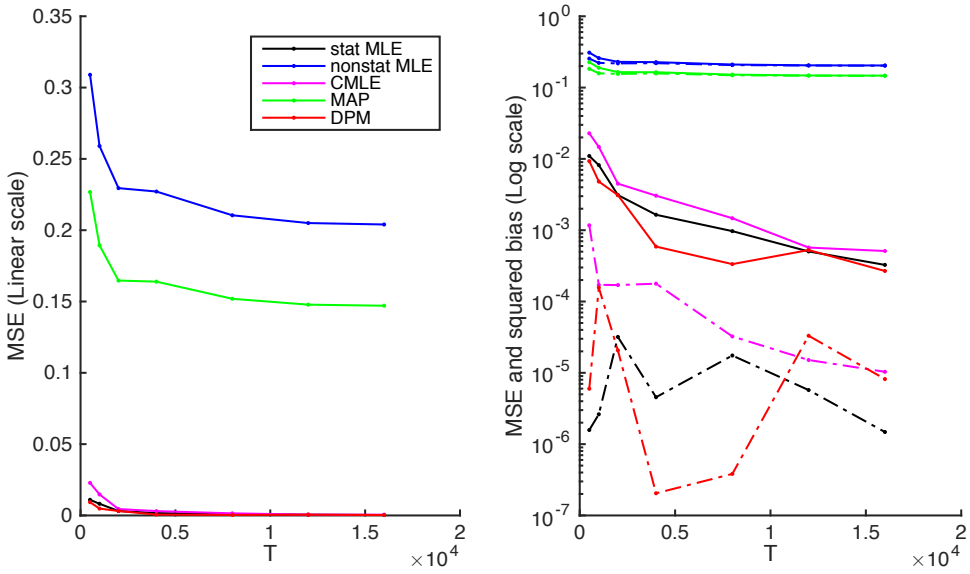


Figure 6.3: Relative MSE (solid) and squared bias (dotted) for different estimates on data generated from uncorrelated firing rates (ν's) : In this case, stationary-MLE, CMLE and DPM estimates appear to be consistent.

6.2.2 DPM results on data with correlated inputs

But now, if we use correlated ν_1^* and ν_2^* to generate the data, the model with the independent DPM priors no longer gives us consistent results.

We set $\nu_w^* = (\nu_{1w}^*, \nu_{2w}^*)$ to be (-1,-3) or (-2,-2) randomly with equal probability. At $T =$ 16000 ms, the results for different estimates on the data generated from these correlated inputs is shown in Table 6.1. Figure 6.4 shows the relative MSE and squared bias as T increases. Only the CMLE appears to be consistent.

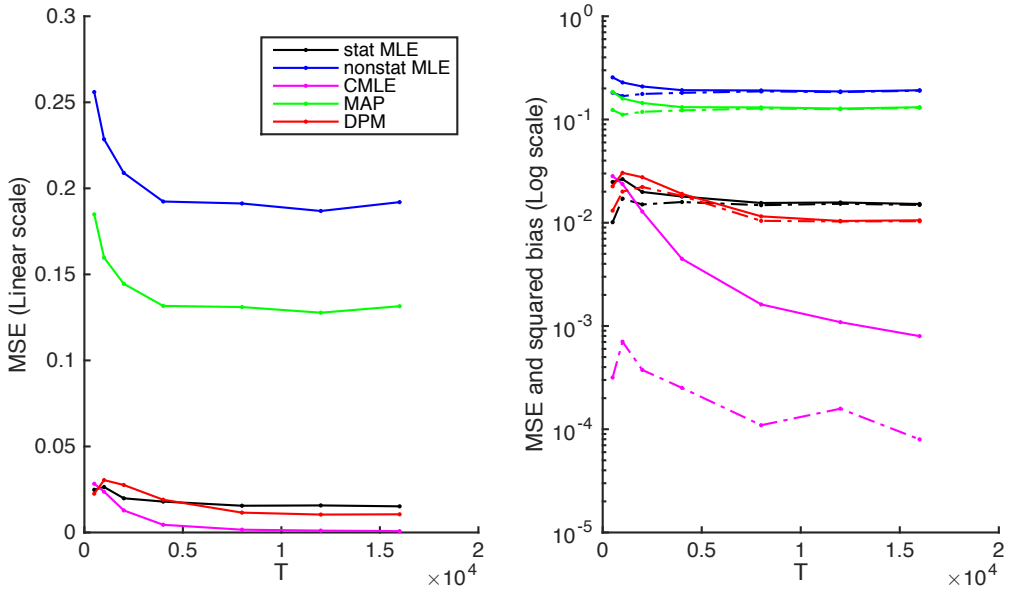


Figure 6.4: Relative MSE (solid) and squared bias (dotted) for different estimates on data generated from correlated firing rates (ν's) : Model with independent-DPM priors on ν_1 and ν_2 gives biased results when inputs are correlated. Only the CMLE appears to be consistent.

$\theta^* = 2$	stat-MLE	nonstat-MLE	CMLE	DPM
Mean	1.7561	2.8721	1.9821	1.7965
Std Dev	0.0364	0.0870	0.0545	0.0308
MSE	0.0608	0.7679	0.0032	0.0423
Rel-MSE	0.0152	0.1920	0.0008	0.0106

Table 6.1: Numerical results on data with correlated inputs at $T = 16000$ ms.

Our model assumes ν_1 and ν_2 are independent and cannot capture their correlation structure. Instead, it inaccurately attributes this correlation to the synchrony parameter θ. In our simulations, ν_1 and ν_2 are negatively correlated; assuming independence between them reduces the inferred value of θ. $\hat{\theta}$ has to account for the reduced correlation in the neuron spikes, which actually comes from the negatively correlated inputs. Hence the observed $\hat{\theta}$ from this DPM model is smaller than the true $\theta^* = 2$. If the inputs had a positive correlation, this model would compensate by estimating a larger value of $\hat{\theta}$.

6.3 Two-dimensional joint DPM prior on $\nu = (\nu_1, \nu_2)^T$

We need a model that can capture the correlation structure between the ν_1 and ν_2 inputs. We can achieve this by using a two-dimensional joint-DPM prior on the $\nu_w = (\nu_{1w}, \nu_{2w})^T$ simultaneously. We still use a similar framework as in (6.1), but now our distributions F and G_0 are two-dimensional distributions.

We can take F to be a multivariate-normal distribution MVN$(\boldsymbol{\mu}, \boldsymbol{\Lambda})$ with mean vector $\boldsymbol{\mu}$ and precision matrix $\boldsymbol{\Lambda}$. We take G_0 to be the conjugate distribution of the MVN distribution, which is the normal-Wishart distribution $\mathcal{NW}$ parameterized by $(a, \boldsymbol{\Lambda_0}, c, \boldsymbol{\mu_0})$.[3] Here, a and c are real numbers, $\boldsymbol{\Lambda}$ is a 2×2 matrix and $\boldsymbol{\mu_0}$ is a two-dimensional vector.

Using this F and G_0, our joint-DPM prior on the $\boldsymbol{\nu}_w = (\nu_{1w}, \nu_{2w})^T$s is given by

$$\begin{aligned} \boldsymbol{\phi}_1 = (\boldsymbol{\mu}_1, \boldsymbol{\Lambda}_1),\ \boldsymbol{\phi}_2 = (\boldsymbol{\mu}_2, \boldsymbol{\Lambda}_2),\ \ldots &\sim \mathcal{NW}(a, \boldsymbol{\Lambda_0}, c, \boldsymbol{\mu_0}) \\ p_1,\ p_2,\ \ldots &\sim \text{Stick-breaking prior}(\alpha) \\ z_1, \ldots, z_W &\sim \text{Discrete}(p_1, p_2, \ldots) \\ \boldsymbol{\nu}_w &\sim \text{MVN}\left(\boldsymbol{\mu}_{z_w}, \boldsymbol{\Lambda}_{z_w}\right) \quad \forall\ w \end{aligned} \tag{6.5}$$

We write this as $\boldsymbol{\nu} = (\nu_1, \nu_2)^T \sim \text{joint-}DPM\left(\text{MVN}(\cdot),\ \mathcal{NW}(a, \boldsymbol{\Lambda_0}, c, \boldsymbol{\mu_0}),\ \alpha\right)$.

We use cluster assignments $z = (z_1, \ldots, z_W)$ and cluster parameters $\boldsymbol{\phi} = (\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_C)$ as described in the algorithm in Section 6.1.1. When we had independent DPM priors on ν_1s and ν_2s, each had separate cluster assignment $\boldsymbol{z}^{(1)} = (z_1^{(1)}, \ldots, z_W^{(1)})$ and $\boldsymbol{z}^{(2)} = (z_1^{(2)}, \ldots, z_W^{(2)})$. In our joint setting using the two-dimensional DPM, our cluster assignments $\boldsymbol{z} = (z_1, \ldots, z_W)$ are common for both ν_1 and ν_2 and the cluster parameters $\boldsymbol{\phi}_k = (\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k)$'s are multivariate which helps to capture the correlation structure between ν_1 and ν_2. This model can be represented by the plate notation shown in Figure 6.5.

[3] A sample $(\boldsymbol{\mu}, \boldsymbol{\Lambda}) \sim \mathcal{NW}(a, \boldsymbol{\Lambda_0}, c, \boldsymbol{\mu_0})$ from a normal-Wishart distribution can be obtained by first drawing $\boldsymbol{\Lambda}$ from a Wishart distribution $\boldsymbol{\Lambda} \mid \boldsymbol{\Lambda_0}, a \sim \mathcal{W}(\boldsymbol{\Lambda_0}, a)$ and then $\boldsymbol{\mu}$ from a multivariate normal distribution $\boldsymbol{\mu} \mid \boldsymbol{\mu_0}, c, \boldsymbol{\Lambda} \sim MVN\left(\boldsymbol{\mu_0}, (c\boldsymbol{\Lambda})^{-1}\right)$.

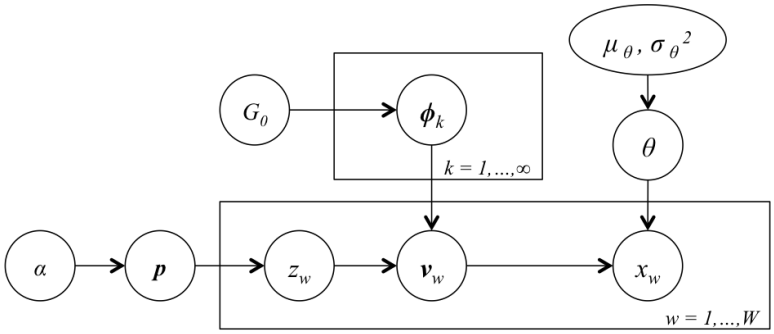


Figure 6.5: Plate notation for model with joint DPM prior

Since F is a multivariate normal distribution and G_0 is the conjugate $\mathcal{NW}$ distribution, the posterior h is a $\mathcal{NW}$ distribution as well, and the prior-predictive $\tilde{f}_{G_0}$ is a multivariate t-distribution.

The overall algorithm steps are as follows:

- Sample $\boldsymbol{z}$ and $\boldsymbol{\phi}$ given the current state $\boldsymbol{\nu} = (\nu_1, \nu_2)^T$ (Gibbs sampling algorithm from Section 6.1.1)
- Sample $\boldsymbol{\nu} = (\nu_1, \nu_2)^T$ from the posterior given x, and current value for θ, $\boldsymbol{z}$ and $\boldsymbol{\phi}$, where the prior for each $\boldsymbol{\nu}_w = (\nu_{1w}, \nu_{2w})^T$ is given by $F(\boldsymbol{\nu}_w|\boldsymbol{\phi}_{z_w}) = \text{MVN}(\boldsymbol{\nu}_w|\boldsymbol{\mu}_{z_w}, \boldsymbol{\Lambda}_{z_w})$ (Metropolis-Hastings)
- Sample θ from the posterior given the data x, and the current sample values of $\boldsymbol{\nu} = (\nu_1, \nu_2)^T$ where the prior on θ is $\mathcal{N}(\mu_\theta, \sigma_\theta^2)$ (Metropolis-Hastings)

6.3.1 Joint-DPM results on data with correlated inputs

When the inputs are correlated, assuming independent DPM priors on ν_1 and ν_2 gives biased results. However using a joint 2-D DPM prior on the ν's can capture the correlation structure and gives more accurate results. The θ estimates using the joint-DPM prior on ν seems to be consistent as seen from results in Table 6.2 and Figure 6.6. Both the CMLE and the Bayesian estimate using joint-DPM prior appear to be consistent.

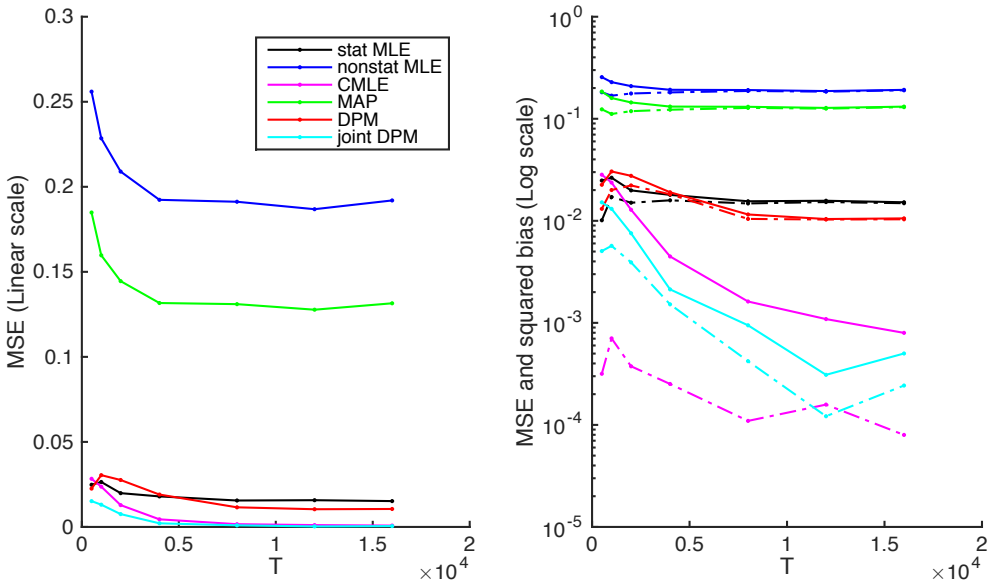


Figure 6.6: Relative MSE (solid) and squared bias (dotted) for different estimates on data generated from correlated firing rates (ν's) : CMLE and the joint-DPM estimates appear to be consistent, whereas independent-DPM gives biased results when inputs are correlated.

$\theta^* = 2$	stat-MLE	nonstat-MLE	CMLE	DPM	joint-DPM
Mean	1.7561	2.8721	1.9821	1.7965	1.9688
Std Dev	0.0364	0.0870	0.0545	0.0308	0.0339
MSE	0.0608	0.7679	0.0032	0.0423	0.0020
Rel-MSE	0.0152	0.1920	0.0008	0.0106	0.0005

Table 6.2: Numerical results on data with correlated inputs at $T = 16000$ ms.

6.3.2 Joint-DPM results on data with uncorrelated inputs

For data generated using uncorrelated inputs, Bayesian estimates using either independent-DPM or joint-DPM priors on the ν's appear to be consistent as seen in Figure 6.7.

A fully Bayesian technique using joint-DPM priors on the incidental parameters gives consistent results, however, it is computationally burdensome even for the two-neuron zero-synchrony case. Comparatively, the CMLE computation takes a fraction of the time, and seems to give similar results. As we scale to larger neurons and introduce lag-lead relationships in the model, the conditional model will scale much better as compared to a nonparametric model with a joint-DPM prior.

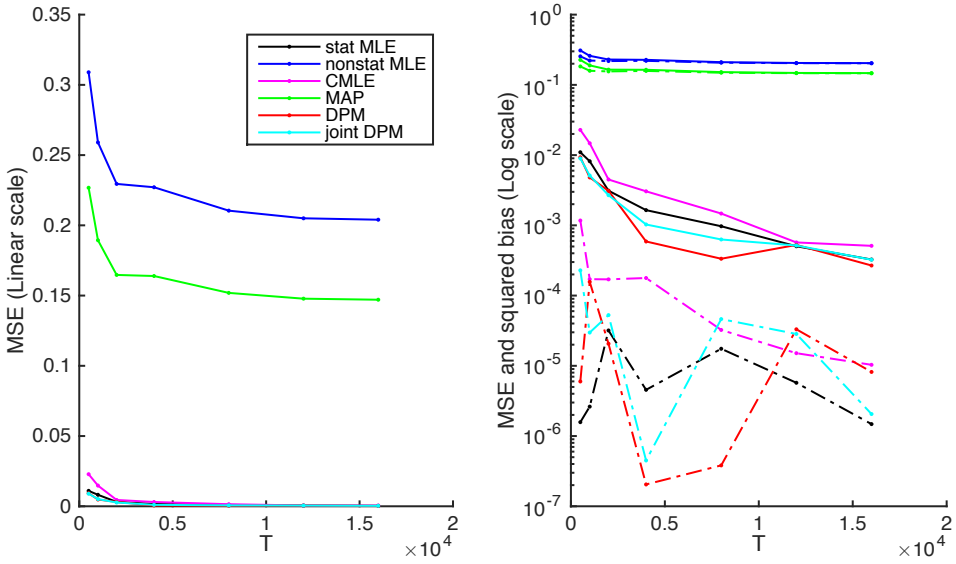


Figure 6.7: Relative MSE (solid) and squared bias (dotted) for different estimates on data generated from uncorrelated firing rates (ν's) : In this case stationary-MLE, CMLE, independent-DPM and joint-DPM appear to be consistent.

Chapter 7

Summary and future direction

Unconditional inference does not work well for realistic data with non-stationary background effects. Difficulties can arise from model misspecification or from incidental parameter problems. In the context of estimating parameters relating to fine-temporal spiking dynamics, conditional inference largely avoids these problems by conditioning away the coarse temporal dynamics. For larger networks, the two-neuron model can be used independently for each pair of neurons, or one can also use a higher-dimensional model to capture all pairwise effects simultaneously. Using a composite likelihood (or pseudo-likelihood) to approximate the conditional distribution will improve computational efficiency when working with larger networks or more complex models. A full Bayesian approach with non-parametric priors on the incidental parameters also seems to work well in this context, but is more computer intensive than the conditioning approach. To be able to use the Bayesian framework for larger networks, we would need to develop better sampling algorithms or explore other non-parametric approaches.

A natural extension for this work would be to develop the conditional inference framework for different and more complex models, for example generalized linear models with covariates. To extend this work to models that incorporate lag-lead relationships in addition to zero lag synchrony, we would need to further investigate the composite likelihood approach

or develop other similar approximations for the conditional likelihood. On the theoretical side, we proved the consistency of the CMLE for the two-neuron model by adapting the proof from [3]; next steps would be to try and adapt the asymptotic normality proof for our model, to extend the consistency proof for a higher dimensional setting, and to investigate the asymptotic properties of the composite likelihood method.

Appendix A

Results

Theorem A.1 (Strong Law of Large Numbers (Kolmogorov Theorem I) [33, p. 114])**.** *Let* $\{X_i\}, i = 1, 2, \ldots$ *be a sequence of independent random variables such that* $\mathbb{E}(X_i) = \mu_i$ *and* $\mathbb{V}(X_i) = \sigma_i^2$*. Then*

$$\sum_{i=1}^{\infty} \frac{\sigma_i^2}{i^2} < \infty \quad \Longrightarrow \quad \overline{X}_n - \overline{\mu}_n = \frac{1}{n}\left(\sum_{i=1}^{n}(X_i - \mu_i)\right) \xrightarrow{a.s} 0$$

i.e. the sequence obeys the Strong Law of Large Numbers.

Theorem A.2 (One-dimensional Delta Method [8, p. 492-493])**.** *If* $\hat{\mu}_n$ *is a real valued random variable and* μ *is a real valued parameter such that the asymptotic distribution of* $\hat{\mu}_n$ *is as follows:*

$$\sqrt{n}(\hat{\mu}_n - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2(\mu))$$

and if g *is a function differentiable at* μ*, then the asymptotic distribution of* $g(\hat{\mu}_n)$ *is given by:*

$$\sqrt{n}\left(g(\hat{\mu}_n) - g(\mu)\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2(\mu)(g'(\mu))^2\right).$$

Theorem A.3 (Multivariate Delta Method [8, p. 492-493])**.** *Let* $\hat{\boldsymbol{\mu}}_{\boldsymbol{n}}$ *be a* k*-dimensional random vector* $\hat{\boldsymbol{\mu}}_{\boldsymbol{n}} = (\hat{\mu}_{n1}, \ldots, \hat{\mu}_{nk})$ *and let* $\boldsymbol{\mu}$ *be a* k*-dimensional vector parameter:* $\boldsymbol{\mu} =$

$(\mu_1, \ldots, \mu_k)$. *Let* $\hat{\boldsymbol{\mu}}_{\boldsymbol{n}}$ *have an asymptotic normal distribution*

$$\sqrt{n}(\hat{\boldsymbol{\mu}}_{\boldsymbol{n}} - \boldsymbol{\mu}) \xrightarrow{d} \mathcal{N}(0, \boldsymbol{\Sigma}(\boldsymbol{\mu}))$$

where $\boldsymbol{\Sigma}(\boldsymbol{\mu})$ *is the* $k \times k$ *asymptotic covariance matrix of* $\hat{\boldsymbol{\mu}}_{\boldsymbol{n}}$. *Let* $\mathbf{g}$ *be a function defined on a open subset of* k*-dimensional space, taking values in* R*-dimensional space:* $\boldsymbol{g}(\boldsymbol{\mu}) = (g_1(\boldsymbol{\mu}), \ldots, g_R(\boldsymbol{\mu}))$ *that is differentiable at* $\boldsymbol{\mu}$. *Then the asymptotic distribution of* $\boldsymbol{g}(\hat{\boldsymbol{\mu}}_{\boldsymbol{n}})$ *is given by:*

$$\sqrt{n}\left(\boldsymbol{g}(\hat{\boldsymbol{\mu}}_{\boldsymbol{n}}) - \boldsymbol{g}(\boldsymbol{\mu})\right) \xrightarrow{d} \mathcal{N}\left(0, (\nabla \boldsymbol{g})' \boldsymbol{\Sigma} (\nabla \boldsymbol{g})\right)$$

where

$$(\nabla \boldsymbol{g})' \boldsymbol{\Sigma} (\nabla \boldsymbol{g}) = \sum_{i=1}^{k} \sum_{j=1}^{k} \Sigma_{ij}(\boldsymbol{\mu}) \frac{\partial \mathbf{g}(\boldsymbol{\mu})}{\partial \mu_i} \frac{\partial \mathbf{g}(\boldsymbol{\mu})}{\partial \mu_j}$$

Appendix B

Neyman-Scott problem for Gaussian distribution

Let us focus on just one-dimensional data coming from a Gaussian probability density function f. Let $X = (X_1, \ldots, X_T)$ be independent Gaussian random variables with common variance θ and means $\mu = (\mu_1, \ldots, \mu_T)$, so that

$$\begin{aligned} X_t \mid \mu, \theta &\sim \mathcal{N}(\mu_t, \theta) \qquad \forall\ t = 1 \ldots, T \\ f_{\mu,\theta}(x) &= \prod_{t=1}^{T} \frac{1}{\sqrt{2\pi\theta}} \exp\left(-\frac{(x_t - \mu_t)^2}{2\theta}\right). \end{aligned}$$

Further suppose also that μ is slow varying and can be taken to be piecewise constant. Partitioning the data into windows w of length D, where wth window is $\{(w-1)D + 1, \ldots, wD\}$, we denote the common value of μ_t in the w-th window to be ν_w. For simplicity, let us assume that T is a multiple of D and let $W = T/D$.

$$\mu_t = \nu_w \qquad \forall\ (w-1)D + 1 \leq t \leq wD, \quad \forall\ w = 1, \ldots, W.$$

This re-parameterization gives

$$X_t \mid \nu, \theta \sim \mathcal{N}(\nu_w, \theta) \qquad \forall\ (w-1)D + 1 \leq t \leq wD, \quad \forall\ w = 1, \ldots, W.$$

We re-index the data with respect to which window they are in and denote X_t by X_{wd} where $w = \lceil t/D \rceil$ and $d = t - (w-1)D$. Hence,

$$X_{wd} \mid \nu, \theta \sim \mathcal{N}(\nu_w, \theta) \qquad \forall\ d = 1, \ldots, D \text{ and } \ w = 1, \ldots, W.$$

This is the classical Neyman-Scott problem from the statistics literature [32]. This can also be written as

$$X_w = \begin{bmatrix} X_{w1} \\ \vdots \\ X_{wD} \end{bmatrix} \mid (\nu, \theta) \sim \mathcal{N}\left(\nu_w\ u_D,\ \theta\ I_D\right) \quad \forall\ w = 1, \ldots, W, \tag{B.1}$$

where I_D is the D-dimensional identity matrix and u_D is a $D \times 1$ vector of all ones.

B.1 Maximum likelihood estimation

The maximum likelihood estimate (we shall denote this as the *MLE* from here onwards) for this is given by

$$\begin{aligned} \hat{\nu}_w^{MLE} &= \frac{1}{D}\sum_{d=1}^{D} x_{wd} \qquad\qquad \forall\ w = 1, \ldots, W \\ \hat{\theta}^{MLE} &= \frac{1}{WD}\sum_{w=1}^{W}\sum_{d=1}^{D} x_{wd}^2 - \frac{1}{WD^2}\sum_{w=1}^{W}\left(\sum_{d=1}^{D} x_{wd}\right)^2 = \frac{DM_2 - M_1}{D} \qquad \text{(B.2)} \\ &= \frac{1}{WD}\sum_{w=1}^{W}\sum_{d=1}^{D} x_{wd}^2 - \frac{1}{WD^2}\sum_{w=1}^{W} s_w^2 \\ &= \frac{1}{WD}\sum_{w=1}^{W}\sum_{d=1}^{D}\left(x_{wd} - \frac{s_w}{D}\right)^2, \qquad \text{(B.3)} \end{aligned}$$

where $s_w \triangleq \sum_d (x_{wd})^2$ and

$$M_2 = \frac{1}{WD}\sum_{w=1}^{W}\sum_{d=1}^{D} x_{wd}^2 \qquad \text{and} \qquad M_1 = \frac{1}{WD}\sum_{w=1}^{W}\left(\sum_{d=1}^{D} x_{wd}\right)^2 = \frac{1}{WD}\sum_{w=1}^{W} s_w^2. \quad \text{(B.4)}$$

Proposition 3. *Suppose that the data was generated with true parameters* $\nu^* = (\nu_w^* : w = 1\ldots, W)$ *and* θ^* *as follows*

$$X_{wd} \mid \nu^*, \theta^* \sim \mathcal{N}(\nu_w^*, \theta^*) \qquad \forall\ d = 1, \ldots, D \quad \textit{and} \quad w = 1, \ldots, W. \quad \text{(B.5)}$$

Then the MLE given by Equation (2.1) *is inconsistent.*

Proof. From (B.5), we get $\mathbb{E}[X_{wd}^2] = {\nu_w^*}^2 + \theta^* \ \forall\ w, d$ and $\mathbb{E}[X_{wd_1}X_{wd_2}] = {\nu_w^*}^2 \ \forall\ w,\ d_1 \neq d_2$.

Let $S_w(X) \triangleq \sum_d X_{wd}$, and hence

$$\begin{aligned}
\mathbb{E}\left[\left(X_{wd} - \frac{S_w}{D}\right)^2\right] &= \mathbb{E}[X_{wd}^2] - \frac{1}{D^2}\,\mathbb{E}[S_w^2] \\
&= \mathbb{E}[X_{wd}^2] - \frac{1}{D^2}\,\mathbb{E}\left[\sum_{d=1}^{D} X_{wd}^2 + \sum_{\substack{d_1 \\ d_1 \neq d_2}}\sum_{d_2} X_{wd_1}X_{wd_2}\right] \\
&= (\nu_w^* + \theta^*) - \frac{1}{D^2}\,\left[D(\nu_w^* + \theta^*) + D(D-1)\nu_w^*\right] \\
&= \frac{D-1}{D}\theta^* \qquad \forall\ w, d.
\end{aligned}$$

Using the above equation and (B.3) we get

$$\hat{\theta}^{MLE} = \frac{1}{WD}\sum_{w=1}^{W}\sum_{d=1}^{D}\left(x_{wd} - \frac{s_w}{D}\right)^2 \longrightarrow \mathbb{E}\left[\left(X_{wd} - \frac{S_w}{D}\right)^2\right] = \frac{D-1}{D}\theta^* \quad \text{(B.6)}$$

and

$$DM_2 - M1 \longrightarrow (D-1)\theta^*. \quad \text{(B.7)}$$

□

Standard MLE gives inconsistent results since we are trying to infer large number of incidental parameters, each from a finite amount of data. For a general window of width D, each ν_w is inferred from only the corresponding D entries in each window. Since these estimates are used to calculate the MLE of θ, it leads to inconsistent results for $\hat{\theta}^{\text{MLE}}$. This is known as the *incidental parameter problem*. This problem has been studied for a long time in statistics and econometric literature and the two main approaches suggested to tackle this problem are conditioning and marginalizing. Let us look at both these approaches with respect to the Neyman-Scott problem.

B.2 Conditioning

The conditional approach involves finding some statistics of the data such that when you condition on these statistics, the remaining conditional likelihood depends only on the structural parameters θ. Conditioning on statistics sufficient for the incidental parameters ν ensures that the conditional likelihood has no ν dependence. One can then infer the structural parameters from the conditional likelihood.

The sum statistic $S_w(X) \triangleq \sum_d X_{wd}$ is sufficient for incidental parameters ν_w. We condition on the observed values of the sum statistics $s_w = S_w(x) = \sum_d x_{wd}$. If f is the density function for X, we can write $f(x) = g(S(x)) \cdot \psi(x|S(x))$ where g is the marginal density of $S(X)$ and ψ is the conditional density. The conditional MLE $\hat{\theta}^{CMLE}$ is given by maximizing the log-conditional-likelihood $\mathcal{L}^{\text{conditional}}(\theta) = \log \psi_\theta(x \mid S_w = s_w \ \forall\ w)$.

Proposition 4. *Suppose that the data was generated with true parameters* $\nu^* = (\nu_w^* : w = 1 \ldots, W)$ *and* θ^* *as described in* (B.5) *. Then the CMLE is consistent:* $\hat{\theta}^{CMLE} \to \theta^*$.

Proof. Using our earlier notation from (B.1), we have $X_w \mid (\nu, \theta) \sim \mathcal{N}(\nu_w\ u_D,\ \theta\ I_D) \quad \forall\ w$ where I_D is the D-dimensional identity matrix and u_D is a $D \times 1$ vector of all ones.

Let us consider the transformation

$$Z_w = \left[\begin{array}{c} \vdots \\ Z_{wa} \\ \vdots \\ \hline Z_{wb} \end{array}\right] \triangleq \left[\begin{array}{c} X_{w1} \\ \vdots \\ X_{w(D-1)} \\ S_w \end{array}\right] = \left[\begin{array}{c|c} I_{(D-1)} & \vec{0} \\ \hline u^T_{(D-1)} & 1 \end{array}\right] X_w = AX_w,$$

where $I_{(D-1)}$ is the $(D-1)$-dimensional identity matrix and $u_{(D-1)}$ is a $(D-1)\times 1$ vector of all ones. For this section, let us drop the subscripts and use the notation I and u to denote $I_{(D-1)}$ and $u_{(D-1)}$.

Now if $X_w \sim \mathcal{N}(\mu_w, \Sigma)$ then we have $Z_w = AX_w \sim \mathcal{N}(A\mu_w, A\Sigma A^T)$, with

$$\begin{aligned} A\Sigma A^T &= \theta \left[\begin{array}{c|c} I & \vec{0} \\ \hline u^T & 1 \end{array}\right]\left[\begin{array}{c|c} I & u \\ \hline \vec{0} & 1 \end{array}\right] = \theta \left[\begin{array}{c|c} I & u \\ \hline u^T & D \end{array}\right] = \left[\begin{array}{cc} \Sigma_{aa} & \Sigma_{ab} \\ {\Sigma_{ab}}^T & \Sigma_{bb} \end{array}\right] \\ A\mu_w &= \nu_w \left[\begin{array}{c|c} I & \vec{0} \\ \hline u^T & 1 \end{array}\right] u_D = \nu_w \left[\begin{array}{c} u \\ \hline 2D-1 \end{array}\right] = \left[\begin{array}{c} \mu_{wa} \\ \mu_{wb} \end{array}\right], \end{aligned}$$

where $\mu_{wa} = \nu_w u$, $\mu_{wb} = (2D-1)\nu_w$ and $\Sigma_{aa} = \theta I$, $\Sigma_{ab} = \theta u$, $\Sigma_{bb} = \theta D$.

Thus we have

$$\begin{aligned} Z_w = \left[\begin{array}{c} Z_{wa} \\ Z_{wb} \end{array}\right] &\sim \mathcal{N}\left(\left[\begin{array}{c} \mu_{wa} \\ \mu_{wb} \end{array}\right], \left[\begin{array}{cc} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ab}^T & \Sigma_{bb} \end{array}\right]\right) \\ &\sim \mathcal{N}\left(\nu_w \left[\begin{array}{c} u \\ D \end{array}\right], \theta \left[\begin{array}{cc} I & u \\ u^T & D \end{array}\right]\right). \end{aligned}$$

The conditional distribution is then given by

$$Z_{wa}|(Z_{wb} = s_w) \sim \mathcal{N}(\mu_{w,a|b}, \Sigma_{a|b}),$$

where

$$\begin{aligned}
\mu_{w,a|b} &= \mu_{wa} + \Sigma_{ab} \cdot \Sigma_{bb}^{-1} \cdot (z_{wb} - \mu_b) = \nu_w u + \theta u \frac{1}{\theta D}(s_w - D\nu_w) = \frac{1}{D} s_w \; u \\
\Sigma_{a|b} &= \Sigma_{aa} - \Sigma_{ab} \cdot \Sigma_{bb}^{-1} \cdot \Sigma_{ba} = \left(\theta I - \theta u \frac{1}{\theta D} \theta u^T\right) = \theta \left(I - \frac{1}{D} u u^T\right).
\end{aligned}$$

The conditional log-likelihood is

$$\begin{aligned}
\mathcal{L}^{\text{conditional}}(\theta) &= \log \psi_\theta(x \mid S_w = s_w \; \forall \; w) = \sum_{w=1}^{W} \log \Phi(z_{wa} \mid \mu_{w,a|b}, \Sigma_{a|b} \;) \\
&= -\sum_{w=1}^{W} \left(\frac{D-1}{2} \log 2\pi + \frac{1}{2} \log |\Sigma_{a|b}| + \frac{1}{2} (z_{wa} - \mu_{w,a|b})^T \Sigma_{a|b}^{-1} (z_{wa} - \mu_{w,a|b}) \right),
\end{aligned}$$

where Φ is the density for the multivariate normal distribution. Using the Sherman–Morrison formula for matrix inversion and matrix determinant lemma, we get

$$\begin{aligned}
|\Sigma_{a|b}| &= \theta^{D-1} \left(1 - \frac{1}{D} u^T u\right) = \frac{1}{D} \theta^{D-1} \\
\Sigma_{a|b}^{-1} &= \frac{1}{\theta} \left(I + \frac{\frac{1}{D}}{1 - \frac{1}{D} u^T u} u u^T \right) \\
&= \frac{1}{\theta} \left(I + u u^T \right).
\end{aligned}$$

Hence the conditional log-likelihood to be minimized is

$$\begin{aligned}
\mathcal{L}^{\text{conditional}}(\theta) &= -\sum_{w=1}^{W} \left(\frac{D-1}{2} \log 2\pi + \frac{1}{2} \log |\Sigma_{a|b}| + \frac{1}{2} (z_{wa} - \mu_{w,a|b})^T \Sigma_{a|b}^{-1} (z_{wa} - \mu_{w,a|b}) \right) \\
&= -\frac{W(D-1)}{2} \log \theta - \frac{1}{2\theta} \sum_{w} \left(z_{wa} - \frac{s_w u}{D} \right)^T (I + u u^T) \left(z_{wa} - \frac{s_w u}{D} \right) - C,
\end{aligned}$$

where C is a constant.

The $\hat{\theta}^{CMLE}$ that maximizes this conditional likelihood is given by

$$\begin{aligned}
\hat{\theta}^{CMLE} &= \frac{1}{W(D-1)} \sum_{w} \left(z_{wa} - \frac{s_w u}{D} \right)^T (I + uu^T) \left(z_{wa} - \frac{s_w u}{D} \right) \\
&= \frac{1}{W(D-1)} \sum_{w=1}^{W} \left[\sum_{d=1}^{D-1} \left(x_{wd} - \frac{s_w}{D} \right)^2 + \left(\sum_{d=1}^{D-1} \left(x_{wd} - \frac{s_w}{D} \right) \right)^2 \right] \\
&\vdots \\
&= \frac{1}{W(D-1)D^2} \sum_{w=1}^{W} \left[D^2 \sum_{d=1}^{D} x_{wd}^2 - D s_w^2 \right] \\
&= \frac{1}{WD(D-1)} \sum_{w=1}^{W} \left[D \sum_{d=1}^{D} x_{wd}^2 - \left(\sum_{d=1}^{D} x_{wd} \right)^2 \right] \\
&= \frac{1}{(D-1)} \left(DM_2 - M_1 \right),
\end{aligned}$$

where M_1 and M_2 are are defined in Equation (B.4). Using (B.7), we get

$$\hat{\theta}^{CMLE} = \frac{DM_2 - M_1}{D-1} \longrightarrow \theta^*. \tag{B.8}$$

Thus, conditional inference gives consistent estimates. □

B.3 Marginalization

B.3.1 Using a (non-informative) uniform prior on ν

A prior often suggested is a non-informative uniform prior. Let us use a uniform prior on the ν's in the Neyman-Scott problem. The likelihood is

$$f_{\nu,\theta}(x) = \prod_{w=1}^{W} \prod_{d=1}^{D} \frac{1}{2\pi\theta} \exp\left(-\frac{(x_{wd} - \nu_w)^2}{2\theta} \right).$$

Integrating each ν_w over $\mathbb{R}$ with respect to an improper uniform prior gives

$$\begin{aligned} f_\theta^{marginal}(x) &= \prod_{w=1}^{W} \int \prod_{d=1}^{D} \frac{1}{\sqrt{2\pi\theta}} \exp\left(-\frac{(x_{wd}-\nu_w)^2}{2\theta}\right) d\nu_w \\ &= \prod_{w=1}^{W} \frac{1}{(2\pi\theta)^{\frac{D}{2}}} \int \exp\left(-\frac{1}{2\theta}\sum_d (x_{wd}-\nu_w)^2\right) d\nu_w \\ &= \prod_{w=1}^{W} \frac{1}{(2\pi\theta)^{\frac{D}{2}}} \int \exp\left(-\frac{1}{2\theta}\sum_d \left((x_{wd})^2 - 2x_{wd}\nu_w + \nu_w^2\right)\right) d\nu_w \end{aligned}$$

Let $r_w = \sum_d (x_{wd})^2$ and $s_w = \sum_d x_{wd}$

$$\begin{aligned} f_\theta^{marginal}(x) &= \prod_{w=1}^{W} \frac{1}{(2\pi\theta)^{\frac{D}{2}}} \int \exp\left(-\frac{r_w - 2s_w\nu_w + D\nu_w^2}{2\theta}\right) d\nu_w \\ &= \prod_{w=1}^{W} \frac{1}{(2\pi\theta)^{\frac{D}{2}}} \exp\left(-\frac{r_w}{2\theta} + \frac{s_w^2}{2\theta D}\right) \int \exp\left(-\frac{D}{2\theta}\left(\nu_w - \frac{s_w}{D}\right)^2\right) d\nu_w \\ &= \prod_{w=1}^{W} \frac{1}{(2\pi\theta)^{\frac{D}{2}}} \exp\left(-\frac{r_w}{2\theta} + \frac{s_w^2}{2\theta D}\right) \left(\frac{2\pi\theta}{D}\right)^{\frac{1}{2}} \\ &\vdots \\ &= \prod_{w=1}^{W} \frac{1}{(2\pi\theta)^{\frac{D-1}{2}}\sqrt{D}} \exp\left(-\frac{1}{2\theta}\sum_d \left(x_{wd} - \frac{s_w}{D}\right)^2\right). \end{aligned} \tag{B.9}$$

Since we get this likelihood over θ by integrating the ν's out, this is known as the integrated likelihood [6] or the marginal likelihood. Now we maximize the this integrated/marginal likelihood to get the marginal maximum likelihood estimate (we call it MMLE).

Instead of maximizing the marginal-likelihood, we can maximize the log of the marginal-likelihood given by

$$\mathcal{L}^{\text{marginal}}(\theta) = \log f_\theta^{marginal}(x) = -\frac{W(D-1)}{2}\log\theta - \frac{1}{2\theta}\sum_{w=1}^{W}\left(\sum_d \left(x_{wd} - \frac{s_w}{D}\right)^2\right) + C,$$

where C is a constant. The maximum likelihood estimate that maximizes this log-marginal-

likelihood is given by

$$\begin{aligned}\hat{\theta}^{MMLE,unif} &= \frac{1}{W(D-1)}\sum_{w=1}^{W}\left(\sum_{d}\left(x_{wd}-\frac{s_w}{D}\right)^2\right)\\ &= \frac{1}{W(D-1)}\sum_{w=1}^{W}\left(\sum_{d}x_{wd}^2-\frac{\left(\sum_d x_{wd}\right)^2}{D}\right)\\ &= \frac{1}{(D-1)}\left(DM_2-M_1\right). \qquad \text{(B.10)}\end{aligned}$$

Note that this is the same as the CMLE in Equation (B.8) and hence

$$\hat{\theta}^{MMLE,unif} \longrightarrow \theta^*.$$

Thus for the Neyman-Scott problem, conditioning and marginalizing (using a uniform prior) approaches lead to the same inference.

B.3.2 Using a (non-conjugate) Gaussian prior on ν

Now let us try to integrate out ν with some other prior. Let us assume the following prior:

$$\nu_w \mid c \sim \mathcal{N}(0,c) \quad \forall\ w.$$

Our original model was

$$X_{wd} \mid \nu,\theta \sim \mathcal{N}(\nu_w,\theta) \quad \forall \quad d=1,2,\ldots,D, \quad w=1,\ldots,W.$$

Then we can write

$$X_{wd} = \nu_w + Z_{wd} \quad \forall d,\ \forall\ w,$$

where each Z_{wd} iid $\sim \mathcal{N}(0,\theta)$ $\forall w,\ \forall d$ and the Z's are independent from the ν's. Hence $\forall$w

$$\begin{aligned}
\mathbb{E}[X_w] &= \mathbb{E}\nu_w + \mathbb{E}Z_{wd} = 0 \\
\mathbb{E}[X_{wd}^2] &= \mathbb{E}\nu_w^2 + \mathbb{E}Z_{wd}^2 + \mathbb{E}\nu_w Z_{wd} = c + \theta \\
\mathbb{E}[X_{wd_1}X_{wd_2}] &= \mathbb{E}\nu_w^2 + \mathbb{E}\nu_w Z_{wd_1} + \mathbb{E}\nu_w Z_{wd_2} + \mathbb{E}Z_{wd_1}Z_{wd_2} = c \qquad \forall\ d_1 \neq d_2.
\end{aligned}$$

Thus we get the model

$$X_w \mid (\theta, c) \sim \mathcal{N}\left(\vec{0}, Q\right) \quad \forall\ w = 1, \ldots, W, \tag{B.11}$$

where

$$X_w = \begin{bmatrix} X_{w1} \\ X_{w2} \\ \vdots \\ X_{wD} \end{bmatrix}, \qquad Q = \begin{bmatrix} \theta + c & c & \ldots & c \\ c & \theta + c & & \vdots \\ \vdots & & \ddots & c \\ c & \ldots & c & \theta + c \end{bmatrix}.$$

Log-marginal-likelihood is given by

$$\begin{aligned}
\mathcal{L}_c^{\text{marginal}}(\theta) = \log f_{\theta,c}(x) &= -\frac{WD}{2}\log 2\pi - \frac{W}{2}\log|Q| - \frac{1}{2}\sum_{w=1}^{W} x_w^T Q^{-1} x_w \\
-\frac{2}{W}\mathcal{L}_c^{\text{marginal}}(\theta) &= D\log 2\pi + \log|Q| + \frac{1}{W}\sum_{w=1}^{W} x_w^T Q^{-1} x_w.
\end{aligned}$$

We can write the log-likelihood explicitly in terms of c and θ by using the fact that $Q = \theta I + cuu^T$ where I is the D-dimensional identity matrix and u is a $D \times 1$ vector of all ones. Using the Sherman–Morrison formula for matrix inversion and the Matrix Determinant lemma, we get

$$\begin{aligned}
|Q| &= \theta^{D-1}(\theta + cD) \\
Q^{-1} &= \frac{1}{\theta}\left(I - \frac{c}{\theta + cD}uu^T\right).
\end{aligned}$$

This gives

$$
\begin{aligned}
-\frac{2}{W}\mathcal{L}_c^{\text{marginal}}(\theta) &= D\log 2\pi + \log(\theta^{(D-1)}(\theta+cD)) + \frac{1}{\theta}\left(\frac{1}{W}\sum_{w=1}^{W} x_w^T\left(I - \frac{c}{\theta+cD}uu^T\right)x_w\right) \\
&= D\log 2\pi + (D-1)\log\theta + \log(\theta+cD) + \frac{1}{\theta}\frac{1}{W}\left(N_2 - \frac{c}{\theta+cD}N_1\right) \quad \text{(B.12)} \\
&= D\log 2\pi + (D-1)\log\theta + \log(\theta+cD) + \frac{M_2 D}{\theta} - \frac{M_1 cD}{\theta(\theta+cD)} \quad \text{(B.13)} \\
&= D\log 2\pi + (D-1)\log\theta + \log(\theta+cD) + \frac{M_2 D}{\theta} - \frac{M_1}{\theta} + \frac{M_1}{\theta+cD}, \quad \text{(B.14)}
\end{aligned}
$$

where

$$
\begin{aligned}
M_2 &= \frac{1}{WD}\sum_{w=1}^{W} {x_w}^T x_w = \frac{1}{WD}\sum_{w=1}^{W}\sum_{d=1}^{D} x_{wd}^2 \quad \text{(B.15)} \\
M_1 &= \frac{1}{WD}\sum_{w=1}^{W} {x_w}^T uu^T x_w = \frac{1}{WD}\sum_{w=1}^{W}\left(\sum_{d=1}^{D} x_{wd}\right)^2 \\
&= \frac{1}{WD}\sum_{w=1}^{W}\left(\sum_{d=1}^{D} x_{wd}^2 + \sum_{\substack{d_1 \\ d_1\neq d_2}}\sum_{d_2} x_{wd_1}x_{wd_2}\right). \quad \text{(B.16)}
\end{aligned}
$$

Let the true prior on ν_w be $\nu_w \sim \mathcal{N}(0,c^*) \quad \forall\ w$ which gives that the true distribution of the data is

$$
X_w \mid (\theta^*, c^*) \ \sim\ \mathcal{N}\left(\vec{0}, Q^*\right) \quad \forall\ w = 1,\ldots,W, \quad \text{(B.17)}
$$

where

$$
\begin{aligned}
Q^* = \mathbb{E}[X_w X_w^T] &= \begin{bmatrix} \theta^*+c^* & c^* & \ldots & c^* \\ c^* & \theta^*+c^* & & \vdots \\ \vdots & & \ddots & c^* \\ c^* & \ldots & c^* & \theta^*+c^* \end{bmatrix} \\
\mathbb{E}[X_{wd}^2] &= \theta^*+c^* \qquad \forall\ w,d \\
\mathbb{E}[X_{wd_1}X_{wd_2}] &= c^* \qquad \forall\ w, d_1 \neq d_2.
\end{aligned}
$$

Using this in equations (B.15) and (B.16) gives

$$\left.\begin{aligned} M_1 &\longrightarrow \frac{1}{D}\left(\sum_{d=1}^{D}\mathbb{E}[X_{wd}^2] + \sum_{\substack{d_1 \\ d_1\neq d_2}}\sum_{d_2}\mathbb{E}[X_{wd_1}X_{wd_2}]\right) = \theta^* + c^*D \\ M_2 &\longrightarrow \frac{1}{D}\sum_{d=1}^{D}\mathbb{E}[X_{wd}^2] = \theta^* + c^*. \end{aligned}\right\} \tag{B.18}$$

MMLE

If we treat the problem as estimating θ with hyper-parameter c assumed to be a known constant, then we will prove (see Proposition 5 below) that $\hat{\theta}^{MMLE} \to \theta^*$ if and only if $c = c^*$.

With c fixed, the MMLE is given by the argument that maximizes Equation (B.13), which is the solution to the following equation

$$\begin{aligned} 0 &= \frac{\partial}{\partial\theta}\left(D\log 2\pi + (D-1)\log\theta + \log(\theta+cD) + \frac{M_2D - M_1}{\theta} + \frac{M_1}{\theta+cD}\right) \\ &= \frac{(D-1)}{\theta} + \frac{1}{(\theta+cD)} - \frac{M_2D-M_1}{\theta^2} - \frac{M_1}{(\theta+cD)^2} \\ &= \frac{\theta^3 + (2cD - c - M_2)\,\theta^2 + (c^2D(D-1) - 2c(DM_2 - M_1))\,\theta - c^2D(DM_2 - M_1)}{\theta^2(\theta+cD)^2}. \end{aligned}$$

Hence $\hat{\theta}^{MMLE}$ is a root of the cubic polynomial:

$$h_W(\theta) = \theta^3 + (2cD - c - M_2)\,\theta^2 + (c^2D(D-1) - 2c(DM_2 - M_1))\,\theta - c^2D(DM_2 - M_1).$$

Since $\hat{\theta}^{MMLE}$ is the maximum likelihood estimator of an exponential family, we know that the likelihood is convex and that the MLE exists and is unique.

Note that M_1 and M_2 are dependent on window size W. In the limit we have $M_1 \to (\theta^* + c^*D)$ and $M_2 \to (\theta^* + c^*)$ from equations (B.18). Hence we have the polynomial

$h_W(\theta) \to h^*(\theta)$, where

$$h^*(\theta) = \theta^3 + (2cD - c - c^* - \theta^*)\theta^2 + (c^2D(D-1) - 2c(D-1)\theta^*)\theta - c^2D(D-1)\theta^*.$$

We want are interested in investigating the conditions under which $\hat{\theta}^{MMLE} \to \theta^*$.

Proposition 5. *If the data is generated from the true model as described in (B.17), then $\hat{\theta}^{MMLE} \to \theta^*$ if and only if $c = c^*$:*

Proof. Let us first prove $\hat{\theta}^{MMLE} \to \theta^*$ only if $c = c^*$:
Suppose that the sequence $\hat{\theta}^{MMLE}$ converges to the true θ^*: $\hat{\theta}^{MMLE} \to \theta^*$. We know that $\hat{\theta}^{MMLE}$ are the roots of h_W and that $h_W \to h^*$. The roots of a polynomial vary continuously as a function of the coefficients [24]. This implies that the sequence θ^* has to converge to a fixed root of h^*. Hence θ^* has to be a root of h^*, i.e.

$$\begin{aligned} 0 &= (\theta^*)^3 + (2cD - c - c^* - \theta^*)(\theta^*)^2 + (c^2D(D-1) - 2c(D-1)\theta^*)(\theta^*) - c^2D(D-1)\theta^* \\ &= (\theta^*)^3 + (2cD - c - c^*)(\theta^*)^2 - (\theta^*)^3 + c^2D(D-1)\theta^* - 2c(D-1)(\theta^*)^2 - c^2D(D-1)\theta^* \\ &= (c - c^*)(\theta^*)^2. \end{aligned}$$

Since $\theta^* > 0$, we have $c = c^*$. Thus $\hat{\theta}^{MMLE} \to \theta^*$ only when $c = c^*$.

Now we prove $\hat{\theta}^{MMLE} \to \theta^*$ if $c = c^*$:
$c = c^*$ implies

$$\begin{aligned} h^*(\theta) &= \theta^3 + (2cD - 2c^* - \theta^*)\theta^2 + (c^{*2}D(D-1) - 2c^*(D-1)\theta^*)\theta - c^{*2}D(D-1)\theta^* \\ &= (\theta - \theta^*)\{(\theta + c^*(D-1))^2 + c^{*2}(D-1))\}. \end{aligned} \tag{B.19}$$

Since the second factor of the polynomial in (B.19) is always positive, h^* has two complex roots and one real root given by θ^*. Since the roots of a polynomial are continuous functions of the coefficients, and since $h_W \to h^*$, there exists a sequence $\hat{\theta}_W$ of roots of h_W that

converge to the real root of h^*, i.e. $\hat{\theta}_W \to \theta^*$. Now a cubic polynomial can have either one real and two complex roots, or three real roots. There also exist sequences of roots of h_W that converge to the two complex roots of h^*. For sufficiently large W, these will have to be non-real and therefore h_W will have only one real root. Hence the sequence that converges to the real root θ^* is given by the unique real root of h_W for sufficiently large W. Therefore $\hat{\theta}^{MMLE} \to \theta^*$.

Hence we have $\hat{\theta}^{MMLE} \to \theta^*$ if and only if $c = c^*$, i.e. we know the prior distribution on the ν's exactly.

□

HMLE

In the hierarchical model, we treat the hyper-parameter c as unknown as well. We can maximize jointly over the (θ, c) space to get the maximum likelihood estimate (we call this the HMLE, short for hierarchical maximum likelihood estimate). Maximizing Equation (B.12) or (B.13) over c and θ simultaneously, and using (B.18) gives

$$\begin{aligned} \hat{c} &= \frac{M_1 - M_2}{(D-1)} \longrightarrow c^* \\ \hat{\theta}_{HMLE} &= \frac{DM_2 - M_1}{(D-1)} \longrightarrow \theta^*, \end{aligned} \tag{B.20}$$

if the data is generated from the true model as described in (B.17). Hence, $\hat{\theta}_{HMLE}$ is consistent. Note that the HMLE in this case is the same as CMLE from Equation (B.8).

We could have put a prior on c and θ to get Bayesian estimates. As we get more and more data, the effect of the prior will eventually die out, and the posterior mean and the posterior mode will both converge to the HMLE.

B.3.3 Using a conjugate prior on ν and θ

Let us explore if similar results hold if we assume a different prior on the parameters. The conjugate prior setting for (ν, θ) is as follows:

$$\begin{aligned}
\theta \mid (a,b) &\sim \Gamma^{-1}(a,b) \\
\nu_w \mid (\theta,c) &\sim \mathcal{N}(0,\theta c) \quad \forall\ w \\
\Pi(\nu,\theta) &= \frac{b^a}{\Gamma(a)}\left(\frac{1}{\theta}\right)^{a+1}\exp\left(-\frac{b}{\theta}\right)\prod_{w=1}^{W}\left(\frac{1}{\sqrt{2\pi c\theta}}\exp\left(-\frac{\nu_w^2}{2c\theta}\right)\right).
\end{aligned}$$

Using this conjugate prior on ν we get the model similar to equation (B.11)

$$X_w = \begin{bmatrix} X_{w1} \\ \vdots \\ X_{wD} \end{bmatrix} \mid (\theta,c) \sim \mathcal{N}\left(\vec{0}, \theta R\right) \quad \forall\ w = 1,\ldots,W,$$

where

$$R = \begin{bmatrix} 1+c & c & \ldots & c \\ c & 1+c & & \vdots \\ \vdots & & \ddots & c \\ c & \ldots & c & 1+c \end{bmatrix}.$$

Log-likelihood is given by

$$\mathcal{L}^{\text{marginal}}(\theta,R) = \log f_{\theta,R}(x) = -\frac{W}{2}\left(D\log 2\pi + \log|R| + D\log\theta + \frac{1}{\theta W}\sum_{w=1}^{W} x_w^T R^{-1} x_w\right).$$

We can write R as $R = I + cuu^T$ where u is the $D \times 1$ vector of all ones. Using the Sherman–Morrison formula for matrix inversion and the matrix determinant lemma, we get

$$\begin{aligned}
|R| &= 1 + cD \\
R^{-1} &= I - \frac{c}{1+cD}uu^T.
\end{aligned}$$

Using this we get

$$\begin{aligned}
-\frac{2}{W}\mathcal{L}^{\text{marginal}}(\theta, R) &= D\log 2\pi + \log|R| + D\log\theta + \frac{1}{\theta}\frac{1}{W}\sum_{w=1}^{W} x_w^T R^{-1} x_w \\
-\frac{2}{W}\mathcal{L}^{\text{marginal}}(\theta, c) &= D\log 2\pi + \log(1+cD) + D\log\theta \\
&\qquad + \frac{1}{\theta}\frac{1}{W}\sum_{w=1}^{W} x_w^T\left(I - \frac{c}{1+cD}uu^T\right)x_w \\
&= D\log 2\pi + \log(1+cD) + D\log\theta \cdots \\
&\qquad + \frac{1}{\theta}\frac{1}{W}\sum_{w=1}^{W}\left(\sum_d x_{wd}^2 - \frac{c}{1+cD}\left(\sum_d x_{wd}\right)^2\right) \\
&= D\log 2\pi + \log(1+cD) + D\log\theta + \frac{D}{\theta}\left(M_2 - \frac{c}{1+cD}M_1\right), \quad \text{(B.21)}
\end{aligned}$$

where M_1 and M_2 are as defined before in equations (B.15),(B.16).

Let the true prior for ν_w have parameter c^* and let the true structural parameter be θ^*, which gives that the true distribution of the data is

$$X_w \sim \mathcal{N}\left(\vec{0}, \theta^* R^*\right), \quad \forall\ w \tag{B.22}$$

where

$$R^* = \theta^* \begin{bmatrix} 1+c^* & c^* & \dots & c^* \\ c^* & 1+c^* & & \vdots \\ \vdots & & \ddots & c^* \\ c^* & \dots & c^* & 1+c^* \end{bmatrix}.$$

This gives

$$\begin{aligned}
\mathbb{E}[X_w X_w^T] &= \theta^* \cdot R^* \\
\mathbb{E}[X_{wd}^2] &= \theta^*(1+c^*) \qquad \forall\ w, d \\
\mathbb{E}[X_{wd_1} X_{wd_2}] &= \theta^* c^* \qquad \forall\ w, d_1 \neq d_2.
\end{aligned}$$

Using this in equations (B.15) and (B.16) gives

$$\left.\begin{aligned} M_1 &\longrightarrow \frac{1}{D}\left(\sum_{d=1}^{D}\mathbb{E}[X_{wd}^2] + \sum_{\substack{d_1 \\ d_1\neq d_2}}\sum_{d_2}\mathbb{E}[X_{wd_1}X_{wd_2}]\right) = \theta^*(1+c^*D) \\ M_2 &\longrightarrow \frac{1}{D}\sum_{d=1}^{D}\mathbb{E}[X_{wd}^2] = \theta^*(1+c^*D). \end{aligned}\right\} \tag{B.23}$$

MMLE

We can treat c as known and maximize the marginal likelihood over θ. Then the maximum marginal likelihood estimate (MMLE) is given by minimizing the negative log marginal likelihood (B.21):

$$\hat{\theta}^{MMLE} = M_2 - \frac{c}{1+cD}M_1 \longrightarrow \theta^*\left(1+\frac{c^*-c}{1+cD}\right) \tag{B.24}$$

under the true model (B.22). We will get $\hat{\theta}^{MMLE} \to \theta^*$ if and only if $c = c^*$, i.e. our prior assumption is correct.

HMLE

Instead, we can treat the hyper-parameter c as unknown and try to estimate both c and θ. The joint MLE is given my minimizing the negative log-marginal-likelihood (B.21) over the (θ, c) space, and we get

$$\begin{aligned} \hat{c} &= \frac{M_1 - M_2}{DM_2 - M_1} \longrightarrow c^* \\ \hat{\theta}_{HMLE} &= \frac{DM_2 - M_1}{(D-1)} \longrightarrow \theta*, \end{aligned} \tag{B.25}$$

if the data is generated from true model (B.22). This HMLE with conjugate prior is also the same as the CMLE from Equation (B.8).

Appendix C

Asymptotic results for two-neuron model with $D = 2$ and $D = 3$

C.1 Asymptotic results for window width $D = 2$

Let us first get expressions for the different estimates, and then we shall explore their asymptotic behavior.

C.1.1 Expressions for the three estimators - stationary MLE, non-stationary MLE and CMLE

Stationary MLE

The MLE for this stationary model for θ is given by (3.3)

$$\hat{\theta}^{\text{stat}} = \log \frac{l\,(T - s_1 - s_2 + l)}{(s_1 - l)\,(s_2 - l)},$$

where s_1, s_2 and l are the observed sufficient statistics given by $s_1 = S_1(x) = \sum x_{1t}$, $s_2 = S_2(x) = \sum x_{2t}$ and $l = L(x) = \sum x_{1t}x_{2t}$. Each $x_t = (x_{1t}, x_{2t})' \in \{0,1\}^{2\times 1} = \{(0,0)', (0,1), (1,0)', (1,1)'\}$. Let us change our notation slightly and use N_{00}, N_{01}, N_{10}, N_{11} to count each possible combination. Hence $N_{10} = s_1 - l$, $N_{01} = s_2 - l$, $N_{11} = l$ and $N_{00} = T - s_1 - s_2 + l$. In our new notation, the stationary MLE takes the form

$$\hat{\theta}^{\text{stat}} = \log \frac{N_{00} \cdot N_{11}}{N_{10} \cdot N_{01}}. \tag{C.1}$$

Nonstationary MLE with $D = 2$

The MLE is given by solving (3.10) and (3.11) to get $\hat{\nu}_{1w}$ and $\hat{\nu}_{2w}$ as a function of θ for each window with $s_w \in \mathbb{S}_{good}$ and then substituting all these estimates into (3.12) to get $\hat{\theta}^{MLE}$.

With window width $D = 2$, possible spike count sums in each window for each neuron are 0, 1 or 2. We ignore the degenerate windows with sums 0 or 2. Hence the only non-degenerate windows have sums $s_w = (s_{1w}, s_{2w})' = (1,1)'$ and $\mathbb{S}_{good} = \{(1,1)'\}$. Let us denote the number of these non-degenerate windows as W_{11}. l_w can take only 2 values: 0 or 1. Let W_{110} and W_{111} denote the number of windows with $(s_{1w}, s_{2w}, l_w) = (1,1,0)$ and $(s_{1w}, s_{2w}, l_w) = (1,1,1)$ respectively. Note that $W_{11} = W_{110} + W_{111}$.

Substituting $s_{1w} = 1$, and $s_{2w} = 1$ in equations (3.10) and (3.11) we get $e^{\hat{\nu_1}_w} = e^{\hat{\nu_2}_w} = 1/\sqrt{e^{\hat{\theta}}}$ which can then be substituted back in (3.12) to get the non-stationary MLE:

$$\begin{aligned}
\sum_{w:s_w=(1,1)'} \frac{L_w}{D} &= \sum_{w:s_w=(1,1)'} \frac{e^{\hat{\nu}_{1w}+\hat{\nu}_{2w}+\hat{\theta}}}{1 + e^{\hat{\nu}_{1w}} + e^{\hat{\nu}_{2w}} + e^{\hat{\nu}_{1w}+\hat{\nu}_{2w}+\hat{\theta}}} \\
\frac{(1)W_{111} + (0)W_{110}}{2} &= (W_{110} + W_{111}) \frac{\sqrt{e^{\hat{\theta}}}}{2(1 + \sqrt{e^{\hat{\theta}}})} \\
\hat{\theta}^{\text{nonstat}} &= 2 \log \frac{W_{111}}{W_{110}}.
\end{aligned} \tag{C.2}$$

Conditional MLE with $D = 2$

Pruning away the degenerate windows, we are left with only the non-degenerate ones with window sums $s_{1w} = 1,\ s_{2w} = 1$. The hypergeometric coefficients in Equation (3.15) for $D = 2$ are $h_{1,1}(0) = 2$, $h_{1,1}(1) = 2$ and $h_{1,1}(2) = 0$.

Using (3.17) and (3.20), the conditional distribution to be optimized is given by

$$\psi_{good}(x|S = s) = \prod_{w:s_w \in \mathbb{S}_{good}} \frac{\exp(\theta l_w)}{\sum_{k=0}^{D} \exp(\theta k) \cdot h_{s_{1w},s_{2w}}(k)} = \prod_{w:s_w=(1,1)'} \frac{e^{(\theta l_w)}}{2 + 2e^{\theta}} \quad \text{(C.3)}$$

$$\log \psi_{good}(x|S = s) = \sum_{w:s_w=(1,1)'} \Big(\theta l_w - \log(2 + 2e^{\theta})\Big)$$

The conditional MLE is given by equating the derivative to 0, we get

$$\begin{aligned} \sum_{w:s_w=(1,1)'} l_w &= \sum_{w:s_w=(1,1)'} \frac{e^{\hat{\theta}}}{1 + e^{\hat{\theta}}} \\ (1)W_{111} + (0)W_{110} &= (W_{111} + W_{110})\frac{e^{\hat{\theta}}}{1 + e^{\hat{\theta}}} \\ \hat{\theta}^{CMLE} &= \log \frac{W_{111}}{W_{110}} \end{aligned} \quad \text{(C.4)}$$

Note that in this case we have

$$\hat{\theta}^{CMLE} = \frac{1}{2} \cdot \hat{\theta}^{nonstat}$$

C.1.2 Results on stationary data

Let us assume that the data comes from a stationary model with true parameters ν_1^*, ν_2^*, and θ^*. Hence the true data probability is

$$\mathbb{P}(X = x) = \prod_{t=1}^{T} \frac{e^{\nu_1^* x_{1t} + \nu_2^* x_{1t} + \theta^* x_{1t} x_{2t}}}{z^*} \quad \text{(C.5)}$$

where the normalizing constant is given by $z^* = 1 + e^{\nu_1^*} + e^{\nu_2^*} + e^{\nu_1^*+\nu_2^*+\theta^*}$.

At each time t, $x_t \in \{(0,0)', (0,1)', (1,0)', (1,1)'\}$ has a categorical distribution with probabilities $(p_{00}, p_{01}, p_{10}, p_{11})$ respectively given by

$$p_{00} = \frac{1}{z^*}, \qquad p_{10} = \frac{e^{\nu_1^*}}{z^*}, \qquad p_{01} = \frac{e^{\nu_2^*}}{z^*}, \qquad p_{11} = \frac{e^{\nu_1^*+\nu_2^*+\theta^*}}{z^*}. \tag{C.6}$$

Stationary MLE for stationary data

When data is generated by the true stationary distribution, we expect the stationary MLE to be consistent.

Proposition 6. *For stationary data as described in* (C.5)*, the stationary MLE given by* (C.1) *is consistent:* $\hat{\theta}^{stat} \to \theta^*$*, and asymptotically normal:*

$$\sqrt{T}\left(\hat{\theta}^{\ stat} - \theta^*\right) \to \mathcal{N}\left(0, \frac{1}{p_{00}} + \frac{1}{p_{10}} + \frac{1}{p_{01}} + \frac{1}{p_{11}}\right).$$

Proof. The stationary MLE for $D = 2$ was given by (C.1) as

$$\hat{\theta}^{stat} = \log \frac{N_{00} \cdot N_{11}}{N_{10} \cdot N_{01}}.$$

We know the true data $(X_t : t = 1, \ldots, T)$ is iid from the discrete categorical distribution (C.6). As $T \to \infty$, by Law or Large Numbers (LLN) we get

$$\begin{aligned}
\frac{N_{00}}{T} &= \frac{1}{T}\sum_t \mathbb{1}(X_{1t} = 0, X_{2t} = 0) \quad \to \quad \mathbb{E}[\mathbb{1}(X_1 = 0, X_2 = 0)] = p_{00} = \frac{1}{z^*} \\
\frac{N_{10}}{T} &= \frac{1}{T}\sum_t \mathbb{1}(X_{1t} = 1, X_{2t} = 0) \quad \to \quad \mathbb{E}[\mathbb{1}(X_1 = 1, X_2 = 0)] = p_{10} = \frac{e^{\nu_1^*}}{z^*} \\
\frac{N_{01}}{T} &= \frac{1}{T}\sum_t \mathbb{1}(X_{1t} = 0, X_{2t} = 1) \quad \to \quad \mathbb{E}[\mathbb{1}(X_1 = 0, X_2 = 1)] = p_{01} = \frac{e^{\nu_2^*}}{z^*} \\
\frac{N_{11}}{T} &= \frac{1}{T}\sum_t \mathbb{1}(X_{1t} = 1, X_{2t} = 1) \quad \to \quad \mathbb{E}[\mathbb{1}(X_1 = 1, X_2 = 1)] = p_{11} = \frac{e^{\nu_1^*+\nu_2^*+\theta^*}}{z^*},
\end{aligned}$$

and hence

$$\hat{\theta}^{stat} = \log \frac{N_{00} \cdot N_{11}}{N_{10} \cdot N_{01}} \quad \rightarrow \quad \log \frac{p_{00} \cdot p_{11}}{p_{10} \cdot p_{01}} = \log \frac{e^{\nu_1^* + \nu_2^* + \theta^*}}{e^{\nu_1^*} \cdot e^{\nu_2^*}} = \theta^*. \tag{C.7}$$

This gives us consistency. Let us now look at asymptotic normality. Let us define $\boldsymbol{M_T}$, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ as follows:

$$\boldsymbol{M_T} = \begin{bmatrix} \frac{N_{00}}{T} \\ \frac{N_{10}}{T} \\ \frac{N_{01}}{T} \\ \frac{N_{11}}{T} \end{bmatrix}, \quad \boldsymbol{\mu} = \begin{bmatrix} p_{00} \\ p_{10} \\ p_{01} \\ p_{11} \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} p_{00}(1-p_{00}) & -p_{00}p_{10} & -p_{00}p_{01} & -p_{00}p_{11} \\ -p_{00}p_{10} & p_{10}(1-p_{10}) & -p_{10}p_{01} & -p_{10}p_{11} \\ -p_{00}p_{01} & -p_{10}p_{01} & p_{01}(1-p_{01}) & -p_{01}p_{11} \\ -p_{00}p_{11} & -p_{10}p_{11} & -p_{01}p_{11} & p_{11}(1-p_{11}) \end{bmatrix}.$$

Using the Central Limit Theorem for multivariate random variables we get

$$\sqrt{T}(\boldsymbol{M_T} - \boldsymbol{\mu}) \rightarrow \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}).$$

For $\boldsymbol{y} = (y_1, y_2, y_3, y_4)' \in \mathbb{R}^{4\times 1}$, let us define the function $g(\boldsymbol{y}) = \log y_1 + \log y_4 - \log y_2 - \log y_3$. The value of $g(.)$ at $\boldsymbol{\mu}$ and $\boldsymbol{M_T}$ is

$$\begin{aligned} g(\boldsymbol{M_T}) &= \hat{\theta}^{stat} \\ g(\boldsymbol{\mu}) &= \log p_{00} + \log p_{11} - \log p_{10} - \log p_{01} = \theta^*. \end{aligned}$$

Now we can use the Multivariate Delta Method (A.3) to get the limiting distribution of the stationary MLE $\hat{\theta}^{stat}$ as follows:

$$\sqrt{T}\left(\hat{\theta}^{stat} - \theta^*\right) = \sqrt{T}\left(g(\boldsymbol{M_T}) - g(\boldsymbol{\mu})\right) \rightarrow \mathcal{N}\left(0\ ,\ (\nabla g)'\boldsymbol{\Sigma}(\nabla g)\right),$$

where $(\nabla g)'\boldsymbol{\Sigma}(\nabla g) = \sum_i \sum_j \Sigma_{ij} \frac{\partial g(\boldsymbol{\mu})}{\partial \mu_i} \frac{\partial g(\boldsymbol{\mu})}{\partial \mu_j}$. Partial derivatives are continuous at $\boldsymbol{\mu}$ and

given by

$$\begin{aligned}
\frac{\partial g(\boldsymbol{\mu})}{\partial \mu_1} &= +\frac{1}{\mu_1} = +\frac{1}{p_{00}} & \qquad \frac{\partial g(\boldsymbol{\mu})}{\partial \mu_2} &= -\frac{1}{\mu_2} = -\frac{1}{p_{10}} \\
\frac{\partial g(\boldsymbol{\mu})}{\partial \mu_3} &= -\frac{1}{\mu_3} = -\frac{1}{p_{01}} & \qquad \frac{\partial g(\boldsymbol{\mu})}{\partial \mu_4} &= +\frac{1}{\mu_4} = +\frac{1}{p_{11}},
\end{aligned}$$

and we can evaluate the new variance to be

$$\begin{aligned}
(\nabla g)'\boldsymbol{\Sigma}(\nabla g) &= \sum_i \sum_j \Sigma_{ij} \frac{\partial g(\boldsymbol{\mu})}{\partial \mu_i} \frac{\partial g(\boldsymbol{\mu})}{\partial \mu_j} \\
&= \frac{1}{p_{00}} + \frac{1}{p_{10}} + \frac{1}{p_{01}} + \frac{1}{p_{11}} \\
&= \left(1 + e^{\nu_1^*} + e^{\nu_2^*} + e^{\nu_1^*+\nu_2^*+\theta^*}\right)\left(1 + e^{-\nu_1^*} + e^{-\nu_2^*} + e^{-(\nu_1^*+\nu_2^*+\theta^*)}\right) \\
&= \frac{\left(1 + e^{\nu_1^*} + e^{\nu_2^*} + e^{\nu_1^*+\nu_2^*+\theta^*}\right)\left(1 + e^{\nu_1^*\theta^*} + e^{\nu_2^*\theta^*} + e^{\nu_1^*+\nu_2^*+\theta^*}\right)}{e^{\nu_1^*+\nu_2^*+\theta^*}}. \qquad \text{(C.8)}
\end{aligned}$$

Hence we get

$$\sqrt{T}\left(\hat{\theta}^{\text{ stat}} - \theta^*\right) \to \mathcal{N}\left(0, \frac{1}{p_{00}} + \frac{1}{p_{10}} + \frac{1}{p_{01}} + \frac{1}{p_{11}}\right). \qquad \text{(C.9)}$$

□

Non-stationary MLE with $D = 2$ for stationary data

Proposition 7. *For stationary data as described in* (C.5)*, the non-stationary MLE given by* (C.2) *is not consistent.*

Proof. The non-stationary MLE for $D = 2$ was given by (C.2) as

$$\hat{\theta}^{\text{nonstat}} = 2 \log \frac{W_{111}}{W_{110}}.$$

When $D = 2$, the probability for each non-degenerate window is

$$\mathbb{P}(X_w = x_w) = \frac{e^{\nu_1^* S_{1w} + \nu_2^* S_{1w} + \theta^* L_w}}{(z^*)^2}.$$

Hence we get

$$\begin{aligned} p_{110} &\triangleq \mathbb{P}(S_{1w}=1, S_{1w}=1, L_w=0) \\ &= \sum_{x_w} \frac{e^{\nu_1^* S_{1w}+\nu_2^* S_{1t}+\theta^* L_w)}}{(z^*)^2} \mathbb{1}(S_{1w}=1, S_{1w}=1, L_w=0) \\ &= \frac{2e^{\nu_1^*+\nu_2^*}}{(z^*)^2} = 2 \cdot p_{10} \cdot p_{01} \end{aligned} \tag{C.10}$$

$$\begin{aligned} p_{111} &\triangleq \mathbb{P}(S_{1w}=1, S_{1w}=1, L_w=1) \\ &= \sum_{x_w} \frac{e^{\nu_1^* S_{1w}+\nu_2^* S_{1t}+\theta^* L_w)}}{(z^*)^2} \mathbb{1}(S_{1w}=1, S_{1w}=1, L_w=0) \\ &= \frac{2e^{\nu_1^*+\nu_2^*+\theta^*}}{(z^*)^2} = 2 \cdot p_{11} \cdot p_{00}. \end{aligned} \tag{C.11}$$

The Law of Large Numbers gives $W_{110}/W \to p_{110}$,$W_{111}/W \to p_{111}$ and

$$\hat{\theta}^{\text{nonstat}} = 2 \log \frac{W_{111}}{W_{110}} \longrightarrow 2 \log \frac{p_{111}}{p_{110}} = 2 \log \frac{2e^{(\nu_1{}^*+\nu_2^*+\theta^*)}}{2e^{(\nu_1{}^*+\nu_2^*)}} = 2\theta^* \tag{C.12}$$

and shows that the non-stationary MLE is not consistent. □

Using the Central limit theorem we get

$$\sqrt{W}\left(\frac{1}{W}\begin{bmatrix} W_{111} \\ W_{110} \end{bmatrix} - \begin{bmatrix} p_{111} \\ p_{110} \end{bmatrix}\right) \to \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} p_{111}(1-p_{111}) & -p_{110}p_{111} \\ -p_{110}p_{111} & p_{110}(1-p_{110}) \end{bmatrix}\right),$$

and using the multivariate delta method from Theorem A.3 above with the function $g(\boldsymbol{y}) = 2\log y_1 - 2\log y_2$ for $\boldsymbol{y} \in \mathbb{R}^{2\times 1}$ gives the asymptotic distribution of the non-stationary MLE:

$$\boldsymbol{\mu} = \begin{bmatrix} p_{111} \\ p_{110} \end{bmatrix}, \qquad g(\boldsymbol{\mu}) = \theta^*, \qquad g\left(\frac{1}{W}\begin{bmatrix} W_{111} \\ W_{110} \end{bmatrix}\right) = \hat{\theta}^{\text{nonstat}}, \qquad \nabla g(\boldsymbol{\mu}) = 2\begin{bmatrix} +\frac{1}{p_{111}} \\ -\frac{1}{p_{110}} \end{bmatrix},$$

$$(\nabla g)'\Sigma(\nabla g) = \sum_i \sum_j \Sigma_{ij} \frac{\partial g(\boldsymbol{\mu})}{\partial \mu_i} \frac{\partial g(\boldsymbol{\mu})}{\partial \mu_j} = 4\left(\frac{1}{p_{111}} + \frac{1}{p_{110}}\right).$$

This gives

$$\sqrt{W}\left(\hat{\theta}^{\text{nonstat}} - 2\theta^*\right) \to \mathcal{N}\left(0, \frac{4}{p_{111}} + \frac{4}{p_{110}}\right)$$

$$\sqrt{T}\left(\hat{\theta}^{\text{nonstat}} - 2\theta^*\right) \to \mathcal{N}\left(0, \frac{8}{p_{111}} + \frac{8}{p_{110}}\right), \tag{C.13}$$

where

$$\frac{2}{p_{111}} + \frac{2}{p_{110}} = \frac{(z^*)^2}{e^{\nu_1^*+\nu_2^*}} + \frac{(z^*)^2}{e^{\nu_1^*+\nu_2^*+\theta^*}} = \frac{(z^*)^2}{e^{\nu_1^*+\nu_2^*+\theta^*}}(e^{\theta^*}+1). \tag{C.14}$$

CMLE with $D = 2$ for stationary data

Proposition 8. *For stationary data as described in (C.5), the CMLE given by (C.4) is consistent and asymptotically normal.*

The CMLE for $D = 2$ was given by (C.4) as

$$\hat{\theta}^{CMLE} = \log\frac{W_{111}}{W_{110}} = \frac{1}{2}\cdot\hat{\theta}^{nonstat}.$$

Using (C.12) and (C.13), we get

$$\hat{\theta}^{\text{CMLE}} \longrightarrow \theta^* \tag{C.15}$$

and

$$\sqrt{T}\left(\hat{\theta}^{\text{CMLE}} - \theta^*\right) \to \mathcal{N}\left(0, \frac{2}{p_{111}} + \frac{2}{p_{110}}\right). \tag{C.16}$$

The CMLE is consistent and asymptotically normal.

Simulation results for stationary data

We show results on simulated data in the main Section 3.4.1.

C.1.3 Results on non-stationary data

Now let us explore how these estimates behave asymptotically if the data is generated from a non-stationary data. The simplest case would be if the data comes from one set of parameters for the first half and a different set of parameters for the second half. Let us assume that T is divisible by 4 (we want each half to break into windows of width 2). Let the parameters be $(\nu_1^*, \nu_2^*, \theta^*)$ for $t \leq T/2$, and $(\lambda_1^*, \lambda_2^*, \theta^*)$ for $t > T/2$. Hence the true data probability is

$$\mathbb{P}(X = x) = \prod_{t=1}^{T/2} \frac{\exp\left(\nu_1^* x_{1t} + \nu_2^* x_{1t} + \theta^* x_{1t} x_{2t}\right)}{z_\nu^*} \prod_{t=\frac{T}{2}+1}^{T} \frac{\exp\left(\lambda_1^* x_{1t} + \lambda_2^* x_{1t} + \theta^* x_{1t} x_{2t}\right)}{z_\lambda^*}, \tag{C.17}$$

where the normalizing constants are given by

$$z_\nu^* = 1 + e^{\nu_1^*} + e^{\nu_2^*} + e^{\nu_1^* + \nu_2^* + \theta^*} \qquad \text{and} \qquad z_\lambda^* = 1 + e^{\lambda_1^*} + e^{\lambda_2^*} + e^{\lambda_1^* + \lambda_2^* + \theta^*}.$$

The categorical probabilities for x_t for the first half for $\forall\ t \leq T/2$ are given by

$$p_{00}^1 = \frac{1}{z_\nu^*}, \qquad p_{10}^1 = \frac{e^{\nu_1^*}}{z_\nu^*}, \qquad p_{01}^1 = \frac{e^{\nu_2^*}}{z_\nu^*}, \qquad p_{11}^1 = \frac{e^{\nu_1^* + \nu_2^* + \theta^*}}{z_\nu^*} \tag{C.18}$$

and

$$p_{00}^2 = \frac{1}{z_\lambda^*}, \qquad p_{10}^2 = \frac{e^{\lambda_1^*}}{z_\lambda^*}, \qquad p_{01}^2 = \frac{e^{\lambda_2^*}}{z_\lambda^*}, \qquad p_{11}^2 = \frac{e^{\lambda_1^* + \lambda_2^* + \theta^*}}{z_\lambda^*} \tag{C.19}$$

for $t > T/2$.

Let us use N_{00}^1, N_{01}^1, N_{10}^1, N_{11}^1 to count each possible combination for $t \leq T/2$ and N_{00}^2, N_{01}^2, N_{10}^2, N_{11}^2 to count each possible combination for $t > T/2$.

As $T \to \infty$, by Law or Large Numbers we get

$$\begin{aligned}
\frac{N_{00}}{T} = \frac{1}{2}\left(\frac{N_{00}^1}{\frac{T}{2}} + \frac{N_{00}^2}{\frac{T}{2}}\right) &\to \frac{1}{2}\left(p_{00}^1 + p_{00}^2\right) = \frac{1}{2}\left(\frac{1}{z_\nu^*} + \frac{1}{z_\lambda^*}\right) \\
\frac{N_{10}}{T} = \frac{1}{2}\left(\frac{N_{10}^1}{\frac{T}{2}} + \frac{N_{10}^2}{\frac{T}{2}}\right) &\to \frac{1}{2}\left(p_{10}^1 + p_{10}^2\right) = \frac{1}{2}\left(\frac{e^{\nu_1^*}}{z_\nu^*} + \frac{e^{\lambda_1^*}}{z_\lambda^*}\right) \\
\frac{N_{01}}{T} = \frac{1}{2}\left(\frac{N_{01}^1}{\frac{T}{2}} + \frac{N_{01}^2}{\frac{T}{2}}\right) &\to \frac{1}{2}\left(p_{01}^1 + p_{01}^2\right) = \frac{1}{2}\left(\frac{e^{\nu_2^*}}{z_\nu^*} + \frac{e^{\lambda_2^*}}{z_\lambda^*}\right) \\
\frac{N_{11}}{T} = \frac{1}{2}\left(\frac{N_{11}^1}{\frac{T}{2}} + \frac{N_{11}^2}{\frac{T}{2}}\right) &\to \frac{1}{2}\left(p_{11}^1 + p_{11}^2\right) = \frac{1}{2}\left(\frac{e^{\nu_1^*+\nu_2^*+\theta^*}}{z_\nu^*} + \frac{e^{\lambda_1^*+\lambda_2^*+\theta^*}}{z_\lambda^*}\right).
\end{aligned}$$

If we define $\boldsymbol{M}_{1T}$, $\boldsymbol{\mu}_1$, $\boldsymbol{M}_{2T}$, $\boldsymbol{\mu}_2$ $\boldsymbol{\Sigma}_1$, and $\boldsymbol{\Sigma}_2$ as

$$\boldsymbol{M}_{1T} = \frac{2}{T}\begin{bmatrix} N_{00}^1 \\ N_{10}^1 \\ N_{01}^1 \\ N_{11}^1 \end{bmatrix}, \qquad \boldsymbol{\mu}_1 = \begin{bmatrix} p_{00}^1 \\ p_{10}^1 \\ p_{01}^1 \\ p_{11}^1 \end{bmatrix}, \qquad \boldsymbol{M}_{2T} = \frac{2}{T}\begin{bmatrix} N_{00}^1 \\ N_{10}^1 \\ N_{01}^1 \\ N_{11}^1 \end{bmatrix}, \qquad \boldsymbol{\mu}_2 = \begin{bmatrix} p_{00}^2 \\ p_{10}^2 \\ p_{01}^2 \\ p_{11}^2 \end{bmatrix},$$

$$\boldsymbol{\Sigma}_1 = \begin{bmatrix} p_{00}^1(1-p_{00}^1) & -p_{00}^1 p_{10}^1 & -p_{00}^1 p_{01}^1 & -p_{00}^1 p_{11}^1 \\ -p_{00}^1 p_{10}^1 & p_{10}^1(1-p_{10}^1) & -p_{10}^1 p_{01}^1 & -p_{10}^1 p_{11}^1 \\ -p_{00}^1 p_{01}^1 & -p_{10}^1 p_{01}^1 & p_{01}^1(1-p_{01}^1) & -p_{01}^1 p_{11}^1 \\ -p_{00}^1 p_{11}^1 & -p_{10}^1 p_{11}^1 & -p_{01}^1 p_{11}^1 & p_{11}^1(1-p_{11}^1) \end{bmatrix}, \tag{C.20}$$

$$\boldsymbol{\Sigma}_2 = \begin{bmatrix} p_{00}^2(1-p_{00}^2) & -p_{00}^2 p_{10}^2 & -p_{00}^2 p_{01}^2 & -p_{00}^2 p_{11}^2 \\ -p_{00}^2 p_{10}^2 & p_{10}^2(1-p_{10}^2) & -p_{10}^2 p_{01}^2 & -p_{10}^2 p_{11}^2 \\ -p_{00}^2 p_{01}^2 & -p_{10}^2 p_{01}^2 & p_{01}^2(1-p_{01}^2) & -p_{01}^2 p_{11}^2 \\ -p_{00}^2 p_{11}^2 & -p_{10}^2 p_{11}^2 & -p_{01}^2 p_{11}^2 & p_{11}^2(1-p_{11}^2) \end{bmatrix}, \tag{C.21}$$

using the Central Limit Theorem for multivariate random variables, we get

$$\sqrt{\frac{T}{2}}(\boldsymbol{M}_{1T} - \boldsymbol{\mu}_1) \to \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_1)$$

$$\sqrt{\frac{T}{2}}(\boldsymbol{M}_{2T} - \boldsymbol{\mu}_2) \to \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_2).$$

If $\boldsymbol{M}_T = (\boldsymbol{M}_{1T} + \boldsymbol{M}_{2T})/2 = [N_{00}N_{10}N_{01}N_{11}]'/T$, $\boldsymbol{\mu} = (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)/2$ and $\boldsymbol{\Sigma} = (\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)/2$, then by the independence of $\boldsymbol{M}_{1T}$ and $\boldsymbol{M}_{2T}$, we get

$$\sqrt{T}(\boldsymbol{M_T} - \boldsymbol{\mu}) \to \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}).$$

Stationary MLE for non-stationary data

Proposition 9. *For non-stationary data as described in* (C.17)*, the stationary MLE given by* (C.1) *is not necessarily consistent.*

Proof. The limiting distribution of is given by

$$\hat{\theta}^{stat} = \log \frac{N_{00} \cdot N_{11}}{N_{10} \cdot N_{01}} \longrightarrow \log \frac{\left(p_{00}^1 + p_{00}^2\right)\left(p_{11}^1 + p_{11}^2\right)}{\left(p_{10}^1 + p_{10}^2\right)\left(p_{01}^1 + p_{01}^2\right)} \tag{C.22}$$

$$\longrightarrow \log \frac{\left(z_\lambda^* + z_\nu^*\right)\left(z_\lambda^* e^{\nu_1^* + \nu_2^* + \theta^*} + z_\nu^* e^{\lambda_1^* + \lambda_2^* + \theta^*}\right)}{\left(z_\lambda^* e^{\nu_1^*} + z_\nu^* e^{\lambda_1^*}\right)\left(z_\lambda^* e^{\nu_2^*} + z_\nu^* e^{\lambda_2^*}\right)} \tag{C.23}$$

$$\longrightarrow \theta^* + \log \frac{\left(z_\lambda^* + z_\nu^*\right)\left(z_\lambda^* e^{\nu_1^* + \nu_2^*} + z_\nu^* e^{\lambda_1^* + \lambda_2^*}\right)}{\left(z_\lambda^* e^{\nu_1^*} + z_\nu^* e^{\lambda_1^*}\right)\left(z_\lambda^* e^{\nu_2^*} + z_\nu^* e^{\lambda_2^*}\right)}. \tag{C.24}$$

Let us call this limiting value

$$\begin{aligned}
\theta_0^{stat} &\triangleq \theta^* + \log \frac{\left(z_\lambda^* + z_\nu^*\right)\left(z_\lambda^* e^{\nu_1^* + \nu_2^*} + z_\nu^* e^{\lambda_1^* + \lambda_2^*}\right)}{\left(z_\lambda^* e^{\nu_1^*} + z_\nu^* e^{\lambda_1^*}\right)\left(z_\lambda^* e^{\nu_2^*} + z_\nu^* e^{\lambda_2^*}\right)} \\
&= \theta^* + \log \frac{z_\nu^{*2} e^{\lambda_1^* + \lambda_2^*} + z_\lambda^{*2} e^{\nu_1^* + \nu_2^*} + z_\nu^* z_\lambda^* e^{\lambda_1^* + \lambda_2^*} + z_\nu^* z_\lambda^* e^{\nu_1^* + \nu_2^*}}{z_\nu^{*2} e^{\lambda_1^* + \lambda_2^*} + z_\lambda^{*2} e^{\nu_1^* + \nu_2^*} + z_\nu^* z_\lambda^* e^{\lambda_1^* + \nu_2^*} + z_\nu^* z_\lambda^* e^{\nu_1^* + \lambda_2^*}}.
\end{aligned}$$

The limit on the right will not be θ^* for all choices of parameters (we show a particular case in our simulations in Section C.1.3).

$\theta_0^{stat} = \theta^*$ if and only if

$$\begin{aligned}
& \frac{z_\nu^{*2} e^{\lambda_1^*+\lambda_2^*} + z_\lambda^{*2} e^{\nu_1^*+\nu_2^*} + z_\nu^* z_\lambda^* e^{\lambda_1^*+\lambda_2^*} + z_\nu^* z_\lambda^* e^{\nu_1^*+\nu_2^*}}{z_\nu^{*2} e^{\lambda_1^*+\lambda_2^*} + z_\lambda^{*2} e^{\nu_1^*+\nu_2^*} + z_\nu^* z_\lambda^* e^{\lambda_1^*+\nu_2^*} + z_\nu^* z_\lambda^* e^{\nu_1^*+\lambda_2^*}} = 1 \\
\text{iff} \quad & e^{\lambda_1^*+\lambda_2^*} + e^{\nu_1^*+\nu_2^*} = e^{\lambda_1^*+\nu_2^*} + e^{\nu_1^*+\lambda_2^*} \\
\text{iff} \quad & (e^{\nu_1^*} - e^{\lambda_1^*})(e^{\nu_2^*} - e^{\lambda_2^*}) = 0 \\
\text{iff} \quad & \nu_1^* = \lambda_1^* \quad \text{or} \quad \nu_2^* = \lambda_2^*.
\end{aligned}$$

For our choice of parameters, $\hat{\theta}^{stat}$ is consistent if only if $\nu_1^* = \lambda_1^*$ or $\nu_2^* = \lambda_2^*$. Thus the stationary MLE need not be consistent. □

Using the function $g(\boldsymbol{y}) = \log y_1 + \log y_4 - \log y_2 - \log y_3$, we get $g(\boldsymbol{M_T}) = \theta_0^{stat}$ and $g(\boldsymbol{\mu}) = \log((p_{00}^1 + p_{00}^2) \cdot (p_{11}^1 + p_{11}^2)) - \log((p_{10}^1 + p_{10}^2) \cdot (p_{01}^1 + p_{01}^2)) = \theta_0^{stat}$. The multivariate delta method (A.3) gives

$$\sqrt{T}\left(\hat{\theta}^{\text{stat}} - \theta_0^{stat}\right) = \sqrt{T}\left(g(\boldsymbol{M_T}) - g(\boldsymbol{\mu})\right) \longrightarrow \mathcal{N}\left(0\ ,\ (\nabla g)'\boldsymbol{\Sigma}(\nabla g)\right), \tag{C.25}$$

where

$$\begin{aligned}
\nabla g(\boldsymbol{\mu}) &= \left[\frac{1}{(p_{00}^1 + p_{00}^2)} \quad \frac{-1}{(p_{10}^1 + p_{10}^2)} \quad \frac{-1}{(p_{01}^1 + p_{01}^2)} \quad \frac{1}{(p_{11}^1 + p_{11}^2)}\right]' \\
\boldsymbol{\Sigma} &= \frac{\boldsymbol{\Sigma}^1 + \boldsymbol{\Sigma}^2}{2}.
\end{aligned}$$

with $\boldsymbol{\Sigma}^1$ and $\boldsymbol{\Sigma}^2$ as defined previously in equations (C.20) and (C.21).

Non-stationary MLE with $D = 2$ for non-stationary data

Proposition 10. *For non-stationary data as described in* (C.17)*, the non-stationary MLE given by* (C.1) *is not consistent.*

Proof. Let us denote the number of windows with $(s_{1w}, s_{2w}, l_w) = (1, 1, 0)$ in the first half

as W^1_{110} and in the second half as W^2_{110}, and the number of windows with $(s_{1w}, l_{2w}, L_w) = (1,1,1)$ in the first half as W^1_{111} and in the second half as W^1_{111}.

Each half is a stationary distribution with different sets of true parameters. The Law of Large Numbers gives

$$\begin{aligned}
\frac{W_{110}}{W} = \frac{1}{2}\left(\frac{W^1_{110}}{\frac{W}{2}} + \frac{W^2_{110}}{\frac{W}{2}}\right) &\rightarrow \frac{1}{2}\left(p^1_{110} + p^2_{110}\right) = \frac{1}{2}\left(\frac{2e^{\nu_1^*+\nu_2^*}}{(z^*_\nu)^2} + \frac{2e^{\lambda_1^*+\lambda_2^*}}{(z^*_\lambda)^2}\right) \\
\frac{W_{111}}{W} = \frac{1}{2}\left(\frac{W^1_{111}}{\frac{W}{2}} + \frac{W^2_{111}}{\frac{W}{2}}\right) &\rightarrow \frac{1}{2}\left(p^1_{111} + p^2_{111}\right) = \frac{1}{2}\left(\frac{2e^{\nu_1^*+\nu_2^*+\theta^*}}{(z^*_\nu)^2} + \frac{2e^{\lambda_1^*+\lambda_2^*+\theta^*}}{(z^*_\lambda)^2}\right)
\end{aligned}$$

and the limiting value of the non-stationary MLE is given by

$$\hat{\theta}^{\text{nonstat}} = 2\log\frac{W_{111}}{W_{110}} \longrightarrow 2\log\frac{p^1_{111}+p^2_{111}}{p^1_{110}+p^2_{110}} = 2\log\frac{z^{*2}_\lambda e^{\nu_1^*+\nu_2^*+\theta^*} + z^{*2}_\nu e^{\lambda_1^*+\lambda_2^*+\theta^*}}{z^{*2}_\lambda e^{\nu_1^*+\nu_2^*} + z^{*2}_\nu e^{\lambda_1^*+\lambda_2^*}} = 2\theta^*. \tag{C.26}$$

□

Using the Central limit theorem (CLT),

$$\begin{aligned}
\sqrt{\frac{W}{2}}\left(\frac{2}{W}\begin{bmatrix}W^1_{111}\\ W^1_{110}\end{bmatrix} - \begin{bmatrix}p^1_{111}\\ p^1_{110}\end{bmatrix}\right) &\rightarrow \mathcal{N}(\mathbf{0}, \mathbf{\Sigma_1}) = \mathcal{N}\left(\mathbf{0}, \begin{bmatrix}p^1_{111}(1-p^1_{111}) & -p^1_{111}p^1_{110}\\ -p^1_{111}p^1_{110} & p^1_{110}(1-p^1_{110})\end{bmatrix}\right) \\
\sqrt{\frac{W}{2}}\left(\frac{2}{W}\begin{bmatrix}W^2_{111}\\ W^2_{110}\end{bmatrix} - \begin{bmatrix}p^2_{111}\\ p^2_{110}\end{bmatrix}\right) &\rightarrow \mathcal{N}(\mathbf{0}, \mathbf{\Sigma_2}) = \mathcal{N}\left(\mathbf{0}, \begin{bmatrix}p^2_{111}(1-p^2_{111}) & -p^2_{111}p^2_{110}\\ -p^2_{111}p^2_{110} & p^2_{110}(1-p^2_{110})\end{bmatrix}\right)
\end{aligned}$$

By independence of $[W^1_{111}\ W^1_{110}]'$ and $[W^2_{111}\ W^2_{110}]'$, we get

$$\sqrt{W}\left(\frac{1}{W}\begin{bmatrix}W_{111}\\ W_{110}\end{bmatrix} - \boldsymbol{\mu}\right) \rightarrow \mathcal{N}\left(\mathbf{0}, \frac{\mathbf{\Sigma_1} + \mathbf{\Sigma_2}}{2}\right), \tag{C.27}$$

where $\boldsymbol{\mu} = \begin{bmatrix}\frac{p^1_{111}+p^2_{111}}{2} & \frac{p^1_{110}+p^2_{110}}{2}\end{bmatrix}'$. We use the multivariate delta method (A.3) with the

function $g(y) = 2\log y_1 - 2\log y_2$ to get the asymptotic distribution

$$\sqrt{T}\left(\hat{\theta}^{\text{nonstat}} - 2\theta^*\right) = \sqrt{T}\left(g\left(\frac{1}{W}\begin{bmatrix} W_{111} \\ W_{110} \end{bmatrix}\right) - g(\boldsymbol{\mu})\right) \longrightarrow \mathcal{N}\left(0\ ,\ (\nabla g)'\boldsymbol{\Sigma}(\nabla g)\right), \tag{C.28}$$

where

$$\begin{aligned} \nabla g(\boldsymbol{\mu}) &= 2\begin{bmatrix} \frac{1}{(p^1_{111} + p^2_{111})} & \frac{-1}{(p^1_{110} + p^2_{110})} \end{bmatrix}' \\ \boldsymbol{\Sigma} &= \frac{\boldsymbol{\Sigma}^1 + \boldsymbol{\Sigma}^2}{2} \end{aligned}$$

with $\boldsymbol{\Sigma}^1$ and $\boldsymbol{\Sigma}^2$ as defined previously in equations (C.20) and (C.21).

CMLE with $D = 2$ for non-stationary data

Proposition 11. *For non-stationary data as described in* (C.17)*, the CMLE given by* (C.4) *is consistent and asymptotically normal.*

Proof.

$$\hat{\theta}^{\text{CMLE}} = \log\frac{W_{111}}{W_{110}} \longrightarrow \log\frac{p^1_{111} + p^2_{111}}{p^1_{110} + p^2_{110}} = \log\frac{z^{*2}_\lambda e^{\nu^*_1 + \nu^*_2 + \theta^*} + z^{*2}_\nu e^{\lambda^*_1 + \lambda^*_2 + \theta^*}}{z^{*2}_\lambda e^{\nu^*_1 + \nu^*_2} + z^{*2}_\nu e^{\lambda^*_1 + \lambda^*_2}}. = \theta^* \tag{C.29}$$

Using Equation (C.27) and the multivariate delta method (A.3) with the function $g(y) = \log y_1 - \log y_2$ we get

$$\sqrt{T}\left(\hat{\theta}^{\text{CMLE}} - \theta^*\right) = \sqrt{T}\left(g\left(\frac{1}{W}\begin{bmatrix} W_{111} \\ W_{110} \end{bmatrix}\right) - g(\boldsymbol{\mu})\right) \longrightarrow \mathcal{N}\left(0\ ,\ (\nabla g)'\boldsymbol{\Sigma}(\nabla g)\right), \tag{C.30}$$

where

$$\begin{aligned} \nabla g(\boldsymbol{\mu}) &= \left[\frac{1}{(p^1_{111}+p^2_{111})} \quad \frac{-1}{(p^1_{110}+p^2_{110})}\right]' \\ \boldsymbol{\Sigma} &= \frac{\boldsymbol{\Sigma}^1+\boldsymbol{\Sigma}^2}{2} \end{aligned}$$

with $\boldsymbol{\Sigma}^1$ and $\boldsymbol{\Sigma}^2$ as defined previously in equations (C.20) and (C.21). □

Simulation results for non-stationary data

We show results on simulated data in Section 3.4.2.

C.1.4 Results on data generated from conditional distribution given the window sums $(1,1)$

Let us assume now that the true data comes from a conditional model such that in each window of width $D=2$ the spike count sums are $s_{1w}=s_{2w}=1$ and the true synchrony parameter is θ^*. Using (C.3), the true conditional distribution of the data given θ^* is

$$\psi(x|S=s) = \prod_{w=1}^{W} \frac{e^{(\theta^* L(x_w))}}{(2+2e^{\theta^*})},$$

since each window will have $s_w=(1,1)'$. Thus the probability of each possible x_w is as follows:

$\begin{pmatrix}0&1\\1&0\end{pmatrix}, \begin{pmatrix}1&0\\0&1\end{pmatrix}$ each with probability $\frac{1}{2+2e^{\theta*}}$ and $\begin{pmatrix}0&1\\0&1\end{pmatrix}, \begin{pmatrix}1&0\\1&0\end{pmatrix}$ each with probability $\frac{e^{\theta^*}}{2+2e^{\theta*}}$.

Let W_{110} count the first two possibilities and W_{111} count the last two. In this case we also have $N_{00}=N_{11}=W_{111}$ and $N_{10}=N_{01}=W_{110}$.

If $q_{110} = \mathbb{P}\left(S_1(x_w) = 1, S_2(x_w) = 1, L(x_w) = 0\right)$ and $q_{111} = \mathbb{P}\left(S_1(x_w) = 1, S_2(x_w) = 1, L(x_w) = 1\right)$, by the Law of Large Numbers,

$$\begin{aligned}
\frac{N_{00}}{W} = \frac{N_{11}}{W} = \frac{W_{111}}{W} \quad &\longrightarrow \quad q_{111} = 2 \times \frac{e^{\theta^*}}{2 + 2e^{\theta^*}} = \frac{e^{\theta^*}}{1 + e^{\theta^*}} \\
\frac{N_{10}}{W} = \frac{N_{01}}{W} = \frac{W_{110}}{W} \quad &\longrightarrow \quad q_{110} = 2 \times \frac{1}{2 + 2e^{\theta^*}} = \frac{1}{1 + e^{\theta^*}}.
\end{aligned}$$

Hence the limiting value for the three estimators are:

$$\hat{\theta}^{\text{stat}} = \hat{\theta}^{\text{nonstat}} = \log \frac{N_{00} \cdot N_{11}}{N_{10} \cdot N_{01}} = 2 \log \frac{W_{111}}{W_{110}} \quad \longrightarrow \quad 2\theta^* \tag{C.31}$$

$$\hat{\theta}^{\text{CMLE}} = \log \frac{W_{111}}{W_{110}} \quad \longrightarrow \quad \theta^* \tag{C.32}$$

Also applying the Central Limit Theorem and the delta method as before, we get the asymptotic distribution as

$$\begin{aligned}
\sqrt{T}\left(\hat{\theta}^{\text{stat}} - 2\theta^*\right) \quad &\longrightarrow \quad \mathcal{N}\left(0, \frac{8(1 + e^{\theta^*})^2}{e^{\theta^*}}\right) \\
\sqrt{T}\left(\hat{\theta}^{\text{nonstat}} - 2\theta^*\right) \quad &\longrightarrow \quad \mathcal{N}\left(0, \frac{8(1 + e^{\theta^*})^2}{e^{\theta^*}}\right) \\
\sqrt{T}\left(\hat{\theta}^{\text{CMLE}} - \theta^*\right) \quad &\longrightarrow \quad \mathcal{N}\left(0, \frac{2(1 + e^{\theta^*})^2}{e^{\theta^*}}\right)
\end{aligned} \tag{C.33}$$

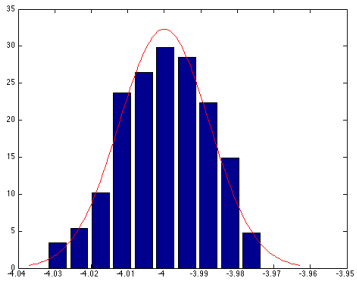
Simulation results

Instead of generating data from the conditional distribution, we generate the data from a stationary distribution, and ignore all windows except the ones with sums (1,1) which will give data generated from the conditional distribution. We generated stationary data from $\nu_1^* = \log(4)$, $\nu_2^* = \log(2)$ and $\theta^* = -2$. We choose a different value for θ^* here because we do not want the probability distributions p_{00} or p_{11} to dominate over the others. For this set of true values we have $p_{00} = 0.124$, $p_{10} = 0.495$, $p_{01} = 0.247$, $p_{11} = 0.134$. (For the earlier set, when $\theta^* = 1$, we had $p_{00} = 0.035$, $p_{10} = 0.139$, $p_{01} = 0.069$, $p_{11} = 0.757$.) We generate 250 samples with a smaller $T = 5 \times 10^5$ (computationally more intensive since we sample more data, and throw away the windows with sums 0 or 2). The stat-MLE and

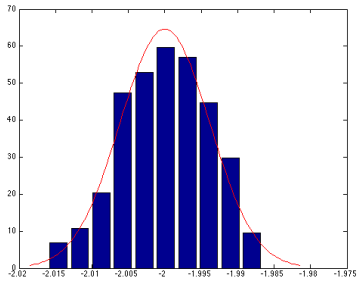
nonstat-MLE are identical:

$$\hat{\theta}^{\text{stat}} = \hat{\theta}^{\text{nonstat}} \xrightarrow{d} \mathcal{N}\left(-4,\ 1.52 \times 10^{-4}\right)$$
$$\hat{\theta}^{\text{CMLE}} \xrightarrow{d} \mathcal{N}\left(-2,\ 3.81 \times 10^{-5}\right)$$

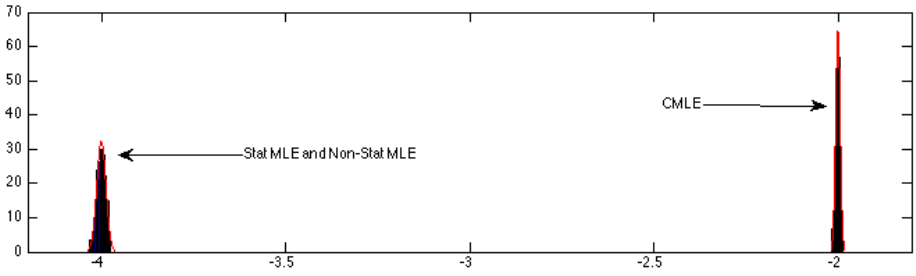
In simulations, the variances for $\hat{\theta}^{\text{ stat}} = \hat{\theta}^{\text{nonstat}}$ and $\hat{\theta}^{\text{CMLE}}$ are 1.455×10^{-4} and 3.636×10^{-5} respectively.



(a) Stat-MLE & Non-stat MLE, $\mathcal{N}\left(-4, 1.52 \times 10^{-4}\right)$



(b) CMLE, $\mathcal{N}\left(-2, 3.81 \times 10^{-5}\right)$



(c) all together

Normalized distribution of stat-MLE, MLE and CMLE on data generated from the conditional distribution with 250 samples at $T = 2 \times 10^5$.

C.2 Asymptotic results for window width $D = 3$

With window width $D = 3$, possible spike count sums in each window for each neuron are 0, 1, 2 or 3. For MLE and CMLE calculations, we ignore the degenerate windows with sums 0 or 3. Hence the only non-degenerate windows have sums $s_w = (s_{1w}, s_{2w})' \in \mathbb{S}_{good} = \{(1,1)',\ (1,2)',\ (1,1)', (2,2)'\}$. Let us denote the number of these non-degenerate windows

as W_{11}, W_{12}, W_{21} and W_{22} respectively, and in general we use W_{ijk} to count the number of windows with $(s_{1w}, s_{2w}, l_w) = (i, j, k)$.

The possible synchrony counts for windows with sums $(1,1)$, $(1,2)$ and $(2,1)$ are 0 and 1 and for windows with sums $(2,2)$, the possible synchronies are 1 and 2. Hence $W_{11} = W_{110} + W_{111}$, $W_{12} = W_{120} + W_{121}$, $W_{21} = W_{210} + W_{211}$ and $W_{22} = W_{221} + W_{222}$.

C.2.1 Non-stationary MLE with $D = 3$

The MLE is given by solving (3.10) and (3.11) to get $\hat{\nu}_{1w}$ and $\hat{\nu}_{2w}$ as a function of θ for each window with $s_w \in \mathbb{S}_{good}$ and then substituting all these estimates in (3.12) to get $\hat{\theta}^{MLE}$.

Let us denote $\exp(\theta)$ as ρ. For each possible type of window, we get the estimates for $\hat{\nu}_1$ and $\hat{\nu}_2$ in terms of ρ by solving (3.10) and (3.10), which can then be substituted in (3.12) to get $\hat{\rho} = \exp(\hat{\theta})$. We use the notation

$$\gamma_{ij}(\rho) \triangleq \frac{e^{\hat{\nu}_{1w}+\hat{\nu}_{2w}+\hat{\theta}}}{1 + e^{\hat{\nu}_{1w}} + e^{\hat{\nu}_{2w}} + e^{\hat{\nu}_{1w}+\hat{\nu}_{2w}+\hat{\theta}}}$$

for each term in (3.12) and hence $\hat{\rho}$ is given by solving

$$\sum_{w:s_w \in \mathbb{S}_{good}} \frac{l_w}{D} = \sum_{w:s_w \in \mathbb{S}_{good}} \gamma_{s_{1w},s_{2w}}(\rho). \tag{C.34}$$

The left hand side of the above equation is given by $(W_{111}+W_{121}+W_{211}+W_{221}+2W_{222})/D$.

For windows with sums $(1,1)$, we get

$$e^{\hat{\nu_{1w}}} = e^{\hat{\nu_{2w}}} = \frac{-1+\sqrt{1+8\rho}}{4\rho}, \qquad \gamma_{11}(\rho) = \frac{2\rho+1+\sqrt{1+8\rho}}{6(\rho-1)}.$$

For windows with sums $(1,2)$, we get

$$e^{\nu_{\hat{1}w}} = \frac{-\sqrt{\rho}+\sqrt{8+\rho}}{4\sqrt{\rho}},\ e^{\nu_{\hat{2}w}} = \frac{\sqrt{\rho}+\sqrt{8+\rho}}{2\sqrt{\rho}},\ \ \gamma_{12}(\rho) = \frac{4\sqrt{\rho}}{9\sqrt{\rho}+3\sqrt{8+\rho}} = \frac{3\rho-\sqrt{\rho(8+\rho)}}{6(\rho-1)}.$$

For windows with sums $(2,1)$, we get

$$e^{\nu_{\hat{1}w}} = \frac{\sqrt{\rho}+\sqrt{8+\rho}}{2\sqrt{\rho}},\ e^{\nu_{\hat{2}w}} = \frac{-\sqrt{\rho}+\sqrt{8+\rho}}{4\sqrt{\rho}},\ \ \gamma_{21}(\rho) = \frac{4\sqrt{\rho}}{9\sqrt{\rho}+3\sqrt{8+\rho}} = \frac{3\rho-\sqrt{\rho(8+\rho)}}{6(\rho-1)}.$$

For windows with sums $(2,2)$, we get

$$e^{\nu_{\hat{1}w}} = e^{\nu_{\hat{2}w}} = \frac{1+\sqrt{1+8\rho}}{2\rho}, \qquad \gamma_{22}(\rho) = \frac{4\rho-1-\sqrt{1+8\rho}}{6(\rho-1)}.$$

Substituting these in the above Equation (C.34) gives

$$\frac{W_{111}+W_{121}+W_{211}+W_{221}+2W_{222}}{3} - W_{11}\gamma_{11}(\rho)+W_{12}\gamma_{12}(\rho)+W_{21}\gamma_{21}(\rho)+W_{22}\gamma_{22}(\rho) = 0.$$

Let us define

$$\begin{aligned} h(\rho) \ \triangleq\ & \frac{W_{111}+W_{121}+W_{211}+W_{221}+2W_{222}}{W}\cdots \\ & -\ 3\ \frac{W_{11}\gamma_{11}(\rho)+W_{12}\gamma_{12}(\rho)+W_{21}\gamma_{21}(\rho)+W_{22}\gamma_{22}(\rho)}{W}. \end{aligned} \quad \text{(C.35)}$$

If $\hat{\rho}$ is a root of $h(\rho)$, then the non-stationary MLE is given by $\hat{\theta}^{\mathrm{MLE}} = \log\hat{\rho}$.

Asymptotic behavior of non-stationary MLE on stationary data

Proposition 12. *Let us assume that the data comes from a stationary model with true parameters ν_1^*, ν_2^*, and θ^*:*

$$\mathbb{P}(X=x) = \prod_{t=1}^{T} \frac{e^{\left(\nu_1^* x_{1t}+\nu_2^* x_{2t}+\theta^* x_{1t}x_{2t}\right)}}{z^*} = \prod_{t=1}^{T} \frac{a^{x_{1t}}\cdot b^{x_{2t}}\cdot c^{x_{1t}x_{2t}}}{z^*},$$

where $a = e^{\nu_1^*}$, $b = e^{\nu_2^*}$ *and* $c = e^{\theta^*}$ *and the normalizing constant is given by* $z^* = 1 + e^{\nu_1^*} + e^{\nu_2^*} + e^{\nu_1^*+\nu_2^*+\theta^*} = 1 + a + b + abc$. *Then the non-stationary MLE is inconsistent.*

Proof. We now explore how the non-stationary MLE behaves asymptotically. We know that probability for each window is

$$\mathbb{P}(X_w = x_w) = \frac{e^{\left(\nu_1^* s_{1w} + \nu_2^* s_{2w} + \theta^* l_w\right)}}{(z^*)^D} = \frac{a^{s_{1w}} \cdot b^{s_{2w}} \cdot c^{l_w}}{z^{*3}}.$$

Let us define

$$p_{ijk} = \mathbb{P}\left(s_{1w} = i, s_{2w} = j, l_w = k\right) = \frac{a^i \cdot b^j \cdot c^k}{z^{*3}} \cdot \#\{y : S_1(y) = i, S_2(y) = j, L(y) = k\},$$

where the last term counts all possible combinations of $y \in \{0,1\}^{2\times 3}$ that have the statistics $(S_1(y), S_2(y), L(y)) = (i, j, k)$. By Law of Large Numbers, we get

$$\left.\begin{array}{lll}
\frac{W_{111}}{W} \to p_{111} = \frac{3ab}{z^{*3}}(c) & \frac{W_{110}}{W} \to p_{110} = \frac{3ab}{z^{*3}}(2) & \frac{W_{11}}{W} \to \frac{3ab}{z^{*3}}(c+2) \\
\frac{W_{121}}{W} \to p_{121} = \frac{3ab^2}{z^{*3}}(2c) & \frac{W_{120}}{W} \to p_{120} = \frac{3ab^2}{z^{*3}}(1) & \frac{W_{12}}{W} \to \frac{3ab^2}{z^{*3}}(2c+1) \\
\frac{W_{211}}{W} \to p_{211} = \frac{3a^2b}{z^{*3}}(2c) & \frac{W_{210}}{W} \to p_{210} = \frac{3a^2b}{z^{*3}}(1) & \frac{W_{22}}{W} \to \frac{3a^2b^2c}{z^{*3}}(c+2) \\
\frac{W_{222}}{W} \to p_{222} = \frac{3a^2b^2c}{z^{*3}}(c) & \frac{W_{221}}{W} \to p_{221} = \frac{3a^2b^2c}{z^{*3}}(2) & \frac{W_{22}}{W} \to \frac{3a^2b^2c}{z^{*3}}(c+2)
\end{array}\right\} \quad \text{(C.36)}$$

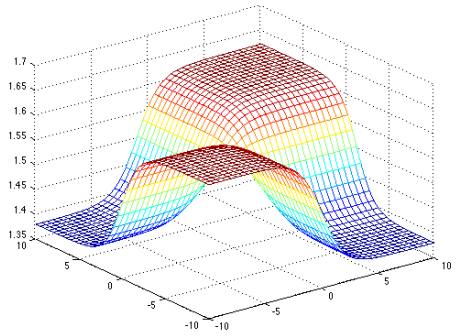
Using (C.35) and the asymptotic values of the ratios from above we get

$$h(\rho) \quad \longrightarrow \quad h^*(\rho), \qquad \text{(C.37)}$$

where $h^*(\rho)$ is given by

$$\begin{aligned} h^*(\rho) &= \frac{3ab}{z^{*3}}\left[c-3(c+2)\gamma_{11}(\rho)\right]+\frac{3ab}{z^{*3}}(a+b)\left[2c-3(2c+1)\gamma_{12}(\rho)\right] \\ &\quad +\frac{3ab}{z^{*3}}abc\left[2(c+1)-3(c+2)\gamma_{22}(\rho)\right] \\ &= \frac{3ab}{z^{*3}}\left(\left[c-(c+2)\frac{2\rho+1-\sqrt{8\rho+1}}{2(\rho-1)}\right]+(a+b)\left[2c-(2c+1)\frac{3\rho-\sqrt{\rho(8+\rho)}}{2(\rho-1)}\right]\right. \\ &\quad \left.+abc\left[2(c+1)-(c+2)\frac{4\rho-1-\sqrt{8\rho+1}}{2(\rho-1)}\right]\right). \end{aligned} \tag{C.38}$$

If $\hat{\rho}$ is a root of $h(\rho)$ and ρ_0 is a root of the limiting function $h^*(\rho)$, then $\hat{\rho}\to\rho_0$, since the roots of a polynomial vary continuously with the co-efficients [24]. This gives $\hat{\theta}^{\mathrm{MLE}}\to\theta_0=\log\rho_0$. Thus for the stationary distribution, θ_0 (the limiting value of $\hat{\theta}^{\mathrm{MLE}}$) depends on the true values of the nuisance parameters as well. Hence we cannot use any sort of a clever correction to correct the MLE (as we can do for the Gaussian case in the Neyman -Scott problem). For varying ${\nu_1}^*$ and ${\nu_2}^*$, with $\theta^*=1$, the mesh plot of the values of $\theta_0=\log\rho_0$ is shown in Figure below. For different values of ${\nu_1}^*$ and ${\nu_2}^*$, we get different limiting θ_0, and hence $\hat{\theta}^{\mathrm{MLE}}$ is not consistent. □



Limiting values of $\hat{\theta}^{MLE}$ as a function of varying ${\nu_1}^*$ and ${\nu_2}^*$ with $\theta^*=1$: $\hat{\theta}^{MLE}\nrightarrow\theta^*=1$.

C.2.2 CMLE with $D = 3$

Using (3.17) and (3.20), the conditional distribution to be optimized is given by

$$\psi_{good}(x|S=s) \quad = \quad \prod_{w:s_w \in \mathbb{S}_{good}} \frac{\exp(\theta l_w)}{\sum_{k=0}^{D} \exp(\theta k) \cdot h_{s_{1w},s_{2w}}(k)}.$$

Pruning away the degenerate windows, we are left with only the non-degenerate ones with window sums $(1,1)$, $(1,2)$, $(2,1)$ and $(2,2)$. The hypergeometric coefficients for $D = 3$ are given by:

$h_{1,1}(0) = h_{1,2}(1) = h_{2,1}(1) = h_{2,2}(1) = 6$ and $h_{1,1}(1) = h_{1,2}(0) = h_{2,1}(0) = h_{2,2}(2) = 3$

Hence the conditional distribution to be optimized is

$$\begin{aligned}
\psi_{good}(x|S=s) &= \frac{e^{W_{111}}}{(6+3e^{\theta})^{W_{11}}} \cdot \frac{e^{W_{121}}}{(3+6e^{\theta})^{W_{12}}} \cdot \frac{e^{W_{211}}}{(3+6e^{\theta})^{W_{21}}} \cdot \frac{e^{W_{221}+2W_{222}}}{(6e^{\theta}+3e^{2\theta})^{W_{22}}} \\
&\propto \frac{e^{W_{111}}}{(2+e^{\theta})^{W_{11}}} \cdot \frac{e^{W_{121}}}{(1+2e^{\theta})^{W_{12}}} \cdot \frac{e^{W_{211}}}{(1+2e^{\theta})^{W_{21}}} \cdot \frac{e^{W_{222}}}{(2+e^{\theta})^{W_{22}}}
\end{aligned}$$

$$\begin{aligned}
\mathcal{L}(\theta) = \log \psi_{good} &= \Big(\theta W_{111} - W_{11}\log(2+e^{\theta})\Big) + \Big(\theta W_{121} - W_{12}\log(2+e^{\theta})\Big) \\
&\quad + \Big(\theta W_{211} - W_{21}\log(2+e^{\theta})\Big) + \Big(\theta W_{222} - W_{22}\log(2+e^{\theta})\Big) + C,
\end{aligned}$$

where C is a constant.

The derivative of the above equation is

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \theta} &= W_{111} - W_{11}\frac{\rho}{2+\rho} + W_{121} - W_{12}\frac{2\rho}{1+2\rho} + W_{211} - W_{21}\frac{2\rho}{1+2\rho} + W_{222} - W_{22}\frac{\rho}{2+\rho} \\
&= \frac{2W_{111} - \rho W_{110}}{2+\rho} + \frac{W_{121} - 2\rho W_{120}}{1+2\rho} + \frac{W_{211} - 2\rho W_{210}}{1+2\rho} + \frac{2W_{222} - \rho W_{221}}{2+\rho} \\
&= \frac{(A+2B) + (2A+B-2C-D)\rho - (C+2D)\rho^2}{(2+\rho)(1+2\rho)}, \qquad \text{(C.39)}
\end{aligned}$$

where $A = 2(W_{111} + W_{222})$, $B = W_{121} + W_{211}$, $C = 2(W_{120} + W_{210})$ and $D = W_{110} + W_{221}$.

$\hat{\rho}$ is given by the positive root of the quadratic equation $(C+2D)\rho^2 - (2A+B-2C-$

$D)\rho - (A + 2B) = 0$ which is

$$\hat{\rho} = \frac{(2A + B - 2C - D) + \sqrt{(2A + B - 2C - D)^2 + 4(A + 2B)(C + 2D)}}{2(C + 2D)}. \tag{C.40}$$

Then the CMLE is given by $\hat{\theta}^{CMLE} = \log \hat{\rho}$:

$$\hat{\theta}^{CMLE} = \log \frac{(2A + B - 2C - D) + \sqrt{(2A + B - 2C - D)^2 + 4(A + 2B)(C + 2D)}}{2(C + 2D)}. \tag{C.41}$$

Asymptotic behavior of CMLE on stationary data

Using the asymptotic equations from (C.36), we get

$$\begin{aligned}
A = 2(W_{111} + W_{222}) &\rightarrow \frac{6ab}{z^{*3}} \cdot (1 + abc) \cdot c \\
B = W_{121} + W_{211} &\rightarrow \frac{6ab}{z^{*3}} \cdot (a + b) \cdot c \\
C = 2(W_{120} + W_{210}) &\rightarrow \frac{6ab}{z^{*3}} \cdot (a + b) \\
D = W_{110} + W_{221} &\rightarrow \frac{6ab}{z^{*3}} \cdot (1 + abc).
\end{aligned}$$

Let us denote $\frac{6ab}{z^{*3}}$ as κ, which gives

$$\begin{aligned}
(2A + B - 2C - D) &\rightarrow \kappa\left((2c - 1)(1 + abc) + (c - 2)(a + b)\right) \\
4(A + 2B)(C + 2D) &\rightarrow 4\kappa^2\left(2c(1 + abc)^2 + 2c(a + b)^2 + 5c(1 + abc)(a + b)\right)
\end{aligned}$$

and substituting in Equation (C.40), we get

$$\begin{aligned}
\hat{\rho} &= \frac{(2A + B - 2C - D) + \sqrt{(2A + B - 2C - D)^2 + 4(A + 2B)(C + 2D)}}{2(C + 2D)} \\
&\rightarrow \frac{(2c - 1)(1 + abc) + (c - 2)(a + b) + (2c + 1)(1 + abc) + (c + 2)(a + b)}{4(1 + abc) + 2(a + b)} \\
&\rightarrow \frac{4c(1 + abc) + 2c(a + b)}{4(1 + abc) + 2(a + b)} \\
&\rightarrow c.
\end{aligned}$$

Hence $\hat{\theta}^{\text{CMLE}} \to \log c = \theta^*$. Thus for a stationary distribution, the CMLE with $D = 3$ is consistent.

C.2.3 Simulation results

Results on simulated data are shown in Section 3.5.2.

Bibliography

[1] Asohan Amarasingham, Matthew T Harrison, Nicholas G Hatsopoulos, and Stuart Geman. Conditional modeling and the jitter method of spike resampling. *Journal of neurophysiology*, 107(2):517–531, 2012.

[2] Erling B Andersen. The asymptotic distribution of conditional likelihood ratio tests. *Journal of the American Statistical Association*, 66(335):630–633, 1971.

[3] Erling Bernhard Andersen. Asymptotic properties of conditional maximum-likelihood estimators. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 283–301, 1970.

[4] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.

[5] Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pages 1165–1188, 2001.

[6] James O Berger, Brunero Liseo, Robert L Wolpert, et al. Integrated likelihood methods for eliminating nuisance parameters. *Statistical Science*, 14(1):1–28, 1999.

[7] Christopher M Bishop et al. *Pattern recognition and machine learning*, volume 4. springer New York, 2006.

[8] Fienberg S.E Bishop, Y.M.M and P.W Holland. Discrete multivariate analysis: Theory and practice. 1975.

[9] David Blackwell and James B MacQueen. Ferguson distributions via pólya urn schemes. *The annals of statistics*, pages 353–355, 1973.

[10] David R Brillinger. Nerve cell spike train data analysis: a progression of technique. *Journal of the American Statistical Association*, 87(418):260–271, 1992.

[11] Carlos D Brody. Slow covariations in neuronal resting potentials can lead to artefactually fast cross-correlations in their spike trains. *Journal of Neurophysiology*, 80(6):3345–3351, 1998.

[12] Emery N Brown, Robert E Kass, and Partha P Mitra. Multiple neural spike train data analysis: state-of-the-art and future challenges. *Nature neuroscience*, 7(5):456–461, 2004.

[13] George Casella and Roger L Berger. *Statistical inference*, volume 2. Duxbury Pacific Grove, CA, 2002.

[14] Gary Chamberlain. Analysis of covariance with qualitative data. *The Review of Economic Studies*, 47(1):pp. 225–238, 1980.

[15] ES Chornoboy, LP Schramm, and AF Karr. Maximum likelihood identification of neural point process systems. *Biological cybernetics*, 59(4-5):265–275, 1988.

[16] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[17] David R Cox. Partial likelihood. *Biometrika*, 62(2):269–276, 1975.

[18] DR Cox and N Reid. Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–39, 1987.

[19] DR Cox and N Reid. A note on pseudolikelihood constructed from marginal densities. *Biometrika*, 91(3):729–737, 2004.

[20] Jan De Leeuw and Norman Verhelst. Maximum likelihood estimation in generalized rasch models. *Journal of Educational and Behavioral Statistics*, 11(3):183–196, 1986.

[21] Thomas S Ferguson. A bayesian analysis of some nonparametric problems. *The annals of statistics*, pages 209–230, 1973.

[22] Shigeyoshi Fujisawa, Asohan Amarasingham, Matthew T Harrison, and György Buzsáki. Behavior-dependent short-term assembly dynamics in the medial prefrontal cortex. *Nature neuroscience*, 11(7):823–833, 2008.

[23] Malay Ghosh. Inconsistent maximum likelihood estimators for the rasch model. *Statistics & Probability Letters*, 23(2):165–170, 1995.

[24] Gary Harris and Clyde Martin. Shorter notes: The roots of a polynomial vary continuously as a function of the coefficients. *Proceedings of the American Mathematical Society*, pages 390–392, 1987.

[25] Matthew T Harrison. Accelerated spike resampling for accurate multiple testing controls. *Neural computation*, 25(2):418–449, 2013.

[26] Matthew T Harrison, Asohan Amarasingham, and Wilson Truccolo. Spatiotemporal conditional inference and hypothesis tests for neural ensemble spiking precision. 2014.

[27] Jack Kiefer and Jacob Wolfowitz. Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *The Annals of Mathematical Statistics*, pages 887–906, 1956.

[28] Tony Lancaster. The incidental parameter problem since 1948. *Journal of econometrics*, 95(2):391–413, 2000.

[29] Bruce Lindsay, Clifford C Clogg, and John Grego. Semiparametric estimation in the rasch model and related exponential response models, including a simple latent class model for item analysis. *Journal of the American Statistical Association*, 86(413):96–107, 1991.

[30] Laura Martignon, Gustavo Deco, Kathryn Laskey, Mathew Diamond, Winrich Freiwald, and Eilon Vaadia. Neural coding: higher-order temporal patterns in the neurostatistics of cell assemblies. *Neural Computation*, 12(11):2621–2653, 2000.

[31] Radford M Neal. Markov chain sampling methods for dirichlet process mixture models. *Journal of computational and graphical statistics*, 9(2):249–265, 2000.

[32] Jerzy Neyman and Elizabeth L Scott. Consistent estimates based on partially consistent observations. *Econometrica: Journal of the Econometric Society*, pages 1–32, 1948.

[33] C Radhakrishna Rao. *Linear statistical inference and its applications*, volume 22. John Wiley & Sons, 2009.

[34] Nancy Reid. The roles of conditioning in inference. *Statistical Science*, pages 138–157, 1995.

[35] Elad Schneidman, Michael J Berry, Ronen Segev, and William Bialek. Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature*, 440(7087):1007–1012, 2006.

[36] Jayaram Sethuraman. A constructive definition of dirichlet priors. *Statistica Sinica*, 4:639–650, 1994.

[37] Jayaram Sethuraman and Ram C Tiwari. Convergence of dirichlet measures and the interpretation of their parameter. Technical report, Department of Statistics, Florida State University, 1981.

[38] J. Shao. *Mathematical Statistics*. Springer, 1999.

[39] Hideaki Shimazaki, Shun-ichi Amari, Emery N Brown, and Sonja Grün. State-space analysis of time-varying higher-order spike correlation for multiple neural spike train data. *PLoS computational biology*, 8(3):e1002385, 2012.

[40] Jonathon Shlens, Greg D Field, Jeffrey L Gauthier, Matthew I Grivich, Dumitru Petrusca, Alexander Sher, Alan M Litke, and EJ Chichilnisky. The structure of multi-neuron firing patterns in primate retina. *The Journal of neuroscience*, 26(32):8254–8266, 2006.

[41] Wilson Truccolo, Omar J Ahmed, Matthew T Harrison, Emad N Eskandar, G Rees Cosgrove, Joseph R Madsen, Andrew S Blum, N Stevenson Potter, Leigh R Hochberg, and Sydney S Cash. Neuronal ensemble synchrony during human focal seizures. *The Journal of Neuroscience*, 34(30):9927–9944, 2014.

[42] Aeilko H Zwinderman. Pairwise parameter estimation in rasch models. *Applied Psychological Measurement*, 19(4):369–375, 1995.