# Topics in Multivariate Statistics with Dependent Data

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Karl Oskar Ekvall

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Galin L. Jones, Adviser

February 2019

# Acknowledgements

I owe a lot to my advisor Galin Jones whose guidance has been instrumental in the writing of this dissertation. His attention to detail and focus on clarity in both thinking and writing have improved my work greatly. He was patient when I had trouble settling on a dissertation topic, encouraged me when I needed it, and always made time to talk. For this and much more I am most grateful.

I would like to thank Adam Rothman, Charles Doss, Charles Geyer, Dennis Cook, Erik Hjalmarsson, Hui Zou, and Liliana Forzani for interesting and fruitful discussions, several of which led to the improvement of chapters in this dissertation. Thanks also to Mark Fiecas for serving on the examining committee and to Brian Gray for giving me the freedom to work on something that interested me during a consulting project, eventually leading to Chapter 3. Financial support from the University of Minnesota, the Fulbright Commission in Stockholm, the American – Scandinavian Foundation, the Tom Hedelius Foundation, and Sixten Gemzéus's foundation is gratefully acknowledged.

My time in Minnesota would not have been nearly as enjoyable without the fellow graduate students in the School of Statistics. You are too many to list, but a special thanks to Aaron, Adam, Dan, Dootika, Haema, James, Matt, Mitch, and Sakshi.

It would not have been possible for me to write this dissertation without the support of my parents. Tack mamma och pappa för att ni alltid trott på mig och alltid ställt upp när det behövts.

Finally, I am grateful to my family, and in particular my wife Lin Fickling; for her support as I pursued my academic goals, but more importantly for always being there and for all our wonderful times together. It is with great joy I look forward to the next chapter in our life together.

# Dedication

*To Lin, Alma, and Oliver.*

# Contents

**List of Tables** **vi**

**List of Figures** **vii**

**1 Introduction** **1**

1.1 Multivariate regression models . . . . . 1

1.2 Vector autoregressions . . . . . 4

1.3 Markov chain Monte Carlo . . . . . 5

1.3.1 Metropolis–Hastings . . . . . 7

1.3.2 Gibbs samplers . . . . . 11

1.3.3 Variance estimation . . . . . 14

1.3.4 Starting and stopping . . . . . 16

**2 Consistent maximum likelihood estimation using subsets** **18**

2.1 Introduction . . . . . 18

2.2 Consistency using subsets of the full data . . . . . 22

2.3 Application to multivariate mixed models . . . . . 27

2.3.1 Longitudinal linear mixed model . . . . . 27

2.3.2 Logit–normal MGLMM . . . . . 33

2.4 Discussion . . . . . 38

**3 Maximum likelihood estimation of covariance matrices with separable correlation** **40**

3.1 Introduction . . . . . 40
3.2 Maximum likelihood . . . . . 42
3.3 Inference . . . . . 48
3.4 Simulations . . . . . 49
3.4.1 Convergence diagnostics . . . . . 51
3.5 Data example . . . . . 53
3.6 Discussion . . . . . 57

**4 Convergence complexity analysis of a collapsed Gibbs sampler for Bayesian vector autoregressions 58**
4.1 Introduction . . . . . 58
4.2 A collapsed Gibbs sampler . . . . . 62
4.3 Geometric ergodicity . . . . . 65
4.4 Asymptotic stability . . . . . 69
4.5 Discussion . . . . . 78

**References 80**

**A Consistent maximum likelihood estimation using subsets 88**
A.1 Theory . . . . . 88
A.1.1 Preliminary results . . . . . 88
A.1.2 Main results . . . . . 89
A.2 Applications . . . . . 92
A.2.1 Longitudinal linear mixed model . . . . . 94
A.2.2 Logit–normal MGLMM . . . . . 98

**B Maximum likelihood estimation of covariance matrices with separable correlation 112**
B.1 A model for sample means . . . . . 112
B.2 Additional simulations . . . . . 114
B.2.1 Convergence diagnostics . . . . . 114

**C Convergence complexity analysis of a collapsed Gibbs sampler for Bayesian vector autoregressions** **116**

C.1 Preliminary results . . . . . 116

C.2 Main results . . . . . 117

# List of Tables

3.1 Estimation Error and Test Size . . . . . 52

B.1 Estimation Error and Test Size II . . . . . 114

B.2 Convergence proportions . . . . . 115

# List of Figures

1.1 Output from Metropolis–Hastings example . . . . . 11
1.2 Output from Gibbs example . . . . . 13

3.1 Estimated covariances for dissolved oxygen data . . . . . 56
3.2 Estimated variances for dissolved oxygen data . . . . . 56

# Chapter 1

# Introduction

Classical multivariate statistics focuses on vector-valued data where every element in a vector is a measurement from the same object or individual [2]. Here we will use a more encompassing definition and call a statistical model multivariate if it is a model for the joint distribution of a collection of observations; the observations can but need not be from the same object or individual, and the distribution of interest can be conditional on some other set of variables. The unifying theme for the settings we consider is that there is dependence between observations that motivates joint modeling. All models are classical in the sense that observations are real-valued and the interest is in estimating and making inference about an unknown parameter living in a finite-dimensional Euclidean space. We start with an introduction to the multivariate regression models studied in Chapters 2 and 3.

## 1.1 Multivariate regression models

A multivariate regression is a model for the conditional distribution of a collection of response variables (responses) $Y$ given a collection of predictor variables (predictors) $X$. We will work with two classes of such models: multivariate generalized linear mixed models (MGLMMs), which serve as motivating examples in Chapter 2, and multivariate linear regression models with separable correlation which are the focus of Chapter 3.

In Chapter 2 we develop a general theory for consistency of maximum likelihood estimators (MLEs) that does not require the observations to be independent or from a stationary

stochastic process, which is otherwise common in the literature. The theory is applied to two MGLMMs, establishing consistency of MLEs. In an MGLMM, all responses are conditionally independent given random effects and predictors. The responses can be of different types, some discrete and some continuous, for example, and dependence between them is modeled using the random effects. Thus, MGLMMs handle multivariate, mixed-type response regressions without assuming that responses of different types are independent or equally well modeled separately. More formally, given $X = [x_1, \ldots, x_n]^\mathsf{T} \in \mathbb{R}^{n \times p}$, a design matrix $Z \in \mathbb{R}^{n \times r}$, and $U \in \mathbb{R}^r$, a multivariate normal vector of random effects with mean zero and covariance matrix $\Sigma$, the components (elements) in the vector of responses $Y = (Y_1, \ldots, Y_n)$ are conditionally independent with exponential family densities

$$f_{\beta,i}(y_i \mid u, x) = k_i(y_i, \tau_i) \exp\left(\frac{y_i \vartheta_i(l_i) - c_i(\vartheta_i(l_i))}{\tau_i}\right),$$

where for $\beta \in \mathbb{R}^p$ and $i = 1, \ldots, n$, $l_i = x_i^\mathsf{T}\beta + z_i^\mathsf{T}u$, $\vartheta_i$ is the natural parameter (as a function of $l_i$), $c_i$ the cumulant function, $\tau_i$ a dispersion parameter, and $k_i(y_i, \tau_i)$ ensures $f_{\beta,i}(y_i \mid u, x)$ integrates to one. We will often assume canonical links, which in our notation means that $\vartheta_i$ is the identity function, for every $i$. Letting $\phi_\Sigma$ denote the density for the multivariate normal distribution with mean 0 and covariance matrix $\Sigma$, the distribution for $Y \mid X$ has density

$$f_\theta(y \mid x) = \int_{\mathbb{R}^r} \prod_{i=1}^n f_{\beta,i}(y_i \mid u, x) \phi_\Sigma(u) \, \mathrm{d}u.$$

Without further restrictions, the $(p + r(r+1)/2)$-dimensional parameter in this model is $\theta = (\beta, \Sigma)$. More generally, we assume that $\theta \in \Theta \subseteq \mathbb{R}^d$ for some $d \leq p + r(r+1)/2$ and let $\beta = \beta(\theta)$ and $\Sigma = \Sigma(\theta)$. The dispersion parameters $\tau_1, \ldots, \tau_n$ are typically treated as known and hence not included in $\theta$ (but see the LMM example in Chapter 2 for an exception).

When $k_i$, $\tau_i$, and $c_i$ are the same for all $i$, the MGLMM reduces to a GLMM where all components of the response vector are of the same type. An important special case of the GLMM is the LMM. In an LMM $f_{\beta,i}(y_i \mid u, x)$ is normal for every $i$, implying that $Y \mid X, U$

is multivariate normal with diagonal covariance matrix $\text{diag}(\tau_1, \ldots, \tau_n)$. Thus, $Y \mid X$ is multivariate normal as well, with mean $X\beta$ and covariance matrix $\text{diag}(\tau_1, \ldots, \tau_n) + Z\Sigma Z^\mathsf{T}$.

Unless all responses are normal, i.e. the MGLMM is really a LMM, the $r$-dimensional integral in the expression for $f_\theta(y \mid x)$ typically does not admit a closed form expression. When it does not, many things are more complicated. Likelihood-based estimation and inference is often either time-consuming or inexact, depending on if the integral is evaluated with numerical accuracy or crudely approximated, and theoretical properties of the distribution of $Y \mid X$ are often not well understood. In particular, in contrast to an LMM, the effects of the random effects design ($Z$ and $\Sigma$) on the distribution of $Y \mid X$ can be difficult to fully appreciate. Among other things, it is often difficult to assess what happens as $n$ tends to infinity. Recall, $n$ is here the length of the vector of all responses and not the number of independent observations, so it is not clear that classical asymptotics apply. This is one reason MGLMMs serve as motivating examples for the theory in Chapter 2.

The multivariate linear regression in Chapter 3 can be formulated as a special case of the LMM, and hence of the MGLMM. However, doing so leads to an awkward parameterization and it is more natural to directly define the distribution of $Y \mid X$ without using random effects and without stacking all responses in one long vector. Thus, in Chapter 3 we instead assume that there are $n$ independent $q-$dimensional multivariate normal vectors $Y_1, \ldots, Y_n$ with means $\mathbb{E}(Y_i \mid X) = \mathcal{B}^\mathsf{T} x_i$ and common covariance matrix $\text{cov}(Y_i \mid X) = \Sigma$, where $\mathcal{B} \in \mathbb{R}^{p \times q}$ and, as before, $X = [x_1, \ldots, x_n]^\mathsf{T} \in \mathbb{R}^{n \times p}$. The parameterization in Chapter 3 further assumes that the covariance matrix factors as $\Sigma = W(V \otimes U)W$, where $\otimes$ is the Kronecker product, $W$ is a diagonal matrix with positive entries, and $U$ and $V$ are correlation matrices of smaller dimension than $\Sigma$. This parameterization, also known as separable correlation, is often considered in the spatiotemporal literature. Our main contribution is a new algorithm for maximum likelihood estimation of covariance matrices with separable correlation. We also discuss some convenient properties of separable correlation that have been largely ignored in the literature.

The next section introduces the model considered in Chapter 4.

## 1.2 Vector autoregressions

For $t = 1, \ldots, n$, let $x_t \in \mathbb{R}^p$ be a vector of non-stochastic predictors and $Y_t \in \mathbb{R}^r$ be a stochastic process satisfying

$$Y_t = \sum_{i=1}^{q} \mathcal{A}_i^{\mathsf{T}} Y_{t-i} + \mathcal{B}^{\mathsf{T}} x_t + \varepsilon_t,$$

where $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. $\mathcal{N}(0, \Sigma)$, $\mathcal{A}_i \in \mathbb{R}^{r \times r}$ $(i = 1, \ldots, q)$, and $\mathcal{B} \in \mathbb{R}^{p \times r}$. We say that $Y_t$ is a vector autoregression of order $q$, or a VAR($q$). Upon defining $\mathcal{A} = [\mathcal{A}_1^{\mathsf{T}}, \ldots, \mathcal{A}_q^{\mathsf{T}}]^{\mathsf{T}} \in \mathbb{R}^{qr \times r}$ and $z_t = [y_{t-1}^{\mathsf{T}}, \ldots, y_{t-q}^{\mathsf{T}}]^{\mathsf{T}} \in \mathbb{R}^{qr}$ $(t = 1, \ldots, n)$ the defining equation for the VAR($q$) can be written more compactly as $Y_t = \mathcal{A}^{\mathsf{T}} z_t + \mathcal{B}^{\mathsf{T}} x_t + \varepsilon_t$. If $\mathcal{A} = 0$ then the VAR reduces to a model for $n$ independent observations of an $r$-dimensional response in the multivariate linear regression model introduced earlier. More generally, the VAR can be thought of as a regression model where the predictors include not only exogenous variables but also past values of the response. This somewhat loose statement can be formalized by comparing the likelihood functions of the two models. Indeed, assuming that the starting point of the process, $z_1$, is non-stochastic, the likelihood for $Y = [y_1, \ldots, y_n]^{\mathsf{T}} \in \mathbb{R}^{n \times r}$ from the VAR($q$) is

$$f(Y \mid \mathcal{A}, \mathcal{B}, \Sigma) \propto |\Sigma|^{-n/2} \exp\left( -\frac{1}{2} \sum_{i=1}^{n} \left[ Y_t - \mathcal{A}^{\mathsf{T}} z_t - \mathcal{B}^{\mathsf{T}} x_t \right]^{\mathsf{T}} \Sigma^{-1} \left[ Y_t - \mathcal{A}^{\mathsf{T}} z_t - \mathcal{B}^{\mathsf{T}} x_t \right] \right).$$

This is the same likelihood as that for $n$ observations in the classical multivariate linear regression with design matrix $[Z, X]$ and coefficient matrix $[\mathcal{A}^{\mathsf{T}}, \mathcal{B}^{\mathsf{T}}]^{\mathsf{T}}$. It follows that once the data is observed and treated as fixed, much of the analysis in a VAR can be carried over without change to the multivariate linear regression model, and vice versa. However, doing so is not always appropriate, as the work in Chapter 4 illustrates. There, we consider a Bayesian version of the VAR that incorporates prior information on the parameters $\mathcal{A}$, $\mathcal{B}$, and $\Sigma$ . The prior distributions are motivated by applications to macroeconomic and financial time series and are different from common choices for the multivariate linear regression model. The main contributions in Chapter 4 are a new Markov chain Monte Carlo

(MCMC) algorithm for Bayesian VARs with predictors and theoretical guarantees for that algorithm. We prove that the Markov chain generated by our algorithm converges quickly to its stationary distribution and that the algorithm can be expected to work well in both small and large samples. The next section provides a relatively non-technical introduction to MCMC.

## 1.3 Markov chain Monte Carlo

Many statistical methods at some stage require sampling from a probability distribution. In ideal cases, one can generate independent and identically distributed (i.i.d.) observations from the desired distribution. In many practically relevant models, however, this is either not possible or prohibitively time consuming. Fortunately, in a wide range of settings, Markov chain Monte Carlo (MCMC) can be used in place of i.i.d. sampling [cf. 8, 61]. Because of this, after the seminal paper by Gelfand and Smith [22], MCMC has become integral to Bayesian analysis where such complicated distributions often arise, but is also important in some frequentist settings [24].

We illustrate with one of the most common purposes of generating (pseudo-)random numbers. If $h$ is some real-valued function and $X$ is a random variable, then we want to calculate an expectation in the form

$$\mu_h := \mathbb{E}(h(X)) < \infty.$$

For different choices of $h$, many problems both in statistics and other disciplines can be written in this way. If $\mu_h$ is complicated to calculate analytically or by numerical integration, an alternative is to generate $X_1, \ldots, X_m$, independent with the same distribution as $X$, and approximate $\mu_h$ by

$$\hat{\mu}_h := \frac{1}{m} \sum_{i=1}^{m} h(X_i).$$

Here, $X$ can be a random vector or something more general, but the main points of our

discussion are equally well grasped thinking of $X$ as a random number. Several useful properties of $\hat{\mu}_h$ are immediate from classical statistical results: (i) $\hat{\mu}_h$ is unbiased, (ii) the law of large numbers (LLN) says that $\hat{\mu}_h$ is consistent as $m$ tends to infinity, and (iii) if $\mathbb{E}(h(X)^2) < \infty$, the central limit theorem (CLT) says that $\hat{\mu}_h$ is approximately normally distributed for large $m$. The LLN is important because it says, loosely speaking, that we can improve the estimate by simulating longer, and the the CLT is important because it lets us quantify the uncertainty in the estimate. In particular, $\mathrm{var}(\hat{\mu}_h)$ can be estimated by $s_h^2/m$, where $s_h^2 = m^{-1}\sum_{i=1}^m (h(X_i) - \hat{\mu}_h)^2$, and approximate confidence intervals for $\mu_h$ can be created by appealing to the CLT. The only difference between this and classical statistics is that the variables are generated in a computer. Methods that use the generation of random numbers are called Monte Carlo (MC) methods and $\hat{\mu}_n$ is called a MC estimator of $\mu_h$. MC methods with i.i.d. variables are sometimes called ordinary MC or i.i.d. MC.

MCMC is similar to ordinary MC but with the key difference that the generated variables $X_1, \ldots, X_m$ need be neither independent, nor have the same distributions. Rather, as the name suggests, they are generated as a Markov chain. Since i.i.d. variables form a Markov chain, ordinary MC is a special case of MCMC. The power of MCMC, however, is that the useful properties of $\hat{\mu}_h$ discussed above (LLN and CLT) continue to hold for much more general chains. Such chains can be constructed in many cases where i.i.d. sampling is infeasible and, hence, MCMC is more widely applicable than ordinary MC. For an introduction to Markov chain theory see [54, 63].

We say that $X_1, X_2, \ldots$ is a Markov chain if the conditional distribution of $X_i$ given $X_1, \ldots, X_{i-1}$, $i \geq 2$, depends only on $X_{i-1}$; this is known as the Markov property. It follows from the Markov property that a Markov chain is characterized by its initial distribution (the distribution of $X_1$), and its transition kernel $P$ defined by

$$P(x, A) = \mathsf{P}(X_i \in A \mid X_{i-1} = x),$$

for any subset $A$ of the state space, the set in which the Markov chain takes its values. If the initial distribution and kernel is such that the distribution of $X_2$ is the same as that of $X_1$

we say that the initial distribution is invariant for the kernel. More generally, a distribution $F$ is invariant for the transition kernel $P$ if $X_i \sim F$ implies $X_{i+1} \sim F$, $i \geq 1$. Using this definition it can be shown that if the initial distribution is invariant for $P$, then in fact every $X_i$, $i \geq 1$ has the same distribution. Such Markov chains are called stationary, and they are indeed stationary in the usual sense for stochastic processes.

Let $F_X$ denote the distribution of $X$ and suppose we generate a Markov chain with initial distribution $F_X$ and a kernel $P$ for which $F_X$ is invariant. Then by the preceding discussion $X_1, \ldots, X_m$ are possibly dependent but identically distributed random variables with the same distribution as $X$. Hence, $\hat{\mu}_h$ is an unbiased estimator of $\mu_h$. Moreover, it can be shown that under additional conditions $\hat{\mu}_h$ is consistent and asymptotically normal with variance $\kappa_h^2/m$, where, with $\sigma_h^2 = \operatorname{var}(h(X))$,

$$\kappa_h^2 = \sigma_h^2 + 2\sum_{i=1}^{\infty} \operatorname{cov}(h(X_1), h(X_{1+i})).$$

Recall, however, that MCMC is often used precisely because sampling from $F_X$ is infeasible. Hence, generating the stationary chain is also infeasible as it requires $X_1 \sim F_X$. Fortunately, it can be shown that if $\hat{\mu}_h$ is consistent and satisfies a CLT when the initial distribution is $F_X$, then the same is true for any other initial distribution. This tells us that if the simulation is long enough ($m$ is large enough), then the starting value $X_1$, whether selected at random or set to some fixed number, is unimportant. In practice this argument is somewhat problematic because it is hard to know what long enough means, but let us ignore that for now—we will return to the issue of starting values later. Next we outline how to, given a target distribution, generate Markov chains for which that distribution is invariant. The focus is on two of the most common algorithms, the Metropolis–Hastings (MH) algorithm and the Gibbs sampler.

### 1.3.1 Metropolis–Hastings

Suppose that we want to estimate $\mu_h$ and know the target density only up to a normalizing constant, or up to scaling. That is, we know that $X$ has a distribution $F_X$ with density $f_X$

satisfying

$$f_X(x) = cp(x)$$

for some $c > 0$ and function $p$. Of course, the fact that densities must integrate to one tells us that $c = 1/\int p(x)\,\mathrm{d}x$, but if $p$ is complicated to integrate, $c$ is not known in any practical sense. Thus, even though $p$ can be evaluated we cannot compute $c$ or easily sample from $F_X$. Settings like this are exceedingly common in Bayesian statistics.

Given an unnormalized density like $p$, the MH algorithm (Algorithm 1.1) constructs a Markov chain with a transition kernel for which $F_X$ is invariant. That is, the MH algorithm lets us sample approximately from $F_X$ even though we only know the corresponding density $f_X$ up to a normalizing constant. The algorithm transitions between states as follows: given that the chain is at state $X_i = x_i$, a move is proposed to a new state $y$ drawn from some distribution with density $q(y \mid x_i)$. Then, the move is either accepted, which happens with a probability that depends on $p(y)$, $p(x_i)$, $q(y \mid x_i)$, and $q(x_i \mid y)$, or rejected. If the move is rejected, the chain stays in the same place for one iteration and then another move is proposed. Since the proposal distribution and the acceptance probability depend on the current state but no previous states the algorithm indeed generates a Markov chain.

To implement the MH algorithm one needs to select a proposal distribution (density) $q(\cdot \mid \cdot)$. Any proposal distribution having support containing that of $p$ will lead to the chain having the right invariant distribution. However, the convergence properties of the chain are in general affected by the choice. A discussion of standard strategies for selecting the proposal distribution is provided by Robert and Casella [61].

The following example illustrates how the MH algorithm can be used in Bayesian statistics. The example is chosen to be simple enough that no MCMC is actually required, which makes the results easy to verify using numerical integration or i.i.d. sampling, but also complicated enough to convey some key ideas.

**Example 1.3.1** (Bayesian estimation of normal mean and variance with conjugate priors)**.** Suppose we have 30 independent observations $y_1, \ldots, y_{30}$ drawn from a normal distribution

**Algorithm 1.1** Metropolis–Hastings

1: *Input:* Starting value $X_1$ and length of chain $m$
2: **for** $i = 1, \ldots, m$ **do**
3: Given $X_i = x_i$, draw proposal $y$ from a distribution with density $q(y \mid x_i)$.
4: Calculate the Hastings ratio

$$r(x_i, y) = \frac{p(y)q(x_i \mid y)}{p(x_i)q(y \mid x_i)}.$$

5: Randomly pick the next state $X_{i+1}$ by accepting or rejecting proposal $y$:

$$X_{i+1} = \begin{cases} y & \text{w. prob. } \alpha(x_i, y) = \min[1, r(x_i, y)] \\ x_i & \text{w. prob. } 1 - \alpha(x_i, y) \end{cases}$$

6: **end for**

with the unknown mean $\mu^\star = 1$ and variance $1/\tau^\star = 1$, where the stars are used to indicate unknown population values. We wish to incorporate prior information about the parameters and specify the prior distribution in two stages by letting $\mu \mid \tau \sim \mathcal{N}(a, \tau^{-1}b^{-1})$ and $\tau \sim \mathcal{G}(c, d)$, where $\mathcal{G}(c, d)$ denotes the gamma distribution with mean $c/d$. That is,

$$f(\mu \mid \tau) = (2\pi)^{-1/2}(b\tau)^{1/2}e^{-b\tau(\mu-a)^2/2} \quad \text{and} \quad f(\tau) = \frac{d^c}{\Gamma(c)}\tau^{c-1}e^{-\tau d}I(\tau > 0),$$

for hyperparameters $a \in \mathbb{R}$, $b > 0$, $c > 0$, and $d > 0$, where $\Gamma(\cdot)$ denotes the gamma function. In an application the hyperparameters would be chosen to reflect the prior beliefs about $\mu$ and $\tau$. For concreteness, we here somewhat arbitrarily set them to $a = 0$, $b = c = d = 1$.

In Bayesian statistics the interest is in the posterior density $f(\mu, \tau \mid y)$. By standard rules for probability densities, the posterior density satisfies

$$\begin{aligned} f(\mu, \tau \mid y) &= \frac{f(y \mid \mu, \tau)f(\mu \mid \tau)f(\tau)}{f(y)} \\ &\propto f(y \mid \mu, \tau)f(\mu \mid \tau)f(\tau) \end{aligned}$$

where $\propto$ means equality holds up to scaling by a quantity not depending on $\mu$ or $\tau$. Multi-

plying the prior densities by the likelihood

$$f(y \mid \mu, \tau) = (2\pi)^{-n/2}\tau^{n/2} \exp\left(-\frac{\tau}{2}\sum_{i=1}^{n}(y_i - \mu)^2\right)$$

we find that the posterior satisfies

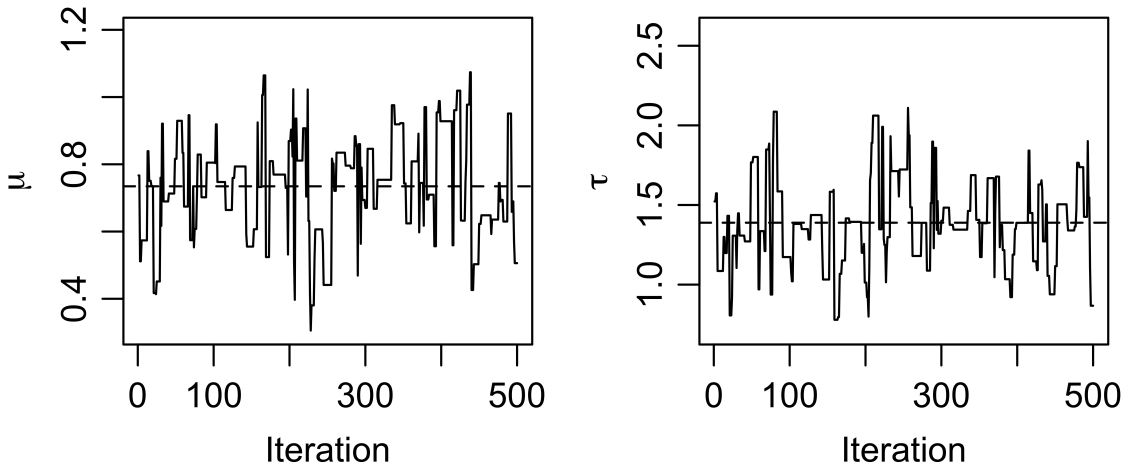
$$f(\mu, \tau \mid y) \propto \tau^{n/2} \exp\left(-\frac{\tau}{2}\sum_{i=1}^{n}(y_i - \mu)^2 - \tau\mu^2/2 - \tau\right). \tag{1.1}$$

Let us define $\theta = (\mu, \tau)$ and denote the expression in (1.1) by $p(\theta; y)$. The density from which we want to sample is $f(\theta \mid y) = p(\theta; y)/\int p(\theta; y)\,\mathrm{d}\theta$. In this example the state space of the Markov chain to be generated is the parameter set of the model for the data $y$—this is typical for Bayesian statistics. Accordingly, for the remainder of this example we let $\Theta = \mathbb{R} \times (0, \infty)$ denote the state space, $\theta_i = (\mu_i, \tau_i)$ the $i$th state of the Markov chain, and $\xi$ the proposed value with conditional density $q(\xi \mid \theta_i)$ in the MH algorithm, so as to not confuse it with the observed data.

We are ready to define a MH algorithm for exploring the distribution with density $f(\theta \mid y)$ on $\Theta$, which as mentioned amounts to selecting a proposal distribution. We take the proposal distribution to be multivariate normal and centered at the current state. More precisely, $\xi \mid \theta_i \sim \mathcal{N}(\theta_i, 0.25I_2)$. The covariance matrix is selected with some experimentation to result in an acceptance rate (the proportion of accepted proposals) of roughly 0.25 (see Rosenthal [66] for a motivation of this number). The starting value is set to $\theta_0 = (\mu_0, \tau_0) = (\bar{y}, 1/s_y^2)$, where $\bar{y}$ is the sample average and $s_y^2$ the biased sample variance; these are the MLEs of the parameters $\mu$ and $\tau$. The hope is that the MLEs are in a high-density region of the posterior.

Figure 1.1 shows the output from running the MH algorithm for 500 iterations. The short, flat segments where the chain stays in the same place for a few iterations correspond to rejected proposals. The sample paths can be used to get an MCMC estimate of $\int h(\theta)f(\theta)\,\mathrm{d}\theta$ for any function $h$ for which the integral exists. If we are, for example, interested in the Bayes estimate of $\mu$, $\mathbb{E}(\mu \mid y)$, then we can take $h(\theta) = h((\mu, \tau)) = \mu$ and estimate $\int h(\theta)f(\theta \mid$

Figure 1.1: Output from Metropolis–Hastings example




Horizontal dashed lines indicate sample averages of the plotted sample paths.

$y)\,\mathrm{d}\theta = \int \mu f(\mu \mid y)\,\mathrm{d}\mu$ by $\sum_{i=1}^{500} h(\theta_i)/500 = \sum_{i=1}^{500} \mu_i/500$. This sample average, which is indicated by a dashed line in 1.1, is 0.73. This can be compared to the MLE $\bar{y} = 0.77$, and the true mean $\mu^\star = 1$. We will return to this example below when implementing a Gibbs sampler.

### 1.3.2 Gibbs samplers

Suppose that the random variable $X$ which distribution $F_X$ we would like to sample from is multivariate, i.e. a random vector. We can then split $X$ into sub-vectors, say $X = (X^{(1)}, \ldots, X^{(s)})$; each $X^{(i)}$ can be univariate or multivariate. We will call $X^{(1)}, \ldots, X^{(s)}$ the components of $X$. To implement a Gibbs sampler, we need to be able to sample from the conditional distribution of any one component given the rest, also known as the component's full conditional distribution. The Gibbs sampler proceeds by updating the states of the components iteratively, drawing new states from the full conditional distributions as detailed in Algorithm 1.2.

To implement a Gibbs sampler one has to select how to partition the vector $X$ into components. In other words, one has to select which elements of the Markov chain to

update together. As an example, if we have a three-dimensional target distribution, then $X_i = (X_{i,1}, X_{i,2}, X_{i,3})$ can be updated in four ways: (i) each element is updated separately, (ii) $X_{i,1}$ and $X_{i,2}$ are updated together (and $X_{i,3}$ is updated by itself), (iii) $X_{i,1}$ and $X_{i,3}$ are updated together, or (iv) $X_{i,2}$ and $X_{i,3}$ are updated together. Which elements are updated together can affect the convergence properties of the chain and hence it can be worthwhile to consider different configurations. There is no general rule to guide the selection, though there is some evidence that strongly correlated components should be updated together. We next illustrate how a Gibbs sampler can be implemented in practice.

---

**Algorithm 1.2** Gibbs sampler

1: *Input:* Starting value $X_1 = x_1$ and length of chain $m$
2: **for** $i = 1, \ldots, m$ **do**
3: **for** $j = 1, \ldots, s$ **do**
4: Draw $x_{i+1}^{(j)}$ from the distribution of

$$X^{(j)} \mid \left( X^{(1)} = x_{i+1}^{(1)}, \ldots, X^{(j-1)} = x_{i+1}^{(j-1)}, X_i^{(j+1)} = x_i^{(j+1)}, \ldots, X^{(s)} = x_i^{(s)} \right)$$

5: **end for**
6: **end for**

---

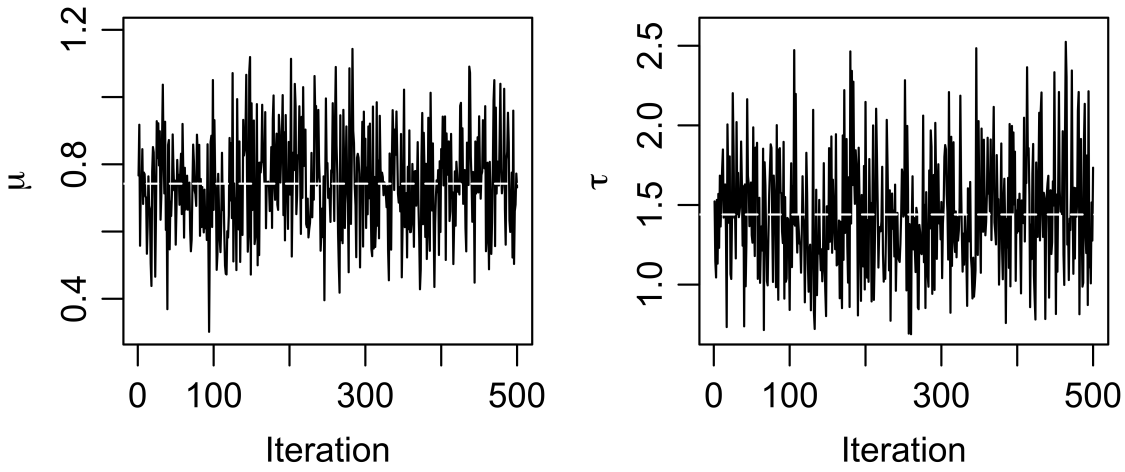
**Example 1.3.1** (continued)**.** Recall, the posterior distribution that we are considering has a density that is known up to scaling:

$$f(\mu, \tau \mid y) \propto \tau^{n/2} \exp\left( -\frac{\tau}{2} \sum_{i=1}^{n} (y_i - \mu)^2 - \tau\mu^2/2 - \tau \right).$$

Since the chain is bivariate, our only choice is to update $\mu$ and $\tau$ separately if we want to implement a Gibbs sampler—there is no other way to split the chain into components. To find the necessary full conditional distributions, notice that for a fixed $\tau$ the exponent of $f(\mu, \tau \mid y)$ is a quadratic function in $\mu$, and for for a fixed $\mu$ the exponent is linear in $\tau$. Using this, one can show that

$$\mu \mid \tau, y \sim \mathcal{N}\left( (n+1)^{-1} \sum_{i=1}^{n} y_i, \tau^{-1}(n+1)^{-1} \right)$$

Figure 1.2: Output from Gibbs example




Horizontal dashed lines indicate sample averages of the plotted sample paths.

and

$$\tau \mid \mu, y \sim \mathcal{G}(n/2+1, \sum_{i=1}^{n}(y_i-\mu)^2/2 + \mu^2/2 + 1).$$

We implement a Gibbs sampler that first updates $\tau$ and then $\mu$. There is no specific reason for this choice—updating $\mu$ first works equally well. As with the MH algorithm, the starting value is $\theta_0 = (\bar{y}, 1/s_y^2)$. Notice, however, that with the Gibbs sampler only the starting value for $\mu$ matters since $\tau$ is updated first, and the distribution from which $\tau_2$ is drawn does not depend on $\tau_1$. Figure 1.2 shows the output from running the Gibbs sampler for 500 iterations. The sample paths there depicted can be used in the same way as those of the MH algorithm. Of course, estimates based on the output in Figure 1.2 will be different from those based on the output in Figure 1.1. In this particular case, one might prefer estimates based on the Gibbs sampler since, for both $\mu$ and $\tau$, the sample path depicted in Figure 1.2 looks more like that of uncorrelated variables than that in Figure 1.1. This indicates the variance of estimators based on the Gibbs chain may be lower than that of those based on the MH chain in this example. The Gibbs chain is also exploring a larger part of the state space than the MH chain in the first 500 iterations.

### 1.3.3 Variance estimation

We have said that MCMC often leads to a consistent and asymptotically normal $\hat{\mu}_h$, and we have shown how to construct Markov chains that have the desired invariant distribution. However, not all chains with the right invariant distribution give a consistent and asymptotically normal $\hat{\mu}_h$. Conditions that ensure these properties are fairly technical and in general have to be verified on a case-by-case basis [40]. For the remainder of this introduction we assume that $\hat{\mu}_h$ is both consistent and asymptotically normal. That is, we assume that approximately for large $m$,

$$\hat{\mu}_h \sim \mathcal{N}(\mu_h, \kappa_h^2/m),$$

where as before

$$\kappa_h^2 = \sigma_h^2 + 2\sum_{i=1}^{\infty} \mathrm{cov}(h(X_1), h(X_{1+i})). \tag{1.2}$$

Under i.i.d. sampling, the infinite sum of autocovariances in (1.2) vanishes and $\kappa_h^2 = \sigma_h^2$ can be estimated by the sample variance $s_h^2$. In contrast, for more general MCMC algorithms the infinite sum typically does not vanish and $\kappa_h^2$ is more challenging to estimate than $\sigma_h^2$. It is also important to notice that $\kappa_h^2$ and $\sigma_h^2$ quantify different things: $\sigma_h^2$ is a characteristic of the invariant distribution only but $\kappa_h^2$ depends on the joint distribution of all the variables in the chain. In particular, two stationary Markov chains with the same invariant distribution but different autocovariances will lead to the same $\sigma_h^2$ but different $\kappa_h^2$. Since $\kappa_h^2$ directly determines the uncertainty in the estimate $\hat{\mu}_h$, it is desirable to, all else equal, pick an algorithm that leads to small (or negative) autocovariances. When we return to our example below, we will see that two perfectly reasonable MCMC algorithms can, for the same problem, generate chains that have the same invariant distribution but substantially different autocovariances.

In most realistic settings $\kappa_h^2$ is unknown and must be estimated using the Markov chain if we hope to say something about the uncertainty in $\hat{\mu}_h$. There are several methods for

estimating $\kappa_h^2$ that use the same chain used to estimate $\mu_h$ [20, 23, 75]. Here, we will give a short introduction to the method of batch means which gives an estimator that is easy and fast to compute. Suppose that $b$ is a divisor of $m$ and define, for $k = 1, \ldots, m_b = m/b$, the batch mean

$$\hat{\mu}_{h,k} = b^{-1} \sum_{i=1}^{b} h(X_{b(k-1)+i}).$$

Thus, $\hat{\mu}_{h,1}$ is the MCMC estimator of $\mu_h$ based on only the first $b$ variables in the chain, $\hat{\mu}_{h,2}$ is that based on only the next $b$ variables, and so on. The batch means estimator of $\kappa_h^2$ is

$$\hat{\kappa}_h^2 = \frac{\sqrt{b}}{m_b} \sum_{i=1}^{m_b} (\hat{\mu}_{h,i} - \hat{\mu}_h)^2.$$

When deciding on the number of batches there is a trade-off between estimating the mean in each batch precisely, which requires a large $b$, and estimating the variability among batches precisely, which requires a large $m_b$. Geyer [25] suggests that 20 – 30 batches is enough for most applications. Another common choice is to pick the number of batches to be approximately $m_b = \sqrt{m}$. After computing $\hat{\kappa}_h^2$, confidence intervals for $\mu_h$, and corresponding tests, can be computed using a $t$-distribution; for $\alpha \in (0, 1)$, a $100(1 - \alpha)\%$ confidence interval for $\mu_h$ is given by

$$\hat{\mu}_h \pm t_{b-1,1-\alpha/2} \sqrt{\hat{\kappa}_h^2 / m},$$

where $t_{b-1,1-\alpha/2}$ is the $(1-\alpha/2)$th quantile of the $t$-distribution with $b-1$ degrees of freedom.

In practice, we are rarely interested in estimating just one characteristic of the target distribution, so $h$ is usually multivariate, or vector-valued. For such settings a multivariate analysis that takes into account the dependence between the components of the multivariate estimator $m^{-1} \sum_{i=1}^{m} h(X_i)$ is appropriate. Methods for multivariate output analysis are available [74].

**Example 1.3.1** (continued)**.** We have previously in this example generated two Markov chains that have the desired invariant distribution and either could be used to estimate $\mu_h$ for some $h$ of interest. Let us continue to focus on the choice $h(\theta) = h((\mu, \tau)) = \mu$, indicating

the quantity of interest is the posterior mean $\mathbb{E}(\mu \mid y)$. For both the MH algorithm and the Gibbs sampler, we estimate the posterior mean by the sample mean of the generated $\mu$-chain, i.e. the first component of the bivariate chain. The 500 iterations of the chains displayed in Figures 1.1 and 1.2 are not enough to get reliable estimates $\mu_h$ or $\kappa_h^2$. After running the chains for 50,000 iterations, the estimate of $\mu_h$ based on the MH chain is 0.743 and the estimate based on the Gibbs chain is 0.742. Moreover, if we calculate the sample variance of each chain, $s_h^2$, they are both approximately 0.024. However, the batch means estimate of $\kappa_h^2$ is 0.160 for the MH chain and 0.021 for the Gibbs chain. This illustrates what was mentioned before, namely that $\mu_h$ and $\sigma_h^2$ are the same for both chains, but $\kappa_h^2$ is different since the autocovariances in the chains are different. The estimates of $\kappa_h^2$ indicate that in this particular example, the infinite sum in (1.2) is larger for the MH chain than the Gibbs sampler. Indeed, we noted earlier that the sample paths in Figure 1.2 look more like those of uncorrelated variables than those in Figure 1.1.

### 1.3.4 Starting and stopping

Whether one is using a MH algorithm, a Gibbs sampler, or something else, deciding where to start the algorithm and when to stop sampling is usually up to the user. It is generally a good idea to pick a starting point in a high-density region of the target distribution. If we could start the chain with the invariant distribution we would likely get a starting value in such a region, and hence, intuition suggests, those points are good to start at. Of course, in many problems we do not have a good guess for a high-density region and then this method is not applicable. Another common practice is to discard a number of the early observations in the chain, say $X_1, \ldots, X_B$ for some $B < m$, and instead estimate $\mu_h$ using only the last $m - B$ observations. This practice is known as burn-in and the idea is that the burn-in should bring the chain closer to its stationary distribution, approximating a situation where the initial value is drawn from the invariant distribution. This intuitive idea is not so easily motivated formally, however, and many authorities consider burn-in questionable [25].

Having selected where to start, one also needs to decide when to stop. In general, a longer chain is better and it is hence uncontroversial to say that $m$ should be as large as

possible. Even so, it is in many settings desirable to have an idea about when $m$ is large enough, in some quantifiable sense. There are ways to measure the distance between the distribution of the chain and the target distribution that can in some cases be used to determine an appropriate $m$ before starting the simulation. This subject is rather technical and we refer the reader to Jones and Hobert [42] for an introduction. One may also construct stopping rules that use, for example, an estimate of $\text{var}(\hat{\mu}_h)$ to decide when to terminate a simulation in real time [41, 74]. For example, one may calculate the width of the confidence interval for $\mu_h$, i.e. $t_{d-1,1-\alpha/2}\sqrt{\hat{\kappa}_h^2/m}$, and terminate when it falls below some pre-specified threshold. We illustrate this idea in the context of the running example.

**Example 1.3.1** (continued)**.** We have considered $\sum_{i=1}^{m} \mu_i/m$ as an estimator for the posterior mean $\mathbb{E}(\mu \mid y)$. In Figures 1.1 and 1.2 we used $m = 500$. When discussing variance estimation we instead used $m = 50000$. To see how to employ a stopping rule to decide on an appropriate $m$ based on the batch means estimate of $\text{var}(\hat{\mu}_h)$, suppose we are content with a width of 0.05 for a 95 % confidence interval for $\mathbb{E}(\mu \mid y)$. To avoid stopping the simulation too early due to poor estimates of $\text{var}(\hat{\mu}_h)$ for small $m$, let us implement a stopping rule as follows: for every $m = 10000, 11000, 12000, \ldots$ calculate a 95% confidence interval for $\mathbb{E}(\mu \mid y)$; if its width is less than 0.05, simulate for another 1000 iterations and try again, and otherwise stop the simulation. Implementing this in our example, we find that the MH algorithm stops after $m = 142000$ iterations while the Gibbs sampler stops after $m = 134000$ iterations.

# Chapter 2

# Consistent maximum likelihood estimation using subsets

## 2.1 Introduction

Mixed models are frequently used in applications and have been the subject of numerous articles and books [15, 35, 53]. Yet, it was unknown until recently whether MLEs are consistent even in some simple generalized linear mixed models (GLMMs) [37]. What complicates proving consistency in some mixed models is the dependence among response variables induced by certain random effects designs. Of course, not all types of dependence between responses are problematic – there is a vast literature on maximum likelihood estimation with dependent observations [5, 14, 30, 32, 68, 76, 79]. But, as we will discuss in more detail below, for some commonly used random effects designs such as those with crossed random effects, existing conditions for consistency of MLEs are hard to verify [37]. In a few GLMMs with crossed random effects, consistency has been proved using a novel argument that relates the likelihood for the full data to that of a subset consisting of independent and identically distributed (i.i.d.) random variables, "the subset argument" [36].

Fundamentally, however, the issue is not unique to GLMMs or even mixed models; any other parametric model appropriate for the same settings may present similar difficulties. Accordingly, it was recognized in the first work on consistency using subsets that the idea has the potential to be extended to more general models [36]. We address this by deriving weaker conditions, based in part on the use of subsets, that are sufficient for consistency of

MLEs, without assuming a particular model. They help explain formally what makes the subset argument work, why it is useful in some settings where more classical ones are not, and when it can fail. We illustrate the usefulness of our conditions by proving consistency of MLEs in two multivariate GLMMs (MGLMMs) to which existing theory has not been applied successfully.

To fix ideas, let $\Theta$ denote a parameter set, $f_\theta^n$ a joint density for the random vector $Y = (Y_1, \ldots, Y_n)$, and $\theta^0$ the "true" parameter. Let also $L_n(\theta; Y) = f_\theta^n(Y)/f_{\theta^0}^n(Y)$ and $\Lambda_n(\theta; Y) = \log L_n(\theta; Y)$. If $\Theta$ is a finite set, then since $L_n(\theta^0; Y) = 1$, a necessary and sufficient condition for consistency of MLEs is that, as $n \to \infty$,

$$\mathsf{P}(L_n(\theta; Y) \geq 1) \to 0 \quad \text{for all } \theta \neq \theta^0. \tag{2.1}$$

When $\Theta$ is not a finite set, (2.1) needs to be amended by a uniformity argument to be sufficient, but the main ideas are the same. There are many ways to establish (2.1). With i.i.d. observations and regularity conditions, (2.1) or stronger results follow from the law of large numbers applied to $n^{-1}\Lambda_n(\theta; Y)$ [12, 16, 18, 77]. More generally, if $Y$ is a stochastic process, $\Lambda_n(\theta; Y)$ may, suitably scaled, satisfy an ergodic theorem, leading again to (2.1) under regularity conditions. In the literature on maximum likelihood estimation with dependent observations, it is often assumed that some such limit law holds, either for $\Lambda_n(\theta; Y)$ or its derivatives [14, 30, 32], or that the moments of $\Lambda_n(\theta; Y)$ converge in an appropriate way [5, 68]. Unfortunately, in many practically relevant settings, it is not clear that any such convergence holds and proving that it does is arguably the main obstacle to establishing consistency of MLEs. Let us illustrate using an MGLMM, commonly considered both in statistics and applied sciences [10, 11, 28, 52, 78].

Let $X = [x_1, \ldots, x_n]^\mathsf{T} \in \mathbb{R}^{n \times p}$ be a matrix of non-stochastic predictors, $Z = [z_1, \ldots, z_n]^\mathsf{T} \in \mathbb{R}^{n \times r}$ a non-stochastic design matrix, and $U \in \mathbb{R}^r$ a multivariate normal vector of random effects, with mean zero and covariance matrix $\Sigma$. For the MGLMM, $\Theta \subseteq \mathbb{R}^d$, for some $d \geq 1$, $\beta = \beta(\theta)$, and $\Sigma = \Sigma(\theta)$. The responses $Y_1, \ldots, Y_n$ are conditionally independent

given $U$, with conditional densities in the form

$$f_{\theta,i}(y_i \mid u) = k_i(y_i, \tau_i) \exp\left(\frac{y_i[x_i^\mathsf{T}\beta + z_i^\mathsf{T}u] - c_i(x_i^\mathsf{T}\beta + z_i^\mathsf{T}u)}{\tau_i}\right),$$

where, for $i = 1, \ldots, n$, $c_i$ is the conditional cumulant function, $\tau_i$ a dispersion parameter, and $k_i(y_i, \tau_i)$ ensures $f_{\theta,i}(y_i \mid u)$ integrates to one. Conditional independence implies $f_\theta^n(y \mid u) = \prod_{i=1}^n f_{\theta,i}(y_i \mid u)$. Several of the responses could be from the same subject, hence the "multivariate", and they can be of mixed type, some continuous and some discrete, for example.

The dependence among the linear predictors is easily characterized since $X\beta + ZU \sim \mathcal{N}(X\beta, Z\Sigma Z^\mathsf{T})$. The relevant density for maximum likelihood estimation, however, is the marginal density,

$$f_\theta^n(y) = \int_{\mathbb{R}^r} f_\theta^n(y \mid u)\phi_\theta^r(u)\mathrm{d}u,$$

where $\phi_\theta^r$ denotes the $r$-dimensional multivariate normal density with mean zero and covariance matrix $\Sigma = \Sigma(\theta)$. The density $f_\theta^n(y)$ typically does not admit a closed form expression. Moreover, the dependence among responses it implies is in general less transparent than that among the linear predictors. What we can say in general is that two responses are dependent only if their corresponding linear predictors are. That is, response component $i$ and $j$ are independent if $z_i^\mathsf{T}\Sigma z_j = 0$.

It is convenient if $Z\Sigma Z^\mathsf{T}$ is, upon possible reordering of the responses, block diagonal since in that case the full vector of responses can be partitioned into independent subvectors. If these are of fixed length as $n$ grows then one is back in the classical setting where the full data consists only of an increasing number of independent vectors. This setting is common to many articles on asymptotic theory in mixed models [29, 56, 57, 70]. Unfortunately, in applications the number of independent response vectors – the number of diagonal blocks in $Z\Sigma Z^\mathsf{T}$ – is often small. For example, Sung and Geyer [70] note that in the famous salamander data [51] there are 3 independent vectors, each of length 120. Thus, in their notation there are $n = 3$ independent observations, but in our notation there

are $n = 3 \times 120 = 360$ possibly dependent observations. It seems more reasonable, then, to assume that $n \to \infty$ without also assuming that the response vector $Y$ consists only of an increasing number of independent sub-vectors. This type of limiting process has, in the context of mixed models, previously only been investigated carefully in special cases that do not allow for predictors or mixed-type responses [36, 55]. To be sure, Jiang's [36] general theory does allow for predictors, but the specific applications do not.

The intuition behind the usefulness of the subset argument can be understood by considering the following simple LMM with crossed random effects. Suppose $Y_{i,j} = \theta + U_i^{(1)} + U_j^{(2)} + E_{i,j}$, where $U_i^{(1)}$, $U_j^{(2)}$, and $E_{i,j}$ are all i.i.d. standard normal, $i = 1, \ldots, N$, $j = 1, \ldots, N$. It is easy to check that the $Y_{i,j}$s cannot be partitioned into independent subsets. However, there are many subsets that, even though there is dependence among them, consist of independent random variables. For example, the two subsets $(Y_{1,1}, Y_{2,2}, \ldots, Y_{N,N})$ and $(Y_{1,2}, Y_{2,3}, \ldots, Y_{N-1,N})$ are dependent, but taken separately they both consist of i.i.d. random variables. The MLE of $\theta$ based on either subset, i.e. a subset sample mean, is consistent as $N \to \infty$. Intuitively, then, the MLE based on all of the $N^2$ variables should be too. Of course, the subset argument is not needed to prove that in this simple example, but the intuition is the same for models where a direct proof is harder. How to formalize this intuition, without actually having to require the subset components to be either independent or identically distributed, is the topic of Section 2.2.

The rest of the paper is organized as follows. We develop our theory for consistency of MLEs using subsets in Section 2.2. These results do not assume any particular model. In Section 2.3 we apply the theory from Section 2.2 to two MGLMMs. Section 2.4 contains a brief discussion of our results and their implications. Many technical details are deferred to the appendices.

## 2.2 Consistency using subsets of the full data

Recall that $Y = (Y_1, \ldots, Y_n)$ denotes a collection of random variables and let

$$W = (W_1, \ldots, W_m)$$

denote a collection of random variables that form a subset of those in $Y$, i.e. $\{W_1, \ldots, W_m\} \subseteq \{Y_1, \ldots, Y_n\}$. We will henceforth call $W$ a subcollection of $Y$ to avoid confusion with other subsets introduced later. The main results in this section give conditions for when subcollections can be used to prove consistency of maximizers of $L_n(\theta; Y)$. Unless otherwise noted, all convergence statements are as $n$ tends to infinity and the number of elements in a subcollection, $m = m(n)$, tends to infinity as a function of $n$.

All discussed random variables are defined on an underlying probability space $(\Omega, \mathcal{F}, \mathsf{P})$, with the elements of $\Omega$ denoted $\omega$. The parameter set $\Theta$ is assumed to be a subset of a metric space $(\mathcal{T}, d_{\mathcal{T}})$. We write, for any $t \in \mathcal{T}$ and $\delta > 0$, $B_\delta(t) = \{t' \in \mathcal{T} : d_{\mathcal{T}}(t, t') < \delta\}$. For any $A \subseteq \mathcal{T}$, $\bar{A}$ denotes its closure and $\partial A$ its boundary. We assume the true parameter $\theta^0$ is the same for all $n$ but the joint density $f_\theta^n(y)$ of $Y$, against a dominating, $\sigma$-finite product measure $\nu = \nu_n$, can depend on $n$ in an arbitrary manner. In particular, our setting allows for a triangular array of responses, $Y_{n,1}, \ldots, Y_{n,n}$, though for convenience we do not make this explicit in the notation.

By $\theta^0$ being the true parameter we mean that $\mathsf{P}(Y \in A) = \int_A f_{\theta^0}^n(y)\nu(\mathrm{d}y)$ for any measurable $A$ in the range space of $Y$. That is, expectations and probabilities with respect to $\mathsf{P}$ are the same as those taken with respect to distributions indexed by $\theta^0$. Densities for the subcollection and its components are denoted by $g$ in place of $f$; for example, $L_m(\theta; W) = g_\theta^m(W)/g_{\theta^0}^m(W)$.

We will use subcollections to establish the following sufficient condition for consistency of maximizers of $L_n(\theta; Y)$:

$$\mathsf{P}\left(\sup_{\theta \in \Theta \cap B_\varepsilon(\theta^0)^c} L_n(\theta; Y) \geq 1\right) \to 0, \quad \forall\, \varepsilon > 0\,. \tag{2.2}$$

The appeal of using subcollections to prove (2.2), instead of directly working with the full data likelihood $L_n(\theta; Y)$, can be explained using the following lemma.

**Lemma 2.2.1.** *For every $c \in (0, \infty)$, $\theta \in \Theta$, and subcollection $W$, $\mathsf{P}$-almost surely,*

$$\mathsf{P}\left(L_n(\theta; Y) \geq c \mid W\right) \leq c^{-1} L_m(\theta; W).$$

Versions of Lemma 2.2.1 are well known [36, 37], but Appendix A contains a proof for completeness. From the lemma it follows that if $L_m(\theta; W) \to 0$, then $\mathbb{E}[\mathsf{P}(L_n(\theta; Y) \geq 1 \mid W)] = \mathsf{P}(L_n(\theta; Y) \geq 1) \to 0$ by dominated convergence. That is, up to a uniformity argument, (2.2) can be established by showing that the likelihood of the subcollection converges to zero in probability, outside of a neighborhood of $\theta^0$. Uniform versions of that convergence will play a crucial role in our results.

**Definition 2.2.1.** We say that a subset $A \subseteq \Theta$ is identified by a subcollection $W$ if $\sup_{\theta \in A} L_m(\theta; W) \xrightarrow{\mathsf{P}} 0$. If $\sup_{\theta \in A} L_m(\theta; W) = O_{\mathsf{P}}(a_n)$ for some sequence of constants $\{a_n\}$, $n = 1, 2, \ldots$, we call $a_n$ an *identification rate*.

To understand this definition better, consider the case where the subcollection $W$ consists of $m$ i.i.d. random variables with common marginal density $g_{\theta,1}$. Suppose also that there is no $\theta \in A$ for which $g_{\theta,1} = g_{\theta^0,1}$ $\nu$-almost everywhere. That is, $\theta^0$ is an identified parameter in the classical sense if we restrict attention to the parameter set $A \cup \{\theta^0\}$. Then, under regularity conditions [18, Theorems 16 and 17], one has $\sup_{\theta \in A} \mathbb{E}[\Lambda_m(\theta; W)] < 0$ and, by a uniform strong law of large numbers,

$$\lim_{m \to \infty} m^{-1} \sup_{\theta \in A} |\Lambda_m(\theta; W) - \mathbb{E}[\Lambda_m(\theta; W)]| = 0.$$

Using this, it is straightforward to show that $A$ is identified by $W$ with an identification rate that is exponentially fast in $m$. That is, with i.i.d. components and regularity conditions, the classical definition of an identified parameter implies identification in the sense of Definition 2.2.1. However, we want to allow for subcollections that do not consist of i.i.d. components, and in that case the classical definition is not as useful. For example, we have independent

but not identically distributed components in one of our MGLMMs. In this and more general cases, a parameter could be identified in the classical sense for all sample sizes $n$, but, loosely speaking, the difference between the distributions for $W$ indexed by some $\theta \in A$ and that indexed by $\theta^0$ could vanish asymptotically, preventing $W$ from identifying $A$ in our sense. Finally, notice also that $A$ being identified by $W$ is essentially equivalent to MLEs based on $W$ with the restricted parameter set $A \cup \{\theta^0\}$ being consistent.

We can now be more precise about how to use subcollections to establish (2.2). The strategy is to first find a subcollection $W$ that identifies $B_\varepsilon(\theta^0)^c \cap \Theta$ for every $\varepsilon > 0$, and then use Lemma 2.2.1 to get the convergence for the full likelihood in (2.2). For this strategy to be useful, showing that $W$ identifies $B_\varepsilon(\theta^0)^c \cap \Theta$ has to be easier than showing that $Y$ does since the latter would directly imply (2.2). That is, one has to be able to pick out a subcollection with more convenient properties than the full data. Our applications in Section 2.3 illustrate how this can be done.

It is useful to allow for several subcollections $W^{(i)}$, consisting of $m_i$ components, and subsets $A_i$, $i = 1, \ldots, s$. By doing so, different subcollections can be used to identify different subsets of the parameter set. For example, if the parameter set is a product space, as is common in applications, then different subcollections can be used to, loosely speaking, identify different elements of the parameter vector. Assumption 1 makes precise what we need to identify $\Theta \cap B_\varepsilon(\theta^0)^c$ using several subcollections.

**Assumption 1.** For every small enough $\varepsilon > 0$, there are subsets $A_i = A_i(\varepsilon) \subseteq \Theta$ and corresponding subcollections $W^{(i)}$, $i = 1, \ldots, s$, such that $\cup_{i=1}^s A_i \supseteq \Theta \cap B_\varepsilon(\theta^0)^c$ and each $A_i$ is identified by $W^{(i)}$ with some identification rate $a_{n,i}$, $n = 1, 2, \ldots$, $i = 1, \ldots, s$.

This assumption is somewhat similar to assumptions A2 and A3 made by Jiang [36], which are also assumptions about parameter identification using several subcollections. However, those assumptions are stated in terms of $\mathbb{E}(\Lambda_{m_i}(\theta; W^{(i)}))$ and $\mathrm{var}(\Lambda_{m_i}(\theta; W^{(i)}))$, $i = 1, \ldots, s$. The fact that we do not have to assume anything about the variances of the log-likelihood ratios is an important improvement. For example, if subcollection $i$ consists of i.i.d. components, the convergence of $m_i^{-1}\Lambda_{m_i}(\theta; W^{(i)})$ is immediate from the law of large

numbers, but calculating its variance may be difficult.

For finite parameter sets, Assumption (1) is enough to give consistency of MLEs via Lemma 2.2.1. For more general cases we also need to control the regularity of the log-likelihood for the full data. The following two assumptions are made to ensure that the uniformity of the convergence detailed in Assumption 1 and Definition 2.2.1 carries over to $\Lambda_n(\theta; Y)$, in the sense of (2.2).

**Assumption 2.** For every $i \in \{1, \dots, s\}$ and $n \in \{1, 2, \dots\}$, $\Lambda_n(\theta; Y)$ is $\mathsf{P}$-almost surely Lipschitz continuous in $\theta$ on the $A_i$ defined in Assumption 1; that is, there exists a random variable $K_{n,i}$ not depending on $\theta$ such that, $\mathsf{P}$-almost surely and for every $\theta, \theta' \in A_i$,

$$|\Lambda_n(\theta; Y) - \Lambda_n(\theta'; Y)| \leq K_{n,i} d_{\mathcal{T}}(\theta, \theta').$$

**Assumption 3.** Each $A_i$ from Assumption 1 can be covered by $M_{n,i}$ balls of radius $\delta_{n,i}$ such that

$$K_{n,i}\delta_{n,i} \xrightarrow{\mathsf{P}} 0 \text{ and } M_{n,i}a_{n,i} \to 0,$$

where $a_{n,i}$ and $K_{n,i}$, $i = 1, \dots, s$, $n = 1, 2, \dots$, are the same as in Assumptions 1 and 2, respectively.

The assumptions give us the convergence in (2.2) and, consequently, the following lemma.

**Lemma 2.2.2.** *If Assumptions 1 – 3 hold, then the probability that there exists a global maximizer of $\Lambda_n(\theta; Y)$ in $B_\varepsilon(\theta^0)^c \cap \Theta$ tends to zero as $n \to \infty$, for every $\varepsilon > 0$.*

*Proof.* We give an outline here and a detailed proof in Appendix A.1. Without loss of generality, we may assume $s = 1$, so there is one subcollection $W$ that identifies $A = \Theta \cap B_\varepsilon(\theta^0)^c$, for arbitrary, small $\varepsilon > 0$, with rate $a_n$. It suffices to prove that $\mathsf{P}(\sup_{\theta \in A} L_n(\theta; Y) \geq 1) \to 0$. For $j = 1, \dots, M_n$ let $\theta^j$ be a point in the intersection of $A$ and the $j$th ball in the cover of $A$ given by Assumption 3. Some algebra and Assumption 2 gives

$$\mathsf{P}\left(\sup_{\theta \in A} L_n(\theta; Y) \geq 1\right) \leq \mathsf{P}\left(\max_{j \leq M_n} L_n(\theta^j; Y) \geq 1/2\right) + \mathsf{P}\left(e^{K_n \delta_n} \geq 2\right).$$

The second term is $o(1)$ by Assumption 3. It remains to deal with the first. By conditioning on the subcollection and using Lemma 2.2.1 one gets

$$\mathsf{P}\left(\max_{j\leq M_n} L_n(\theta^j;Y)\geq 1/2 \mid W\right)\leq 2M_n \sup_{\theta\in A} L_m(\theta;W).$$

The right hand side is $o(1)$ by Assumption 3, so the expectation of the left hand side is also $o(1)$ by dominated convergence, which finishes the proof. □

We will use Lemma 2.2.2 to establish both a Wald-type consistency, meaning consistency of sequences of global maximizers of $L_n(\theta;Y)$, and a Cramér-type consistency, meaning consistency of a sequence of roots to the likelihood equations $\nabla\Lambda_n(\theta;Y)=0$. It follows almost immediately from the lemma that if $L_n(\theta;Y)$ has a global maximizer $\hat{\theta}_n$, $\mathsf{P}$-almost surely for every $n$, then $\hat{\theta}_n \xrightarrow{\mathsf{P}} \theta^0$. In particular, if $\Theta$ is compact one gets Wald-type consistency with an additional continuity assumption. Since Assumption 2 implies $L_n(\theta;Y)$ is continuous at every point except possibly $\theta^0$, assuming continuity also at the unknown $\theta^0$ should be insignificant in any application of interest.

**Theorem 2.2.3.** *If $\Theta$ is compact, $L_n(\theta;Y)$ is $\mathsf{P}$-almost surely continuous on $\Theta$ for every $n$, and Assumptions 1 – 3 hold, then a maximizer $\hat{\theta}_n$ of $L_n(\theta;Y)$ exists $\mathsf{P}$-almost surely for every $n$, and $\hat{\theta}_n \xrightarrow{\mathsf{P}} \theta^0$ for any sequence of such maximizers.*

*Proof.* Since continuous functions attain their suprema on compact sets, $L_n(\theta;Y)$ has a maximizer on $\Theta$, $\mathsf{P}$-almost surely. By Lemma 2.2.2 all maximizers are in $B_\varepsilon(\theta^0)$ with probability tending to one, for all small enough $\varepsilon>0$. □

Though compactness is a common assumption [32, 80], it is sometimes too restrictive or even unnecessary. If $L_n(\theta;Y)$, or more commonly $\Lambda_n(\theta;Y)$, is strictly concave in $\theta$ on a convex $\Theta$, then it is enough to verify the assumptions on a neighborhood of $\theta^0$ (c.f. Theorem 2.2.4) to get consistency of the unique global maximizer. However, a global maximizer need not exist even as $n\to\infty$, or perhaps the assumptions cannot be verified for other reasons. With a few additional assumptions, Lemma 2.2.2 can then be used to get the weaker Cramér-type consistency, which also only requires verifying assumptions for neighborhoods of $\theta^0$.

**Theorem 2.2.4.** *If $\Theta \subseteq \mathbb{R}^d$ for some $d \geq 1$, $L_n(\theta; Y)$ is almost surely differentiable in $\theta$ on a neighborhood of an interior $\theta^0$ for every $n$, and Assumptions 1 – 3 hold with $\Theta$ replaced by $\bar{B}_\varepsilon(\theta^0)$ for all small enough $\varepsilon > 0$, then, with probability tending to one as $n \to \infty$, there exists a local maximizer of $L_n(\theta; Y)$, and hence a root to the likelihood equation $\nabla \Lambda_n(\theta; Y) = 0$, in $B_\varepsilon(\theta^0)$, for all small enough $\varepsilon > 0$.*

*Proof.* Since $\theta^0$ is interior we may assume $\varepsilon > 0$ is small enough that all points of $\bar{B}_\varepsilon(\theta^0)$ are interior. Almost sure differentiability of $L_n(\theta; Y)$ implies almost sure continuity. Thus, $L_n(\theta; Y)$ attains a local maximum on the compact $\bar{B}_\varepsilon(\theta^0)$, $\mathsf{P}$-almost surely. By Lemma 2.2.2, with probability tending to one, there are no such maximizers in $\bar{B}_\varepsilon(\theta^0) \setminus B_\varepsilon(\theta^0) = \partial B_\varepsilon(\theta^0)$. Thus, with probability tending to one, there exists a local maximizer in $B_\varepsilon(\theta^0)$. Since $L_n(\theta; Y)$ and hence $\Lambda_n(\theta; Y)$ is $\mathsf{P}$-almost surely differentiable, any such maximizer must be a root to the likelihood equation $\nabla \Lambda_n(\theta; Y) = 0$. □

In the next section we apply Theorem 2.2.4 to two special cases of the MGLMM described in Section 2.1. We also discuss in more detail how to think about the subcollections and subsets in specific models.

## 2.3 Application to multivariate mixed models

### 2.3.1 Longitudinal linear mixed model

The first model we consider is an extension of the variance components model that has been studied previously [55]. In addition to dependence between subjects induced by crossed random effects the model incorporates autoregressive temporal dependence between measurements from the same subject. To make the discussion clearer we assume easy-to-specify fixed and random effect structures. This allows us to focus on the core issues, that is, on how to select subcollections and subsets that can be used to verify the conditions of our theory. Our model includes a baseline mean and a treatment effect. A general fixed effect design matrix could be treated the same way as in our second example, discussed in Section 2.3.2. Before establishing consistency, we discuss the model definition and how to select

appropriate subcollections.

Suppose for subjects $(i, j)$, $i = 1, \ldots, N$, $j = 1, \ldots, N$, and time points $t = 1, \ldots, T$ we observe the response $Y_{i,j,t}$, where for convenience we assume both $N$ and $T$ are even. Let the stacked vector of responses be

$$Y = [Y_{1,1,1}, \ldots, Y_{1,1,T}, Y_{1,2,1}, \ldots, Y_{N,N,T}]^\mathsf{T} \in \mathbb{R}^n, \ n = TN^2.$$

Recall from the introduction that the MGLMM is specified by the conditional distribution $f_\theta^n(y \mid u)$ and the distribution of the random effects, $\phi_\theta^r(u)$. For a linear mixed model we let $f_\theta^n(y \mid u)$ be a multivariate normal distribution with mean $X\beta + Zu$ and covariance matrix $\theta_3 I_n$, $\theta_3 > 0$, where the two components of $\beta = [\theta_1, \theta_2]^\mathsf{T} \in \mathbb{R}^2$ are a baseline mean and a treatment effect, respectively, and $I_k$ denotes the $k \times k$ identity matrix. Note, in the notation of the introduction, the dispersion parameter in the conditional distribution is $\tau_i = \theta_3$, for all $i$. We treat $\theta_3$ as a parameter to be estimated and not as known, which is otherwise common in the literature.

Let $h_n$ be a vector of zeros and ones where the $i$th element is one if it corresponds to an observation in time $t \leq T/2$ and zero otherwise and let $1_n$ denote an $n$-vector of ones. We take $X = [1_n, h_n] \in \mathbb{R}^{TN^2 \times 2}$, which corresponds to a treatment being applied in the first half of the experiment. Notice that unless $T$ is fixed, which we do not assume, this setup implies the predictors change with $n$. Indeed, as $T$ grows, a particular observation can go from being made in the latter half of the experiment to the earlier half. Thus, the responses form a triangular array.

Partition $U$ into three independent sub-vectors, $U^{(1)} \sim \mathcal{N}(0, \theta_4 I_N)$, $U^{(2)} \sim \mathcal{N}(0, \theta_5 I_N)$, and $U^{(3)} \sim \mathcal{N}(0, \theta_6 I_{N^2} \otimes \Psi)$, where $\Psi = (\Psi_{i,j}) = (\theta_7^{|i-j|})$ is a first order autoregressive correlation matrix, $\theta_i > 0$, $i = 4, 5, 6$, and $\theta_7 \in (-1, 1)$. We will use $U^{(1)}$ and $U^{(2)}$ as crossed random effects, inducing dependence between subjects, and $U^{(3)}$ to get temporal dependence within subjects. To that end, let $Z_1 = I_N \otimes 1_N \otimes 1_T$, $Z_2 = 1_N \otimes I_N \otimes 1_T$, and $Z = [Z_1, Z_2, I_{TN^2}]$. Then, with $J_k = 1_k 1_k^\mathsf{T}$, the covariance matrix of the linear predictors

$X\beta + ZU$ is

$$Z\Sigma Z^{\mathsf{T}} = \theta_4 I_N \otimes J_{NT} + \theta_5 J_N \otimes I_N \otimes J_T + \theta_6 I_{N^2} \otimes \Psi.$$

More transparently, for the elements of $\mathbb{E}(Y \mid U) = X\beta + ZU$, it holds that

$$\operatorname{cov}[\mathbb{E}(Y_{i,j,t} \mid U), \mathbb{E}(Y_{i',j',t'} \mid U)] = \begin{cases} \theta_4 + \theta_5 + \theta_6 \theta_7^{|t-t'|} & i = i', j = j' \\ \theta_4 & i = i', j \neq j' \\ \theta_5 & i \neq i', j = j' \\ 0 & \text{otherwise} \end{cases}.$$

The marginal density $f_\theta^n(y)$ admits a closed form expression in this example. Specifically, the marginal distribution for $Y$ is multivariate normal with mean $m(\theta) = X\beta(\theta)$ and covariance matrix $C(\theta) = \theta_3 I_{TN^2} + Z\Sigma(\theta)Z^{\mathsf{T}}$. Note that the structure of $C(\theta)$ is similar to that of the covariance matrix of the linear predictors just discussed. In particular, there are many zeros in the covariance matrix $C(\theta)$, i.e. there are many independent observations, but $Y$ cannot be partitioned into independent vectors.

**Subcollection selection**

The model definitions imply that $\Theta = \mathbb{R} \times \mathbb{R} \times (0,\infty) \times (0,\infty) \times (0,\infty) \times (0,\infty) \times (-1,1)$, a subset of $(\mathcal{T}, d_{\mathcal{T}}) = (\mathbb{R}^7, \|\cdot\|)$, where $\|\cdot\|$ denotes the Euclidean norm when applied to vectors. We write $\theta = (\theta_1, \ldots, \theta_7)$.

Subcollections are selected for the purpose of verifying Assumption 1. The main idea guiding selection is suggested by the fact that identification follows, under regularity conditions, if the subcollection's log-likelihood satisfies a law of large numbers. We will use $s = 2$ such subcollections and require that they together identify $\theta$ in the classical sense. By this we mean that, letting $\nu_\theta^i$ denote the distribution of subcollection $i$ implied by parameter $\theta$,

$$\{\theta \in \Theta : \nu_\theta^1 = \nu_{\theta^0}^1\} \cap \{\theta \in \Theta : \nu_\theta^2 = \nu_{\theta^0}^2\} = \{\theta^0\}.$$

With these properties in mind, we take $W^{(1)}$ to consist of the vectors

$$W_i^{(1)} = (Y_{2i-1,2i-1,1}, Y_{2i,2i,T}) \in \mathbb{R}^2, i = 1, \ldots, N/2.$$

Because these vectors do not share any random effects, they are independent. In fact, they are i.i.d. multivariate normal with common mean $m_1(\theta) = [\theta_1 + \theta_2, \theta_1]^\mathsf{T}$ and common covariance matrix $C_1(\theta) = I_2(\theta_3+\theta_4+\theta_5+\theta_6)$. Clearly, $\theta_1$ and $\theta_2$ are identified in the classical sense by this subcollection, but not $\theta_3, \ldots, \theta_7$. Note that even though the predictors, and hence the distributions, do not change with $N$ for this subcollection, it is strictly speaking a triangular array unless $T$ is fixed.

To identify the remaining parameters, take $W^{(2)}$ to consist of the vectors

$$W_i^{(2)} = (Y_{2i-1,2i-1,1}, Y_{2i-1,2i-1,2}, Y_{2i-1,2i-1,3}, Y_{2i-1,2i,1}, Y_{2i,2i-1,1}), \quad i = 1, \ldots, N/2.$$

These are also i.i.d. multivariate normal, with common mean $m_2(\theta) = (\theta_1 + \theta_2)1_5$ and common covariance matrix

$$C_2(\theta) = \begin{bmatrix} \sum_{i=3}^6 \theta_i & \theta_4 + \theta_5 + \theta_6\theta_7 & \theta_4 + \theta_5 + \theta_6\theta_7^2 & \theta_4 & \theta_5 \\ \cdot & \sum_{i=3}^6 \theta_i & \theta_4 + \theta_5 + \theta_6\theta_7 & \theta_4 & \theta_5 \\ \cdot & \cdot & \sum_{i=3}^6 \theta_i & \theta_4 & \theta_5 \\ \cdot & \cdot & \cdot & \sum_{i=3}^6 \theta_i & 0 \\ \cdot & \cdot & \cdot & \cdot & \sum_{i=3}^6 \theta_i \end{bmatrix}.$$

It is straightforward to check that $C_2(\theta) = C_2(\theta')$ implies $\theta_i = \theta_i'$, $i = 3, \ldots, 7$.

In summary, the two subcollections together identify $\theta$ in the classical sense. Moreover, since both subcollections consist of i.i.d. multivariate normal vectors, their log-likelihoods satisfy a law of large numbers as $N \to \infty$. With this we are equipped to verify that Assumptions 1 – 3 hold locally, leading to the main result of the section in Theorem 2.3.4.

### Consistency

The purpose of this section is to verify the conditions of Theorem 2.2.4. The interesting part of that is to check that Assumptions 1 – 3 hold with $\Theta$ replaced by $\bar{B}_\varepsilon(\theta^0)$, for all small enough $\varepsilon > 0$. For this purpose we will first prove two lemmas that roughly correspond to Assumptions 1 and 2. The limiting process we consider is that $N$ tends to infinity while $T$ can be fixed or tend to infinity with $N$, at rates discussed below. Thus, the statements $n \to \infty$ and $N \to \infty$ are equivalent. We will need the following result which is proved in Appendix A.2.

**Proposition 2.3.1.** *If $\Theta$ is compact, $L_{m_i}(\theta; W^{(i)})$ is continuous in $\theta$ on $\Theta$ for every $w^{(i)}$ in the support of $W^{(i)}$, $i = 1, \ldots, s$, and $\cap_{i=1}^s \{\theta \in \Theta : \nu_\theta^i = \nu_{\theta^0}^i\} = \{\theta^0\}$, then for any $\varepsilon > 0$ there are compact sets $\tilde{A}_1, \ldots, \tilde{A}_s$ such that $\{\theta \in \Theta : \nu_\theta^i = \nu_{\theta^0}^i\} \cap \tilde{A}_i = \emptyset$, $i = 1, \ldots, s$, and $\cup_{i=1}^s \tilde{A}_i = \Theta \cap B_\varepsilon(\theta^0)^c$.*

Note, when applying the proposition in the present application, $m_i = N$, $s = 2$, and $\Theta$ is replaced by $\bar{B}_\varepsilon(\theta^0)$. The proposition is useful because the $\tilde{A}_i$s it gives are compact, as we will see in the proof of the following lemma. This lemma formalizes verification of Assumption 1.

**Lemma 2.3.2.** *If $\theta^0$ is an interior point of $\Theta$, then for all small enough $\varepsilon > 0$ there exist subsets $A_1$ and $A_2$ such that $A_1 \cup A_2 = \partial B_\varepsilon(\theta^0)$,*

1. *$N^{-1} \sup_{\theta \in A_i} \mathbb{E}[\Lambda_{N/2}(\theta; W^{(i)})] = \sup_{\theta \in A_i} \mathbb{E}[\Lambda_1(\theta; W_1^{(i)})]/2 < 0$,*
2. *$\mathsf{P}$-almost surely, $N^{-1} \sup_{\theta \in A_i} |\Lambda_{N/2}(\theta; W^{(i)}) - \mathbb{E}[\Lambda_{N/2}(\theta; W^{(i)})]| \to 0$, and, consequently;*
3. *$A_i$ is identified by $W^{(i)}$ with an identification rate $a_{n,i} = o(e^{-\epsilon N(n)})$ for some $\epsilon > 0$, $i = 1, 2$.*

*Proof.* We give an outline here and a detailed proof in Appendix A.2. It is easy to check that the requirements of Proposition 2.3.1 are satisfied with $\Theta$ replaced by $\bar{B}_\varepsilon(\theta^0)$. By taking the $A_i$s to be the $\tilde{A}_i$s given by Proposition 2.3.1, proving points 1 – 2 is similar to

proving that MLEs based on subcollection $i$ are consistent if the parameter set is restricted to the compact set $A_i \cup \{\theta^0\}$, $i = 1, 2$. Since the subcollection components are i.i.d., this is straightforward using classical ideas [18, Theorems 16 and 17]. The only difference from the referenced work is that one subcollection is a triangular array and so we use a different strong law. Point 3 follows from points 1 and 2. □

Note that, in this lemma and elsewhere, $\epsilon$ is a small number that is defined in context whereas $\varepsilon$ always denotes the radius of the neighborhood of $\theta^0$ we are considering. It remains to verify the assumptions concerned with the regularity of the log-likelihood of the full data. When the log-likelihood is differentiable, Lipschitz continuity follows from the mean value theorem if the gradient is bounded. The following lemma uses that to verify Assumption 2. The resulting Lipschitz constant, i.e. the bound of the gradient, is the same for both $A_1$ and $A_2$. The lemma also gives a probabilistic bound on the order of this Lipschitz constant as $n \to \infty$ that will be useful when verifying Assumption 3.

**Lemma 2.3.3.** *If $\theta^0$ is an interior point of $\Theta$, then for every $n$ and small enough $\varepsilon > 0$ there exists a random variable $K_n$ such that, $\mathsf{P}$-almost surely,*

$$\sup_{\theta \in \bar{B}_\varepsilon(\theta^0)} \|\nabla \Lambda_n(\theta; Y)\| \leq K_n = o_{\mathsf{P}}(n^b),$$

*for some $b > 0$.*

Proving Lemma 2.3.3 (see Appendix A.2) is largely an exercise in bounding the eigenvalues of the covariance matrix $C(\theta)$ and its inverse on interior points of $\Theta$. We are ready for the main result of the section.

**Theorem 2.3.4.** *If $\theta^0$ is an interior point of $\Theta$ and $T = O(N^k)$ for some $k \geq 0$ as $N \to \infty$, then, $\mathsf{P}$-almost surely, there exists a sequence $\hat{\theta}_n$ of roots to the likelihood equations $\nabla \Lambda_n(\theta; Y) = 0$ such that $\hat{\theta}_n \xrightarrow{\mathsf{P}} \theta^0$.*

*Proof.* We verify the conditions of Theorem 2.2.4. Fix an arbitrary $\varepsilon > 0$. Since $\theta^0$ is interior we may assume $\varepsilon$ is small enough that all points in $\bar{B}_\varepsilon(\theta^0)$ are interior points

of $\Theta$. By Lemma A.2.3 $\ell_n(\theta; Y)$ is $\mathsf{P}$-almost surely differentiable on $\bar{B}_\varepsilon(\theta^0)$, so $\Lambda_n(\theta; Y) = \ell_n(\theta, Y) - \ell_n(\theta^0; Y)$ is, too. By Lemma 2.3.2, Assumption 1 holds with what is there denoted $\Theta$ replaced by $\bar{B}_\varepsilon(\theta^0)$. The identification rate is exponentially fast in $N/2$, $a_n = o(e^{-N\epsilon})$ for some $\epsilon > 0$. Lemma 2.3.3 shows that $\Lambda_n(\theta; Y)$ is $K_n$-Lipschitz on both $A_1$ and $A_2$, and that $K_n = o_{\mathsf{P}}(n^b)$ for some $b > 0$. This verifies Assumption 2. It remains only to verify that the rate conditions in Assumption 3 hold. The $\delta$-covering number of the sphere $\partial B_\varepsilon(\theta^0)$ is $O([\varepsilon/\delta]^{d-1})$ as $\delta \to 0$ [7, Lemma 1]. Thus, since $A_i \subseteq \partial B_\varepsilon(\theta^0)$, by picking $\delta_{i,n} = n^{-b}$ we can have $M_{n,i} = O(n^{[d-1]b})$ as $n \to \infty$, $i = 1, 2$. Our choice of $\delta_{n,i}$ ensures $K_{n,i}\delta_{n,i} = K_n\delta_n = o_{\mathsf{P}}(1)$, which is the first rate condition. Since the identification rate is exponential in $N/2$ for both subcollections, we have that $M_{n,i}a_{n,i} = O(N^{2b[d-1]}T^{b[d-1]}e^{-\epsilon N})$ for some $\epsilon > 0$, which is $o(1)$ as $N \to \infty$ since $T$ is of (at most) polynomial order in $N$. □

Two key components in the proof of Theorem 2.3.4 are that the subcollections consist of i.i.d. random vectors that identify the parameters and that the gradient of the log-likelihood is of polynomial order in $n$. We expect the same proof technique to work in many other cases. Essentially, all that is needed is that the subcollections grow faster than logarithmically in the total sample size $n$ and that the gradient is of polynomial order, no matter how large that order is.

It is possible that the assumption that $T = O(N^k)$, $k \geq 0$, could be relaxed by picking other subcollections that also make use of the variation in the time dimension. It is not trivial, however, since the dependence between any two responses sharing a random effect does not vanish as time between the observations increases. In the next section we examine how predictors and mixed-type responses affect the argument.

### 2.3.2 Logit–normal MGLMM

The model we consider in this section is an extension in several ways of the logistic GLMMs for which the technique based on subcollections was first developed [36]. The random effect structures are similar, i.e. crossed, but we have multivariate, mixed-type responses, and predictors. The main ideas for verifying the assumptions of the theory from Section 2.2

are the same as in our LMM example. However, due to the inclusion of predictors, we use results from empirical process theory in place of the more classical strong laws used for the LMM. Showing existence of appropriate subsets of the parameter space that the subcollections identify also requires more work than with i.i.d. components. As before, we discuss the model definition and subcollection selection before establishing consistency.

Suppose for subjects $(i, j)$, $i = 1, \ldots, N, j = 1, \ldots, N$, there are two responses, $Y_{i,j,1}$ which is continuous and $Y_{i,j,2}$ which is binary. The vector of all responses is

$$Y = [Y_{1,1,1}, Y_{1,1,2}, Y_{1,2,1}, \ldots, Y_{N,N,2}]^\mathsf{T} \in \mathbb{R}^n, \ n = 2N^2.$$

For each subject we observe a vector of non-stochastic predictors $x_{i,j} \in \mathbb{R}^p$, the same for both responses. Similarly, $z_{i,j} \in \mathbb{R}^r$ is the same for both responses. Let $l_{i,j,k} = x_{i,j}^\mathsf{T}\beta_k + z_{i,j}^\mathsf{T}u$ be the linear predictor, $i = 1, \ldots, N$, $j = 1, \ldots, N$, $k = 1, 2$, where $\beta_1 = [\theta_1, \ldots, \theta_p]^\mathsf{T}$, $\beta_2 = [\theta_{p+1}, \ldots, \theta_{2p}]^\mathsf{T}$. We assume that $\|x_{i,j}\| \leq 1$ for all $i, j$. In practice this only rules out the possibility that $\|x_{i,j}\| = \infty$ since our setting allows for the standardization of predictors. The conditional density of the responses given the random effects that we consider is

$$f_\theta^n(y \mid u) = (2\pi)^{-n/2} \exp\left[\sum_{i,j} -(y_{i,j,1} - l_{i,j,1})^2/2 + y_{i,j,2}l_{i,j,2} - \log\left(1 + e^{l_{i,j,2}}\right)\right].$$

Given the random effects, $Y_{i,j,1}$ is normal with mean $l_{i,j,1}$ and variance 1, and $Y_{i,j,2}$ is Bernoulli with success probability $1/(1 + e^{-l_{i,j,2}})$ – a logistic GLMM. The choice of $\tau_i = 1$ for all $i$ is made for identifiability reasons for the Bernoulli responses, and for convenience for the normal responses. Setting the $\tau_i$s to some other known constants does not fundamentally change the results.

Consider two independent vectors, $U^{(1)} \sim \mathcal{N}(0, \theta_d I_N)$ and $U^{(2)} \sim \mathcal{N}(0, \theta_d I_N)$, and corresponding design matrices $Z_1 = I_N \otimes 1_N \otimes 1_2$ and $Z_2 = 1_N \otimes I_n \otimes 1_2$. With $U = [U^{(1)\mathsf{T}}, U^{(2)\mathsf{T}}]^\mathsf{T}$ and $Z = [Z_1, Z_2]$ the linear predictors are $l_{i,j,k} = x_{i,j}^\mathsf{T}\beta_k + u_i^{(1)} + u_j^{(2)}$. Thus, responses from the same subject share two random effects, responses from different subjects with one of the first two indexes in common share one random effect, and other responses

share no random effects and are hence independent. The covariance matrix for the linear predictors is easily computed in the same way as in the LMM. The covariance matrix for responses, however, is less transparent. It is for simplicity that we assume in this section that all random effects have the same variance. It is not necessary for our theory to be operational but this simplification shortens proofs considerably and allows us to focus on the main ideas.

**Subcollection selection**

With $p$ predictors the $(2p+1)$-dimensional parameter set is $\Theta = \mathbb{R}^p \times \mathbb{R}^p \times (0, \infty)$, a subset of the metric space $(\mathbb{R}^d, \|\cdot\|)$. The intuition behind the selection of subcollections is that the normal responses should identify the coefficient $\beta_1$ and the variance parameter $\theta_d$. Similarly, the Bernoulli responses should identify the coefficient vector $\beta_2$. With that in mind we take $W^{(i)} = (Y_{1,1,i}, Y_{2,2,i}, \ldots, Y_{N,N,i})$, $i = 1, 2$. Both of these subcollections consist of independent but not identically distributed random variables – independence follows from the fact that no components in the same subcollection share random effects. Notice that these subcollections are in practice often triangular arrays since the predictors may need to be scaled by $1/\max_{i \leq N, j \leq N} \|x_{i,j}\|$ to satisfy $\|x_{i,j}\| \leq 1$. All responses in the first subcollection have marginal normal distributions and all responses in the second have marginal Bernoulli distributions.

Identification is more complicated than in our previous example. One issue is that there can be many $\theta_d$ and $\beta_2$ that give the same marginal success probability for the components in the second subcollection. A second issue is that, since the predictors can change with $n$, classical identification for a fixed $n$ does not necessarily lead to identification in the sense of Definition 2.2.1. Additionally, the approach used in the LMM to find appropriate subsets $A_1$ and $A_2$ by means of Proposition 2.3.1 only works in general when the subcollection components are i.i.d. Thus, we take a slightly different route to establishing consistency compared to the LMM.

### Consistency

In this section we verify the conditions of Theorem 2.2.4. The limiting process is that $N \to \infty$, which is equivalent to $n \to \infty$ since $n = 2N^2$. We will first prove two lemmas that roughly correspond to Assumptions 1 and 2.

Let $\lambda_{\min}(\cdot)$ denote the minimum eigenvalue of its matrix argument.

**Lemma 2.3.5.** *If $\theta^0$ is an interior point of $\Theta$ and*

$$\liminf_{N\to\infty} \lambda_{\min}\left(N^{-1}\sum_{i=1}^{N} x_{i,i}x_{i,i}^{\mathsf{T}}\right) > 0,$$

*then for all small enough $\varepsilon > 0$ there exist $A_1$ and $A_2$ such that $A_1 \cup A_2 = \partial B_\varepsilon(\theta^0)$,*

1. $\limsup_{N\to\infty} N^{-1} \sup_{\theta\in A_i} \mathbb{E}[\Lambda_N(\theta; W^{(i)})] < 0$,
2. $\sup_{\theta\in A_i} N^{-1} \left|\Lambda_N(\theta; W^{(i)}) - \mathbb{E}[\Lambda_N(\theta; W^{(i)})]\right| \xrightarrow{\mathsf{P}} 0$, *and, consequently;*
3. $A_i$ *is identified by* $W^{(i)}$ *with an identification rate* $a_{n,i} = o(e^{-\epsilon N})$ *for some* $\epsilon > 0$, $i = 1, 2$.

*Proof.* A detailed proof is Appendix B, we here give the proof idea. Let $A_2 = \partial B_\varepsilon(\theta^0) \cap \{\theta : |\theta_d - \theta_d^0| \leq \eta\} \cap \{\|\beta_2 - \beta_2^0\| \geq \varepsilon/2\}$, for some small $\eta > 0$. Let $A_1$ be the closure of $\partial B_\varepsilon(\theta^0) \cap A_2^c$. The idea is that if $\eta$ is small enough, so that $\theta_d \approx \theta_d^0$ and $\|\beta_2 - \beta_2^0\| \geq \varepsilon/2$ on $A_2$, then the distributions of $W^{(2)}$ implied by $\theta \in A_2$ and $\theta^0$ are different if $X = [x_{1,1}, x_{2,2}, \ldots, x_{N,N}]^{\mathsf{T}}$ has full column rank. That is, $W^{(2)}$ should be able to distinguish every $\theta \in A_2$ from $\theta^0$. Moreover, one can show that on $A_1$ it holds either that $|\theta_d - \theta_d^0| \geq \min(\eta, \varepsilon/4)$ or that $\|\beta_1 - \beta_1^0\| \geq \varepsilon/4$. In either case, $W^{(1)}$ should be able to distinguish $\theta \in A_1$ from $\theta^0$. Formalizing this idea leads to point 1. Point 2 follows from checking the conditions of a uniform law of large numbers [59, Theorem 8.2] and point 3 from points 1 and 2. □

The explicit construction of the subsets $A_1$ and $A_2$, as opposed to using Proposition 2.3.1, warrants an additional comment. Recall, the proposition gives compact $\tilde{A}_1$ and $\tilde{A}_2$ such that $\tilde{A}_1 \cup \tilde{A}_2 = \partial B_\varepsilon(\theta^0)$ and $\nu_\theta^i \neq \nu_{\theta^0}^i$, $\theta \in \tilde{A}_i$, $i = 1, 2$. If one takes $A_i = \tilde{A}_i$, then point

1 in Lemma 2.3.2 follows. Moreover, when the subcollection components are i.i.d., this in turn leads to point 1 in Lemma 2.3.5, which is what is really needed. However, when the distributions of the subcollection components are not identical, this last implication is not true in general.

Having selected appropriate subcollections and subsets it remains only to check that the log-likelihood for the full data satisfies the regularity conditions in Assumptions 2 – 3. The following lemma verifies Assumption 2 and establishes a rate needed for the verification of Assumption 3.

**Lemma 2.3.6.** *If $\theta^0$ is an interior point of $\Theta$, then for every $n$ and small enough $\varepsilon > 0$ there exists a random variable $K_n$ such that, $\mathsf{P}$-almost surely,*

$$\sup_{\theta \in \bar{B}_\varepsilon(\theta^0)} \|\nabla \Lambda_n(\theta; Y)\| \leq K_n = o_{\mathsf{P}}(n^b),$$

*for some $b > 0$.*

Upon inspecting the proof one sees that $b$ can be taken to be $1 + \epsilon$, for any $\epsilon > 0$. This is a better (slower) rate than that obtained in the linear mixed model (see the proof of Lemma 2.3.3). We are now ready to state the main result of the section.

**Theorem 2.3.7.** *If $\theta^0$ is an interior point of $\Theta$ and*

$$\liminf_{N \to \infty} \lambda_{\min} \left( N^{-1} \sum_{i=1}^{N} x_{i,i} x_{i,i}^{\mathsf{T}} \right) > 0,$$

*then, $\mathsf{P}$-almost surely, there exists a sequence $\hat{\theta}_n$ of roots to the likelihood equations*

$$\nabla \Lambda_n(\theta; Y) = 0$$

*such that $\hat{\theta}_n \xrightarrow{\mathsf{P}} \theta^0$.*

*Proof.* The proof is similar to that of Theorem 2.3.4 so we skip some details. We may assume all points in $\bar{B}_\varepsilon(\theta^0)$ are interior points of $\Theta$. By Lemma A.2.5 $\Lambda_n(\theta; Y)$ is differentiable on

$\bar{B}_\varepsilon(\theta^0)$. By Lemma 2.3.5, the identification rate is exponentially fast in $N$ and Lemma 2.3.6 shows that $\Lambda_n(\theta; Y)$ is $K_n$-Lipschitz on both $A_1$ and $A_2$, and that $K_n = o_{\mathsf{P}}(n^b)$ for some $b > 0$. This verifies Assumption 2. By picking $\delta_{i,n} = n^{-b}$ we can have $M_{n,i} = O(n^{[d-1]b})$ as $n \to \infty$, $i = 1, 2$. Thus, $K_n\delta_n = o_{\mathsf{P}}(1)$ and $M_{n,i}a_{n,i} = O(N^{2b[d-1]}e^{-\epsilon N})$ for some $\epsilon > 0$, which is $o(1)$ as $N \to \infty$ since $n = 2N^2$. □

## 2.4 Discussion

Our theory develops the current state-of-the-art asymptotic theory based on subcollections to cover more general cases. The assumptions we make highlight what makes the use of subcollections work. In particular, the interplay between the identification rates of subcollections and the regularity of the likelihood function for the full data is made precise. We note that when the subcollections consist of $m \in \{1, 2, \dots\}$ independent random variables, as in our examples, then if $n = o(m^b)$ for some $b > 0$ and $\nabla\Lambda_n(\theta; Y) = o(n^{b'})$ for some $b' > 0$, uniformly on a compact $\Theta$, the rate conditions are satisfied. This so because, under regularity conditions, the identification rate in a subcollection with $m$ independent random variables is exponential in $m = n^{1/b}$. Since this argument works for arbitrarily large $b$ and $b'$ our theory is operational in a wide range of models. Loosely speaking, if the score function is of less than exponential order in the sample size and there are subcollections of independent random variables that grow faster than logarithmically in the sample size, the MLE is consistent. The conditions should be verifiable in many models since they often require only standard asymptotic tools. For example, in the LMM example nothing more than a uniform law of large numbers and strict positivity of the K–L divergence between distributions corresponding to distinct, identified parameters is needed. Though not pursued here, by inspecting the assumptions of our theory one also sees that it has the potential to be extended to allow the dimension of the parameter set, $d$, grow with $n$. The rates required in our assumptions could be satisfied also if $d$ grows, at least if at a slow enough rate.

Consistency of MLEs have not previously been established in either of the two models to which we apply the general theory. In particular, previous work on asymptotic theory

for MLEs in mixed models often either assumes independent replications of a response vector, that there are no predictors, or no mixed-type responses. We have tried to keep the models here as simple as possible while still illustrating key ideas. Crossed random effects, temporal dependence, and predictors are included because they are challenging theoretically and are commonly used in practice. We have refrained from including things that do not require any new methods but make ideas less transparent. For example, it would be straightforward to include random effects that are not crossed, possibly at the expense of using more subcollections or subcollections consisting of independent vectors of larger dimension than what is now necessary. Similarly, adding several crossed random effects does not make things much harder, only less transparent.

Avenues for future research includes the rate of convergence of the MLEs as well as their asymptotic distribution. Intuitively, one expects MLEs based on the full data to converge at least as fast as the slowest of the subcollection MLEs, that is, the estimators one gets from using only a subset of the full data. There is some evidence of this, namely that, under regularity conditions, the Fisher information in the full data is always larger than that in any subcollection [36]. On the other hand, it is easy to show that, for the simple LMM example in the introduction, the full data MLE converges at the same rate as that based on a subcollection of $N = \sqrt{n}$ i.i.d. observations; that is, at the rate $n^{1/4}$. Given the similarities of the random effect structures, that convergence rate may in future work be a reasonable working hypothesis for MLEs in the MGLMM considered here.

# Chapter 3

# Maximum likelihood estimation of covariance matrices with separable correlation

## 3.1 Introduction

Many statistical applications require the estimation of a covariance matrix $\Sigma$ of some random vector $Y \in \mathbb{R}^q$. When the number of observations is not large enough to estimate an unstructured covariance matrix with acceptable precision, it is common to assume that $\Sigma$ is a function of some lower-dimensional parameter vector, $\theta$. If several different parameterizations are plausible, there is usually a balance to strike between estimation precision and model flexibility. Classical structures such as diagonal with constant variance, first order autoregressive, and compound symmetric require only one or two parameters and can hence be reasonably estimated with few observations. On the other hand, they are too restrictive for many situations. Here, we will focus on parameterizations that lie somewhere between the classical ones just mentioned and an unstructured covariance matrix in terms of flexibility and parameter counts; namely, covariance matrices that have corresponding correlation matrices that are separable.

We say that a covariance matrix is separable, or simply that covariance is separable, if $\Sigma = \Sigma_2 \otimes \Sigma_1$ for some $\Sigma_1 \in \mathbb{S}^r_{++}$ and $\Sigma_2 \in \mathbb{S}^c_{++}$, where $\otimes$ denotes the Kronecker product, $\mathbb{S}^k_{++}$ the set of $k \times k$ ($k = 1, 2 \dots$) symmetric, positive definite matrices, and it is assumed

that $\max(r, c) < q$. Separable covariance requires $r(r+1)/2 + c(c+1)/2 - 1$ parameters, to be compared with the $rc(rc+1)/2$ required for unstructured covariance. Separable covariance is commonly considered in, for example, spatiotemporal statistics [13, section 6.1.3]. We say that correlation is separable, or that $\Sigma$ is a covariance matrix with separable correlation, if it holds that $\Sigma = W(V \otimes U)W$ for some $W \in \mathbb{D}^q_{++}$, $U \in \mathbb{C}^r_{++}$, and $V \in \mathbb{C}^c_{++}$, where $\mathbb{D}^k_{++}$ and $\mathbb{C}^k_{++}$ are the sets of $k \times k$ diagonal positive definite matrices and positive definite correlation matrices, respectively. It is immediate from standard rules of Kronecker products that if $\Sigma = W(V \otimes U)W$ and $W = W_2 \otimes W_1$ for $W_1 \in \mathbb{D}^r_{++}$ and $W_2 \in \mathbb{D}^c_{++}$, then $\Sigma = (W_2 V W_2) \otimes (W_1 U W_1)$, implying that separable covariance is a special case of separable correlation.

In spatiotemporal applications the vector of responses $Y$ can often be more intuitively thought of as a matrix $Y^{(m)} = \mathrm{vec}^{-1}(Y) \in \mathbb{R}^{r \times c}$, where the rows index locations and the columns index time points, for example. Here and elsewhere, $\mathrm{vec}(\cdot)$ denotes the vectorization operator and $\mathrm{vec}^{-1}(\cdot)$ its inverse, the range of which should be clear from context. Now, assuming that the matrix $Y^{(m)}$ is matrix normal with scale parameters $\Sigma_1$ and $\Sigma_2$ is equivalent to assuming that $Y$ is multivariate normal with covariance $\Sigma_2 \otimes \Sigma_1$. When covariance is separable, $\Sigma_1$ is sometimes interpreted as the common covariance matrix of the columns of $Y^{(m)}$ and $\Sigma_2$ as the common covariance matrix of the rows of $Y^{(m)}$ [13]. However, without further restrictions, this interpretation is problematic since $\Sigma_1$ and $\Sigma_2$ are only identified up to scaling by a constant. Indeed, for any $\gamma > 0$, $(\Sigma_2/\gamma) \otimes (\gamma\Sigma_1) = \Sigma_2 \otimes \Sigma_1$. Since $\Sigma_1$ and $\Sigma_2$ are unidentified one usually has to focus estimation and interpretation only on their Kronecker product $\Sigma_1 \otimes \Sigma_2$ [69]. For fitting purposes an identifiability constraint such as $\|\Sigma_1\| = 1$ (unit spectral norm) or $(\Sigma_1)_{1,1} = 1$ (unit leading entry) is often imposed. Though this may be computationally convenient it does not in general solve the issue of interpretation unless one has reason to believe, e.g., that the leading element of $Y$ has unit variance. By contrast, with separable correlation the matrices $U$ and $V$ are indeed the correlation matrices of the columns and rows, respectively, and $W$, $U$, and $V$ are all identified without further restrictions; this is proven in Section 3.2.

The idea of separable correlation is not new and it is well known that for certain appli-

cations it is more appropriate than separable covariance [27]. In particular, the fact that $W$ is not in general separable means that the standard deviations of the elements in $Y$ all have their own parameter and they are in that sense unrestricted, unlike with separable covariance. For this flexibility one pays the price of estimating an additional $rc - r - c + 1$ parameters, which means that the number of parameters required for separable correlation and covariance is of the same order as $rc \to \infty$. There are, however, some important differences between the two notions of separability that have not been stressed in the literature. Most importantly for our discussion, algorithms for maximum likelihood estimation that are designed for separable covariance matrices do not carry over in an obvious way to the estimation of covariance matrices with separable correlation. One option is to use standard descent methods such as Newton's. However, Lu and Zimmerman [48] report that for the separable covariance model, Newton's algorithm is as much as 5000 times slower than the popular flip-flop algorithm [17] when $r = c = 6$, and the difference seems to be increasing in $r$ and $c$. The flip-flop algorithm is comparatively fast because, up to a rescaling step, it is a (block) coordinate descent algorithm where all coordinatewise optimization sub-problems admit closed form solutions. Our algorithm builds on similar ideas. It is fully likelihood-based and recovers true MLEs. The algorithm is also stable, decreasing the likelihood at every iteration in a coordinate descent-type fashion.

In the next section we present the algorithm. Section 3.3 contains a brief discussion of inference about covariance structures using a parametric bootstrap, Section 3.4 contains simulations, and Section 3.5 a data example. Section 3.6 concludes with a discussion of the proposed methods and the results.

## 3.2 Maximum likelihood

Assume that $y_1, \ldots, y_n$ are realizations of $n$ independent, multivariate normal $q$-vectors $Y_1, \ldots, Y_n$ with means $\mathbb{E}(Y_i) = \mathcal{B}^\mathsf{T} x_i$, for some vector of non-stochastic predictors $x_i \in \mathbb{R}^p$, parameter matrix $\mathcal{B} \in \mathbb{R}^{p \times q}$, and covariance matrix $\Sigma = W(V \otimes U)W \in \mathbb{S}^q_{++}$. These assumptions may be equivalently stated as: if $Y_i^{(m)}$ is a matrix such that $Y_i = \mathrm{vec}(Y_i^{(m)})$

and $\tilde{Y}_i^{(m)}$ is the matrix obtained by subtracting the mean and dividing by the standard deviation elementwise in $Y_i^{(m)}$, then $\tilde{Y}_i^{(m)}$ has a matrix normal distribution with mean 0 and scale matrices $U$ and $V$.

Twice the negative log-likelihood is, up to an additive constant,

$$g_f(\mathcal{B}, U, V, W; y) = \log |W(V \otimes U)W| + \operatorname{tr}\left(S(\mathcal{B})[W(V \otimes U)W]^{-1}\right), \tag{3.1}$$

where $S(\mathcal{B}) = \sum_{i=1}^{n}(y_i - \mathcal{B}^{\mathsf{T}}x_i)(y_i - \mathcal{B}^{\mathsf{T}}x_i)^{\mathsf{T}}/n$, $|\cdot|$ denotes the determinant when applied to matrices, and subscript $f$ stands for full, so as to distinguish $g_f$ from a partially minimized version to be defined shortly. For simplicity the dependence on the sample $y = (y_1, \ldots, y_n)$ is often implicit and we write $g_f(\mathcal{B}, U, V, W) = g_f(\mathcal{B}, U, V, W; y)$. Formally, $g_f$ is defined on the parameter set $\Theta = \mathbb{R}^{p\times q} \times \mathbb{C}^r_{++} \times \mathbb{C}^c_{++} \times \mathbb{D}^q_{++}$.

Allowing for general non-stochastic predictors $x_i \in \mathbb{R}^p$ makes the development no more difficult than assuming $x_i \equiv 1$. Indeed, assuming that $n > p$, $g_f(\mathcal{B}, U, V, W)$ is uniquely partially minimized in $\mathcal{B}$ at the least squares estimate $\hat{\mathcal{B}} = (X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}[y_1, \ldots, y_n]^{\mathsf{T}}$, where $X = [x_1, \ldots, x_n]^{\mathsf{T}} \in \mathbb{R}^{n\times p}$. Thus, after partially minimizing in $\mathcal{B}$, the objective function of interest is

$$g(U, V, W) = g_f(\hat{\mathcal{B}}, U, V, W).$$

Minimizing $g$ over $\Psi = \mathbb{C}^r_{++} \times \mathbb{C}^c_{++} \times \mathbb{D}^q_{++}$ gives the maximum likelihood estimates of the true parameters that we henceforth denote by $U^\star, V^\star$, and $W^\star$ to distinguish them from the corresponding optimization variables. We write $g(\psi)$ for $g$ evaluated at $\psi = (U, V, W) \in \Psi$.

It is immediate from the introductory discussion that if $W$ is restricted to be separable, with dimensions conforming to those of $U$ and $V$, then the optimization problem is equivalent to maximizing the matrix normal likelihood. This can often be done using the popular flip-flop algorithm, assuming that $n$ is large enough in comparison to $r$ and $c$ [17, 69]. Several useful properties have been proven about the likelihood for the separable covariance model, including conditions that ensure existence of MLEs and the convergence of optimization algorithms to that MLE [69]. Unfortunately, the proofs of those results

do no carry over in an obvious way to the separable correlation model. One issue is that the separable covariance model has a geodesically convex log-likelihood [81], but this is not known to be true for the separable correlation model. Instead, we get by other means the following theorem that gives two useful theoretical properties of $g_f$.

**Theorem 3.2.1.** *If $X$ has full column rank and $S(\hat{\mathcal{B}}) \in \mathbb{S}^q_{++}$, then the parameter vector $\theta = (\mathcal{B}, W, U, V)$ is identified and $g_f$ has a global minimizer over $\Theta$.*

*Proof.* Take arbitrary $\theta = (\mathcal{B}, W, U, V)$ and $\theta' = (\mathcal{B}', W', U', V)$. We must show that if $\theta \neq \theta'$, then $g_f(\theta; y)$ and $g_f(\theta'; y)$ are different on a set of $y$s with positive Lebesgue measure. Since the multivariate normal distribution is characterized by its mean vector and covariance matrix, it suffices to show that $\theta \neq \theta'$ implies either different means or different covariance matrices. If $\mathcal{B} \neq \mathcal{B}'$, then since $X$ has full column rank $X\mathcal{B} \neq X\mathcal{B}'$, and we are done. Similarly, if $W \neq W'$ then there is at least one diagonal entry where $\Sigma = W(V \otimes U)W$ differs from $\Sigma' = W'(V' \otimes V')W'$, and again we are done. Suppose next that $W = W'$ but $U \neq U'$. Since $W$ is invertible, $\Sigma = \Sigma'$ if and only if $V \otimes U = V' \otimes U'$. But the leading $r \times r$ blocks of the two sides are $U$ and $U'$, and hence $V \otimes U \neq V' \otimes U'$. A similar argument shows that $W = W'$ and $V \neq V'$ leads to different covariance matrices, which finishes the proof of identifiability.

The second conclusion of the theorem uses sufficient conditions for existence of MLEs due to Burg et al. [9]. The proof is similar to the proof of Theorem 2 in Roś et al. [64]. The sufficient conditions are that (i) $S(\hat{\mathcal{B}})$ is invertible and that (ii) for any converging sequence $\Sigma^k$, $k = 1, 2 \ldots$, such that $\Sigma^k = W^k(V^k \otimes U^k)W^k$, $W^k \in \mathbb{D}^q_{++}$, $U^k \in \mathbb{C}^r_{++}$, and $V^k \in \mathbb{C}^c_{++}$, it holds that its limit $\bar{\Sigma}$ is either non-negative definite and singular or can be written as $\bar{W}(\bar{R}_2 \otimes \bar{R}_1)\bar{W}$, for some $\bar{W} \in \mathbb{D}^q_{++}$, $\bar{R}_1 \in \mathbb{C}^r_{++}$, and $\bar{R}_2 \in \mathbb{C}^c_{++}$. Condition (i) is by assumption so let us prove that (ii) is satisfied. It is immediate from positive definiteness of $\Sigma^k$, $k = 1, 2, \ldots$, that $\liminf_{k\to\infty} v^\mathsf{T}\Sigma^k v \geq 0$ for any $v \in \mathbb{R}^n$. Thus, since $\Sigma \mapsto v^\mathsf{T}\Sigma v$ is a continuous mapping, $\bar{\Sigma}$ is non-negative definite. Suppose $\bar{\Sigma}$ is positive definite. Since $\Sigma^k$ converges, so do its diagonal entries. Thus, since the entries of $W^k$, for any $k$, and $\bar{W}$ are positive, $W^k \to \bar{W}$. Thus, since $\Sigma^k$ is convergent, the sequence $R^k = V^k \otimes U^k$ is a

convergent sequence of separable covariance matrices. Theorem 2 in Roś et al. [64] says the limit $\bar{R}$ of $R^k$ is also a separable covariance matrix, though not necessarily a correlation matrix. But the diagonal entries of $R^k$ are all unity, for all $k$, so it must be true also for its limit $\bar{R}$, which finishes the proof. □

---

**Algorithm 3.1** Maximum Likelihood with Separable Correlation

*Input:* $U^0, V^0, w_1^0, \ldots, w_q^0, k = 0$
**repeat**
  Set $\tilde{U}^{k+1} = \arg\min_{U \in \mathbb{S}^c_{++}} g(U, V^k, w_1^k, \ldots, w_q^k)$
  Set $\tilde{V}^{k+1} = \arg\min_{V \in \mathbb{S}^r_{++}} g(\tilde{U}^{k+1}, V, w_1^k, \ldots, w_q^k)$
  Set $U^{k+1} = (\tilde{U}^{k+1} \circ I_r)^{-1/2} \tilde{U}^{k+1} (\tilde{U}^{k+1} \circ I_r)^{-1/2}$
  Set $V^{k+1} = (\tilde{V}^{k+1} \circ I_c)^{-1/2} \tilde{V}^{k+1} (\tilde{V}^{k+1} \circ I_c)^{-1/2}$
  Set $\tilde{W}^{k+1} = W^k[(\tilde{U}^{k+1} \otimes \tilde{V}^{k+1}) \circ I_q]^{1/2}$
  **for** $j = 1, \ldots, q$ **do**
    Set $w_j^{k+1} = \arg\min_{w \in (0,\infty)} g(U^{k+1}, V^{k+1}, w_1^{k+1}, \ldots, w, \tilde{w}_{j+1}^{k+1}, \ldots, \tilde{w}_q^{k+1})$
  **end for**
  $k \leftarrow k + 1$
**until** $|g(U^k, V^k, W^k) - g(U^{k-1}, V^{k-1}, W^{k-1})| \leq \epsilon$

---

Our algorithm that we now describe is summarized in Algorithm 3.1. The algorithm uses (block) coordinatewise updating of the optimization variables together with rescaling. When deriving updates for $W$ we consider the diagonal entries individually and write, with some overloading of notation, $g(U, V, w_1, \ldots, w_q) = g(U, V, \mathrm{diag}(w_1, \ldots, w_q))$, where $\mathrm{diag}(w_1, \ldots, w_q) = W$. Some steps in the derivation of the algorithm we propose require extending the function $g$ to the larger domain $\bar{\Psi} = \mathbb{S}^r_{++} \times \mathbb{S}^c_{++} \times \mathbb{D}^q_{++}$. That is, in some steps we allow $U$ and $V$ to be covariance matrices rather than only correlation matrices. On this larger domain there are infinitely many points that give the same $\Sigma$. Indeed, by the same argument as for separable covariance matrices, if the diagonal entries of $U$ and $V$ are not necessarily unity, then one can scale either by a factor $\gamma > 0$ and the other by a factor $1/\gamma$ without affecting the value of the objective function $g$. However, the following proposition shows that we can identify each point in $\bar{\Psi}$ with exactly one point in $\Psi$ that results in the same covariance matrix, and hence objective function value.

**Proposition 3.2.2.** *To every $W' \in \mathbb{D}^q_{++}$, $U' \in \mathbb{S}^r_{++}$, and $V' \in \mathbb{S}^c_{++}$ there correspond unique $W \in \mathbb{D}^q_{++}$, $U \in \mathbb{C}^r_{++}$, and $V \in \mathbb{C}^c_{++}$ such that $W'(V' \otimes U')W' = W(U \otimes V)W$.*

*Proof.* Take $W = W'(I_q \circ [V' \otimes U'])^{1/2}$, $U = U'(I_r \circ U')^{-1/2}$, and $V = V'(I_r \circ V')^{-1/2}$. Now the existence part follows from standard properties of Kronecker products and uniqueness follows from Theorem 3.2.1 □

It is clear from the constructive proof of existence in Proposition 3.2.2 that if we are given a point in $\bar{\Psi}$, the equivalent point in $\Psi$ is easy to compute in practice. We will use this in the derivation of the algorithm that now follows.

Let $\psi^k = (U^k, V^k, w_1^k, \ldots, w_q^k) \in \Psi$ be the current iterate and let $\tilde{U}^{k+1}$ be the solution to $\nabla_{U^{-1}} g(U, V^k, w_1^k, \ldots, w_q^k) = 0$. It is useful for this update to note that $g$ can be equivalently written as

$$g(U, V, W) = \log |W(V \otimes U)W| + n^{-1} \sum_{i=1}^{n} \operatorname{tr}\left[E_i V^{-1} E_i^{\mathsf{T}} U^{-1}\right], \tag{3.2}$$

where the $E_i$ are $r \times c$ matrices defined by $\operatorname{vec}(E_i) = W^{-1}(y_i - \hat{\mathcal{B}}^{\mathsf{T}} x_i)$ $(i = 1, \ldots, n)$. By differentiating (3.2) and solving the resulting first order condition one finds that

$$\tilde{U}^{k+1} = \frac{1}{nc} \sum_{i=1}^{n} E_i^k (V^k)^{-1} (E_i^k)^{\mathsf{T}},$$

where $\operatorname{vec}(E_i^k) = (W^k)^{-1}(y_i - \hat{\mathcal{B}}^{\mathsf{T}} x_i)$. When $\tilde{U}^{k+1}$ has full rank, it is the unique partial minimizer of $g$—this is straightforward to prove formally by differentiating again and noticing that the Hessian for $\operatorname{vec}(U^{-1})$ is positive definite. The matrix $\tilde{U}^{k+1}$ is symmetric, positive semi-definite by construction; we have not been able to prove that $\tilde{U}^{k+1}$ is positive definite with probability one in general. In practice $\tilde{U}^{k+1}$ is positive definite for all $k$ whenever $n$ is large enough, both in our data example and simulations. From here on, unless otherwise noted, we assume that the iterates $\tilde{U}^k$ are indeed positive definite for all $k$.

Next let $\tilde{V}^{k+1}$ be the solution to $\nabla_{V^{-1}} g_1(\tilde{U}^{k+1}, V, w_1^k, \ldots, w_q^k) = 0$, which implies that

$$\tilde{V}^{k+1} = \frac{1}{nr} \sum_{i=1}^{n} (E_i^k)^\intercal (\tilde{U}^{k+1})^{-1} E_i^k.$$

The properties of $\tilde{V}^{k+1}$ are analogous to those of to $\tilde{U}^{k+1}$. Most importantly, $\tilde{V}^{k+1}$ is the unique partial minimizer of $g$ whenever it has full rank.

The point $\tilde{\psi}^{k+1} = (\tilde{U}^{k+1}, \tilde{V}^{k+1}, w_1^k, \ldots, w_q^k)$, typically lies in $\bar{\Psi}$ but not in $\Psi$, meaning $\tilde{U}^{k+1}$ and $\tilde{V}^{k+1}$ are not correlation matrices. However, by construction, $g(\tilde{\psi}^{k+1}) \leq g(\psi^k)$. Using the results in Proposition 3.2.2, the updates $U^{k+1}$ and $V^{k+1}$ are now found by rescaling $\tilde{U}^{k+1}$ and $\tilde{V}^{k+1}$ to be correlation matrices, and $W^k$ is rescaled accordingly so as to make sure the rescaling does not affect the value of the objective function (see lines 5 – 7 in Algorithm 3.1). Let us denote the rescaled version of $W^k$ by $\tilde{W}^k$. Lastly, $w_j^{k+1}$ $(j = 1, \ldots, q)$ are set to the positive solutions of

$$\nabla_{w_j^{-1}} g(U^{k+1}, V^{k+1}, w_1^{k+1}, \ldots, w_{j-1}^{k+1}, w_j, \tilde{w}_{j+1}^k, \ldots, \tilde{w}_q^k) = 0.$$

For less cluttered notation, we ignore the iteration index when motivating this update. Differentiating $g$ with respect to $w_j^{-1}$ twice gives

$$\nabla_{w_j^{-1}} g = -2n \left[ w_j - \sum_{l=1}^{rc} R_{j,l}^{-1} w_l^{-1} S_{l,j} \right] \quad \text{and} \quad \nabla^2_{w_j^{-1}} g = n[w_j^2 + R_{j,j}^{-1} S_{j,j}],$$

where $R = V \otimes U$. Multiplying through the first order condition by $w_j$ gives the quadratic equation $w_j^2 - w_j \sum_{l \neq j} R_{j,l} w_l^{-1} S_{l,j} - R_{j,j}^{-1} S_{j,j} = 0$. Thus, since $w_j$ has to be positive, $w_j = (a + \sqrt{a^2 + b})/2$, where $a = \sum_{l \neq j} R_{j,l}^{-1} w_l^{-1} S_{l,j}$ and $b = 4R_{j,j}^{-1} S_{j,j} \geq 0$. This root is almost surely positive when the iterates $U^{k+1}$ and $V^{k+1}$ are positive definite, since then $(U^{k+1} \otimes V^{k+1})^{-1}_{j,j} > 0$, and $S_{j,j} > 0$ almost surely when $n > 1$. By the same argument, the second derivative is positive, and thus the updates are the unique partial minimizers.

The proposed algorithm by design weakly decreases the objective function at every successful iteration. Every iteration requires of the order $r^3 + c^3 + n(cr^2 + rc^2)$ floating

point operations, where the cubic terms come from Cholesky decompositions of the iterates of $U$ and $V$, and the squared terms from matrix multiplications and solving linear systems when updating $U$ and $V$. The algorithm can be terminated when, for example, the decrease in the objective function reaches below a certain threshold or when the change in the iterates does. Of course, relative changes may also be considered in place of absolute ones.

## 3.3 Inference

In applications it is often of interest to determine if the assumption of separable correlation is reasonable before proceeding to fitting the model. We will consider two hypothesis tests relevant for this purpose, namely (a) $H_0$: Separable correlation v. $H_A$ : Non-separable correlation, and (b) $H_0$ : Separable covariance v. $H_A$ : Separable correlation. For both (a) and (b) we suggest using bootstrapped likelihood ratio tests since, as will be clear from the simulation results, classical likelihood ratio tests for (a) and (b) are in general conservative. For testing (a) we assume that $S(\hat{\mathcal{B}}) \in \mathbb{S}^q_{++}$ since otherwise the log-likelihood under the alternative is unbounded, implying the test would reject with probability one.

Some more notation is needed to define the parametric bootstrap we consider. Let $\hat{\Sigma}^0_N$ denote the MLE of $\Sigma$ under the null hypothesis, using the original data $(y_i, x_i), i = 1, \ldots, n$. For test (a) $\hat{\Sigma}^0_N$ is the separable correlation MLE, and for test (b) it is the separable covariance MLE. Let also $\hat{\Sigma}^0_A$ be the MLE under the alternative, which for test (a) means the sample covariance and for test (b) the separable correlation MLE. For any other integer $k \geq 1$, the estimates $\hat{\Sigma}^k_N$ and $\hat{\Sigma}^k_A$ are similarly defined but with the original data replaced by the $k$th bootstrapped dataset $(y^k_i, x_i), i = 1, \ldots, n$. The parametrically bootstrapped responses $y^k_1, \ldots, y^k_n$ are generated from the null hypothesis model with $\mathcal{B}$ and $\Sigma$ replaced by $\hat{\mathcal{B}}$ and $\hat{\Sigma}^0_N$. Finally, let $\xi^k = \xi(\hat{\Sigma}^k_N, \hat{\Sigma}^k_A)$ denote the likelihood ratio based on the $k$th bootstrapped dataset, where as before $k = 0$ means the original data. The bootstrap procedure is presented in Algorithm 3.2

We examine the performance of this testing procedure in the next section.

**Algorithm 3.2** Parametric Bootstrap Test

*Input:* $(y_i, x_i) \in \mathbb{R}^q \times \mathbb{R}^p$, $i = 1, \ldots, n$, nominal level $\alpha \in (0, 1)$, and $B \in \{1, 2, \ldots\}$
Compute $\hat{\mathcal{B}}$ and $\hat{\Sigma}_N^0$ using $(y_i, x_i), i = 1, \ldots, n$
**for** $k = 1, \ldots, B$ **do**
    Generate $y_1^k, \ldots, y_n^k$ as independent multivariate normal random variables with means $\hat{\mathcal{B}}^\mathsf{T} x_i$, $i = 1, \ldots, n$, and common covariance matrix $\hat{\Sigma}_N^0$
    Compute $\hat{\Sigma}_N^k$ and $\hat{\Sigma}_A^k$ using $(y_i^k, x_i), i = 1, \ldots, n$
    Compute and store the $k$th likelihood ratio $\xi^k$
**end for**
Reject the null hypothesis if $\xi^0$ is smaller than the $\alpha$th empirical quantile of $\{\xi^1, \ldots, \xi^B\}$.

## 3.4 Simulations

All simulation results are summarized in Table 3.1. We present simulation-based estimates of rejection rates of the statistical tests and of spectral norm errors of MLEs of $\Sigma$.

The data generating process in the simulations is a separable correlation model with $x_i = 1$, $\mathcal{B} = 0$, $U^\star = (U^\star_{i,j}) = (1/2^{|i-j|})$ $(i = 1, \ldots, r, j = 1, \ldots, r)$, and $V^\star = (V^\star_{i,j}) = (1/2^{|i-j|})$ $(i = 1, \ldots, c, j = 1, \ldots, c)$. That is, both correlation matrices have a first order autoregressive structure with correlation parameter 1/2. The standard deviation matrix is either $W^\star = I_q$, the $q \times q$ identity matrix, so that the separable covariance model is also correct, or the diagonal entries of $W^\star$ are evenly spaced numbers between 0.1 and 10, which is not separable. We also considered other generating processes with, for example, compound symmetric correlation matrices or those drawn randomly from a Wishart distribution rescaled to have unit diagonal entries, and the results were qualitatively similar. Some additional simulation results can be found in Appendix B. The maximum number of iterations in the algorithm was set to 10000, and the algorithm was terminated if the absolute change in the objective function was less than $\epsilon = 10^{-10}$ in an iteration.

The reported average spectral norm error for any given configuration $(n, r, c)$ is $err_{cor} = m^{-1} \sum_{j=1}^m \|\hat{\Sigma}^j_{cor} - \Sigma^\star\|$, where $\hat{\Sigma}^j_{cor}$ is the separable correlation estimator based on the $j$th simulated dataset. The definitions of $err_{cov}$ and $err_{ur}$ are similar but with the separable correlation estimate replaced by the separable covariance estimate, and the sample covariance (unrestricted ML) estimate, respectively. The columns labeled $rej$ with subscript $cov$,

$(cov, b)$, $cor$, or $(cor, b)$ are estimated rejection rates for statistical tests. The subscript $cov$ or $cor$ indicates if the null hypothesis is separable covariance or correlation, and the subscript $b$ indicates the bootstrap procedure in Algorithm 3.2 was used. If the bootstrap was not used, then the tests are usual likelihood ratio tests. Next we summarize the results and start with the estimation error.

For all configurations where the separable covariance model is correct, the average spectral norm errors are lower for the estimates using this assumption than those assuming only separable correlation (upper panel, Table 3.1). This is unsurprising since separable correlation requires the estimation of more parameters. Comparisons to the unrestricted maximum likelihood estimator are only made when $n - p > q$, i.e. when it exists. The unrestricted estimates have at worst about five times higher average error than the ones based on separable correlation ($n = 320, r = c = 15$), and at best about two times higher ($n = 160, r = c = 5$).

When data is generated with separable correlation (lower panel, Table 3.1), the separable covariance estimates have lower average spectral norm errors when the sample size is small in comparison to $r$ and $c$. For larger sample sizes, the separable correlation estimates have lower average errors. That is, there is a bias-variance trade-off in that fitting the incorrect model may yield lower errors due to higher precision in the estimates when the sample size is small.

Consider next columns 7 – 10 in the upper panel. The parametric bootstrap based test has close to nominal size, 0.05, in all settings considered, whether the null hypothesis is separable covariance or separable correlation. Notice, both of these are correct when covariance is separable. On the other hand, the standard likelihood ratio test only has nominal size when $n$ is large in comparison to $r$ and $c$ and the null hypothesis is separable covariance. When the null hypothesis is separable correlation, the usual likelihood ratio test is very conservative for almost all configurations. Only when $n = 320$ and $r = c = 5$ is the size close to nominal, 0.12 compared to 0.05. To assess large sample validity of the test, we also ran a simulation with $n = 1000$, $r = c = 5$, and that sample size was large enough to yield near-nominal rejection rates for the usual likelihood ratio test.

When data is generated with separable correlation but not separable covariance (lower panel), then the empirical rejection rates for the bootstrap based test with the null hypothesis that covariance is separable varies between 0.04 and 1.0. Power is low when either the number of observations is small in comparison to $r$ and $c$, or one of $r$ and $c$ is small, or both. For small $n$, here $n = 10$ or $n = 20$, the power is lower when $r = 5, c = 15$ than when $r = c = 15$ even though $n$ is smaller in comparison to $q = rc$ in the latter case. The difference in parameter counts between the two models is $15 \times 5 - (15 + 5 - 1) = 56$ in the former case but $15^2 - (15 + 15 - 1) = 196$ in the latter, which is likely the reason rejection rates are higher under the latter regime.

It is illuminating that in some settings, for example when $n = 20$ and $r = c = 5$, the separable covariance model seems the better choice if one is interested in estimating the covariance matrix with small error, but our simulations indicate statistical testing rejects that model with high probability. Thus, the choice between the models should depend on the purpose for which they are to be used, not only on which model is correct.

### 3.4.1 Convergence diagnostics

The proposed algorithm can terminate in four ways: it converges, meaning the change between two iterations in the objective function is less than some threshold $\epsilon$, the maximum number of iterations are reached, an update of $U$ is indefinite, or an update of $V$ is indefinite. We have performed extensive simulations examining what causes the algorithm to terminate for different values of $n, r$, and $c$. A summary of these results are in Appendix B. In general, the algorithm always converges when $n - p$ is slightly greater than $\max(r/c, c/r)$. For example, when $x_i = 1$, $c = 9$, and $r = 2$, simulations indicate the algorithm always converges when $n \geq 10$. For $n \leq 4$, the algorithm terminates because an iterate of $U$ is indefinite in almost every simulation run. For $n = 6$, the algorithm either converges (about 60% of the time), reaches the maximum number of iterations (about 10% of the time), or terminates because and update for $U$ is indefinite. For $n = 8$ all updates are positive definite, but sometimes the algorithm does not converge before the maximum number of iterations (1000) is reached (about 20% of the time). The pattern is similar for different

Table 3.1: Estimation Error and Test Size

| $n$ | $r$ | $c$ | $err_{cor}$ | $err_{cov}$ | $err_{ur}$ | $rej_{cov}$ | $rej_{cov,b}$ | $rej_{cor}$ | $rej_{cor,b}$ |
|---|---|---|---|---|---|---|---|---|---|
| Data generated with separable covariance | | | | | | | | | |
| 10 | 5 | 5 | 2.74 | 2.20 | - | 0.29 | 0.06 | - | - |
| 10 | 5 | 15 | 5.60 | 3.50 | - | 0.90 | 0.02 | - | - |
| 10 | 15 | 15 | 5.52 | 3.58 | - | 0.97 | 0.03 | - | - |
| 20 | 5 | 5 | 1.67 | 1.49 | - | 0.12 | 0.05 | - | - |
| 20 | 5 | 15 | 2.70 | 2.34 | - | 0.37 | 0.06 | - | - |
| 20 | 15 | 15 | 2.92 | 2.39 | - | 0.51 | 0.04 | - | - |
| 160 | 5 | 5 | 0.57 | 0.52 | 1.19 | 0.06 | 0.06 | 0.21 | 0.06 |
| 160 | 5 | 15 | 0.87 | 0.81 | 2.44 | 0.06 | 0.05 | 1.00 | 0.05 |
| 160 | 15 | 15 | 0.92 | 0.81 | - | 0.09 | 0.07 | - | - |
| 320 | 5 | 5 | 0.39 | 0.36 | 0.84 | 0.05 | 0.05 | 0.12 | 0.06 |
| 320 | 5 | 15 | 0.62 | 0.58 | 1.70 | 0.07 | 0.07 | 0.97 | 0.05 |
| 320 | 15 | 15 | 0.65 | 0.58 | 3.51 | 0.05 | 0.05 | 1.00 | 0.04 |
| Data generated with separable correlation | | | | | | | | | |
| 10 | 5 | 5 | 15.62 | 13.41 | - | 0.87 | 0.50 | - | - |
| 10 | 5 | 15 | 34.73 | 20.35 | - | 0.95 | 0.04 | - | - |
| 10 | 15 | 15 | 33.86 | 22.43 | - | 1.00 | 0.14 | - | - |
| 20 | 5 | 5 | 9.87 | 9.69 | - | 0.99 | 0.97 | - | - |
| 20 | 5 | 15 | 15.95 | 13.89 | - | 0.82 | 0.36 | - | - |
| 20 | 15 | 15 | 18.48 | 15.23 | - | 0.94 | 0.50 | - | - |
| 160 | 5 | 5 | 3.28 | 5.16 | 6.53 | 1.00 | 1.00 | 0.23 | 0.06 |
| 160 | 5 | 15 | 5.31 | 5.46 | 13.86 | 1.00 | 1.00 | 1.00 | 0.05 |
| 160 | 15 | 15 | 6.01 | 6.37 | - | 1.00 | 1.00 | - | - |
| 320 | 5 | 5 | 2.37 | 4.82 | 4.63 | 1.00 | 1.00 | 0.12 | 0.06 |
| 320 | 5 | 15 | 3.68 | 4.19 | 9.63 | 1.00 | 1.00 | 0.97 | 0.05 |
| 320 | 15 | 15 | 4.23 | 5.17 | 19.59 | 1.00 | 1.00 | 1.00 | 0.05 |

Columns labeled $err$ show average spectral norm errors. Subscripts indicate separable correlation, separable covariance, and unrestricted estimators. Columns with label $rej$ show empirical rejection rates. The subscripts indicate the null hypotheses covariance separability and correlation separability. A second subscript $b$ indicates the parametric bootstrap was used. With one exception, the largest standard errors for entries in columns 4 – 10 are, respectively, 0.04, 0.001, 0.02, 0.02, 0.02, 0.02, and 0.01. The standard error for $err_{cor}$ with $n = 10$, $r = c = 15$ is 0.08.

configuration of our model. If we take reaching the maximum number of iterations as a sign of there not existing a unique maximum of the log-likelihood, these results are qualitatively consistent with theory developed by Soloveychik and Trushin [69], who show that, with separable covariance and $x_i = 1$, no unique minimum exists if $n < \max(r/c, c/r) + 1$, whereas if $n > r/c + c/r + 1$ a unique minimum exists almost surely, and the flip-flop algorithm converges to this unique minimum almost surely from any starting point. For $n$ in the gap between these two cutoffs, a unique minimum exists with positive probability strictly less than one.

## 3.5 Data example

We illustrate our model using data on dissolved oxygen concentration in the Mississippi River and are interested estimating the covariance matrix for measurements from a number of different areas and time points. Additional details on the modeling choices can be found in Appendix B. We consider a mean structure where the regressors $x_i$ $(i = 1, \ldots, n)$ are those resulting from fitting cubic splines in the year index. Since the choice of predictors does not affect the estimation procedure for $\Sigma$, we do not discuss it in detail.

Our data consist of $n = 21$ years (1994 – 2002, 2004 – 2015) of quarterly measurements from $r = 16$ areas of the Upper Mississippi River. Observations from winter are excluded since water is typically mostly frozen in the northernmost sampling areas in winter, so $c = 3$. The areas represent sampling strata-river reach combinations; the data were collected by the US Army Corps of Engineers' Upper Mississippi River Restoration Program Long Term Resource Monitoring element [39].

We treat dissolved oxygen measurements as independent between years. In the raw data there are several measurements for every year, season, and area. For convenience, we model the sample means of these measurements. That is, we let $Y_{i,j,k}$ $(i = 1, \ldots, n,\ j = 1, \ldots, 16,\ k = 1, 2, 3)$ denote the sample mean of the dissolved oxygen measurement in season $k$, area $j$, and year $i$. For the sample means data, the separable correlation model says that the correlations between mean dissolved oxygen concentrations in spring, summer, and fall are

the same for all areas. Additionally, the correlations between mean concentrations from different areas are the same in all seasons. If separability or some other restrictions were not imposed in addition to independence between years, then the covariance matrix $\Sigma$ would be of size $48 \times 48$ in this example, and we only have $n = 21$ years of data. Thus, in this example, correlation separability is motivated both as a parsimonious parameterization enabling maximum likelihood estimation, and as a reflection of the spatiotemporal structure of the data. It is also of scientific interest to examine whether, in fact, the separable covariance model can be applied to our data.

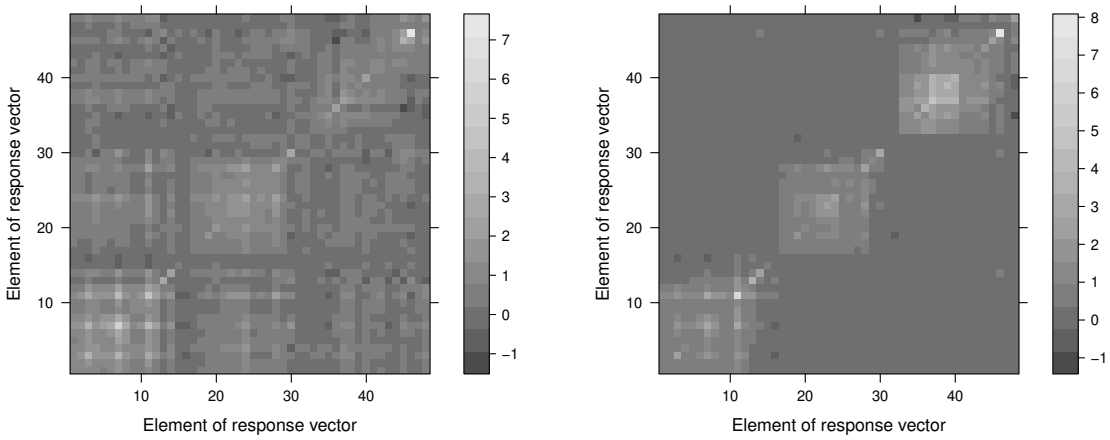
We ran the the parametric bootstrap test of $H_0$ : Separable covariance v. $H_A$ : Separable correlation with $B = 10000$. In none of these datasets did we observe a likelihood ratio test statistic larger than what was observed in the real data, suggesting a $p-$value less than 0.0001. The parametric bootstrap of course relies on the model being a good approximation for the data generating process. It is outside the scope of this article to examine this rigorously, but as a quick check we plotted quantiles of the standardized residuals $r_i = \hat{\Sigma}_{cor}^{-1/2}(y_i - \hat{\mathcal{B}}^T x_i)$ $(i = 1, \ldots, n)$ against those of a normal distribution. This revealed a somewhat heavy left tail but otherwise good agreement between theoretical and empirical quantiles; the plot is excluded for brevity.

The null hypothesis of correlation separability cannot be formally tested with a likelihood ratio test since $n = 21 < 48 = q$. However, inspecting the residual covariance matrix $S(\hat{\mathcal{B}})$ and comparing it to the estimate from the separable correlation model may still give some informal indications of how well the model fits. In the heatmap of the residual covariance matrix $S$ in Fig. 3.1, there is a larger variation in off-block-diagonal covariances than in the heatmap of the separable correlation estimate. The block-diagonal structure with three main blocks is more pronounced in the plot of the estimate with separable correlation. These differences may indicate either a lack of fit or that the residual covariance matrix is picking up on noise in the data. The three blocks, from lower left to upper right, correspond to measurements from spring, summer, and fall. The lower off-block-diagonal covariances indicate dependence between observations from different seasons is in weak in general. The estimated variance of the 46th response element is particularly large, this can be seen also

in Fig. 3.2. At the sampling location corresponding to this element, the river exhibits an unusual combination of deep water and low flow; measurements are also taken in fall when flow is generally lower than in spring and summer [1]. Together, these characteristics might lead to larger swings in average concentrations of dissolved oxygen among years than for other reach-stratum combinations in fall, thus offering a possible physical explanation for our finding. Assuming it is a real finding, it is an important discovery that is inconsistent with separable covariance where any season effect on the variance of measurements has to be the same for every sampling location. Plots of the variance estimates in Fig. 3.2 reveal other elements of the response vector where the difference between the models is even greater. For example, the separable correlation estimates indicate variation is particularly large at sampling location 11 in spring, that is, for element 11 of the response vector. Although outside the scope of this article, investigating further why variability in measurements is particularly large at this location and season can potentially lead to interesting scientific findings that would not be discovered if assuming separable covariance.

[1] J. Rogala, personal communication with Brian Gray, 29 Jan 2018

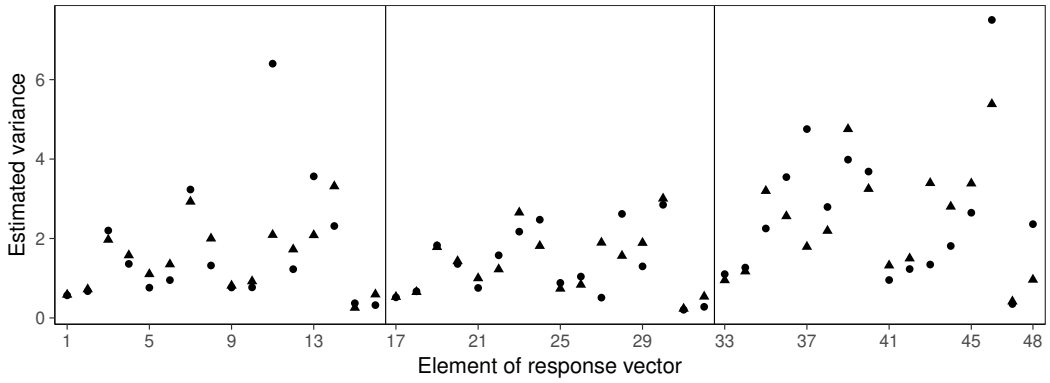
Figure 3.1: Estimated covariances for dissolved oxygen data




(a) Residual covariance matrix (b) Estimate with separable correlation

Grayscale indicates covariance.

Figure 3.2: Estimated variances for dissolved oxygen data




Dots and triangles correspond to estimates from separable correlation and separable covariance models, respectively. The three panels are, from left to right, spring, summer, and fall.

## 3.6 Discussion

In settings where the data have a two-dimensional structure, the separable correlation model is a natural alternative to consider; it occupies a middle ground between separable and unstructured covariance.

Ours is the first algorithm for maximum likelihood estimation specifically designed for covariance matrices with separable correlation. By updating optimization variables in blocks and using rescaling we circumvent the need for constrained optimization methods over sets of correlation matrices. Our algorithm also has closed form updates at every iteration, making it computationally convenient. Lu and Zimmerman [48] mention that the flip-flop algorithm is between 50 and 5000 times faster than a Newton–Raphson algorithm applied to their problem with separable covariance, and the difference increases with the dimension of the optimization problem. Given the similarities between our setting and theirs, we expect our algorithm to also outperform second order descent algorithms even though updating $W$ and rescaling $U$ and $V$ is more time-consuming in our model than in the separable covariance model. We also note that our algorithm can be extended to handle constraints on the correlation matrices $U$ and $V$ by simply changing the corresponding update, without having to change anything else.

Deriving analytical bounds for when minima exist and are unique in our model, similar to those in [69], is an avenue for future research. Their proof does not immediately carry over to our setting since the negative log-likelihood of our model is not geodesically convex in the sense that of the separable covariance model is.

Further years of sampling would permit formal testing of the validity of the assumed separability against an unstructured covariance matrix in our data example. Restrictions on either the spatial or temporal autocorrelation could also be incorporated in such a study, building on the work of [71], as could the inclusion of habitat predictors and an analysis of how the assumed covariance structure affects inference about the regression coefficient $\mathcal{B}$.

# Chapter 4

# Convergence complexity analysis of a collapsed Gibbs sampler for Bayesian vector autoregressions

## 4.1 Introduction

Markov chain Monte Carlo (MCMC) is frequently used in Bayesian statistics to explore the posterior distribution of a parameter $\theta$ given data $Y$. In order to assess or ensure the reliability of an analysis using MCMC it is essential to understand some convergence properties of the chain in use [19, 42, 74]. Here we will only consider irreducible, aperiodic, and Harris recurrent chains and focus the discussion on the rate at which they converge to their stationary distribution. To be more precise, let $\nu = \nu_Y$ denote a probability measure with density $f(\theta \mid Y)$ on a set $\Theta$, and let $P^h$ ($P \equiv P^1$) be the $h$-step transition kernel for a Markov chain with state space $\Theta$, started at a point $\theta \in \Theta$. Throughout, sets on which measures are defined are assumed to be topological and equipped with their Borel $\sigma$-algebra, with the dominating measures for densities being clear from context. In our analysis the set in question will be a subset of $\mathbb{R}^d$ for some $d \geq 1$ and the dominating measure Lebesgue measure. Now, by a chain's convergence rate we mean the rate at which $\|P^h(\theta, \cdot) - \nu(\cdot)\|_{TV}$ approaches zero as $h$ tends to infinity, where $\|\cdot\|_{TV}$ denotes the total variation norm. If this convergence happens at a geometric (or exponential) rate, meaning there exist a $\rho \in [0, 1)$

and an $M : \Theta \to [0, \infty)$ such that for every $\theta \in \Theta$ and $h \in \{1, 2, \dots\}$

$$\|P^h(\theta, \cdot) - \nu(\cdot)\|_{TV} \leq M(\theta)\rho^h, \tag{4.1}$$

then the chain, or the kernel $P$, is said to be geometrically ergodic. Following Qin and Hobert [60] we also define the geometric convergence rate $\rho^\star$ to be the infimum of the set of $\rho \in [0, 1]$ such that (4.1) holds. Since all probability measures have unit total variation norm, $\rho^\star$ is always in $[0, 1]$, and $P$ is geometrically ergodic if and only if $\rho^\star < 1$. Geometric ergodicity plays an important role in the theory of Markov chains as well as in the MCMC literature. Functionals of geometrically ergodic Markov chains satisfy a central limit theorem (CLT) under relatively weak additional conditions [40], and under additional moment conditions the variance in the limiting distribution given by the CLT can be consistently estimated [20, 75]. Numerous articles have been written that establishes geometric ergodicity or similar convergence results for MCMC algorithms used for Bayesian models [e.g. 1, 3, 33, 38, 60, 73]. Lately, there has also been an increasing interest in convergence complexity analysis of MCMC algorithms, the goal of which is an understanding of how the convergence rate $\rho^\star$ behaves as the number of parameters or the number of observations in the underlying model varies [60, 82]. Here, that model will be a vector autoregression (VAR). We will consider a fixed number of parameters and vary the sample size. We establish conditions that ensure $\rho^\star < 1$, both for fixed $n$ and as $n$ tends to infinity. To facilitate an introductory discussion of these results, let us first define the VAR.

A stochastic process $Y_t \in \mathbb{R}^r$, $t = 1, \dots, n$, is a VAR of order $q$ (VAR($q$)) if it satisfies

$$Y_t = \sum_{i=1}^{q} \mathcal{A}_i^\mathsf{T} Y_{t-i} + \mathcal{B}^\mathsf{T} x_t + \varepsilon_t, \tag{4.2}$$

where $\varepsilon_1, \dots, \varepsilon_n$ are i.i.d. $\mathcal{N}(0, \Sigma)$, $x_t \in \mathbb{R}^p$ is a vector of non-stochastic predictors, $\mathcal{A}_i \in \mathbb{R}^{r \times r}$ $(i = 1, \dots, q)$, and $\mathcal{B} \in \mathbb{R}^{p \times r}$. We assume that the starting point $(Y_{-q+1}, \dots, Y_0)$ is non-stochastic. To define prior distributions, let $m \in \mathbb{R}^{qr^2}$, $C \in \mathbb{S}_+^{qr^2}$, $D \in \mathbb{S}_+^r$, and $a \geq 0$ be hyperparameters, where $\mathbb{S}_+^r$ denotes the set of (real) $r \times r$ symmetric positive semi-definite

(SPSD) matrices; we use $\mathbb{S}^r_{++}$ to denote the symmetric positive definite (SPD) ones. Let also $\mathcal{A} = [\mathcal{A}_1^\mathsf{T}, \ldots, \mathcal{A}_q^\mathsf{T}]^\mathsf{T} \in \mathbb{R}^{qr \times r}$ and $\alpha = \mathrm{vec}(\mathcal{A})$, where $\mathrm{vec}(\cdot)$ is the vectorization operator. The vectorization $\alpha$ will often, with some abuse of notation, be used interchangeably with $\mathcal{A}$ as a function argument. We consider priors on $\theta = (\alpha, \mathcal{B}, \Sigma) \in \Theta = \mathbb{R}^{qr^2} \times \mathbb{R}^{p \times r} \times \mathbb{S}^r_{++}$ that have densities in the form $f(\theta) = f(\alpha) f(\mathcal{B}) f(\Sigma)$, with

$$
\begin{aligned}
f(\alpha) &\propto \exp\left(-\frac{1}{2}[\alpha - m]^\mathsf{T} C [\alpha - m]\right), \\
f(\mathcal{B}) &\propto 1,
\end{aligned}
$$

and

$$
f(\Sigma) \propto |\Sigma|^{-a/2} \operatorname{etr}\left(-\frac{1}{2} D \Sigma^{-1}\right) I_{\mathbb{S}^r_{++}}(\Sigma),
$$

where $\operatorname{etr}(\cdot) = \exp(\operatorname{tr}(\cdot))$ and $|\cdot|$ means the determinant when applied to matrices. The flat prior on $\mathcal{B}$ is standard in multivariate scale and location problems, including in particular the multivariate regression model which is recovered when $\mathcal{A} = 0$ in our model. The priors on $\alpha$ and $\Sigma$ are common in macroeconomics [44]. Sometimes $\mathcal{A}$ and $\mathcal{B}$ are grouped as $\Psi = [\mathcal{A}^\mathsf{T}, \mathcal{B}^\mathsf{T}]^\mathsf{T}$ and a proper multivariate normal prior is assigned to $\mathrm{vec}(\Psi)$ [46]. Here, however, we treat $\mathcal{A}$ and $\mathcal{B}$ separately since doing so allows for the use of the standard improper prior on $\mathcal{B}$, while using a proper prior on $\mathcal{A}$ leads to a proper posterior even when $n$ is small in comparison to $r$ and $q$. Consequently, many high-dimensional, or large, VARs considered in the literature are compatible with our assumptions [4, 26, 45]. Finally, the prior on $\Sigma$ includes the inverse Wishart ($D \in \mathbb{S}^r_{++}, a > 2r$) and Jeffreys prior ($D = 0, a = r + 1$) as a special cases.

The literature on convergence properties of MCMC algorithms for Bayesian vector autoregressions is limited. Hobert et al. [33] analyze the convergence of a MCMC algorithm for a multivariate linear regression model that, considering the data as fixed, includes our VAR as a special case. However, they consider the (improper) prior $f(\theta) \propto |\Sigma|^{-a}$ which is not compatible with the large VARs we allow for in the fixed $n$ setting. In the large $n$

setting, i.e. when doing convergence complexity analysis, the data is no longer considered fixed and, hence, the VAR is no longer a special case of the multivariate linear regression. Kadiyala and Karlsson [43] investigate numerical properties of a two-component ($\mathcal{A}$ and $\Sigma$) Gibbs sampler for Bayesian vector autoregressions without predictors. Their analysis is simulation-based and as such does not provide any theoretical guarantees. Our results address this since, as we will discuss in more detail below, the algorithm we propose simplifies to the Gibbs sampler they consider when there are no predictors.

The algorithm we propose (Section 4.2) is a collapsed three-component Gibbs sampler that updates $\mathcal{A}$, $\mathcal{B}$, and $\Sigma$ in separate steps. Our investigation of its properties can be divided into two parts—one in which $n$ is fixed and the data conditioned on (Section 4.3), as is historically more common in the MCMC literature, and one in which $n$ tends to infinity (Section 4.4). In the former case the data can be treated as fixed but in the latter case the stochastic properties of the data are of critical importance. To appreciate the differences between the two settings, suppose first that $n$ is fixed and that having observed $Y = [Y_1, \ldots, Y_n]^\mathsf{T} \in \mathbb{R}^{n \times r}$ we want to explore the posterior density $f(\theta \mid Y)$. Let $P_n$ denote the kernel of the Markov chain generated by our collapsed Gibbs sampler and recall that the Markov chain is geometrically ergodic if and only if its geometric convergence rate $\rho^\star = \rho_n^\star$ is less than one. Geometric ergodicity is often established by using the celebrated drift and minorization conditions due to Rosenthal [65] to get an upper bound $\bar{\rho}_n$ such that $\rho_n^\star \leq \bar{\rho}_n < 1$. In many settings, including ours, the bound $\bar{\rho}_n$ one obtains from the drift and minorization conditions depends on $Y$. This is only of minor interest when $n$ and $Y$ are treated as fixed. However, as $n$ varies it is unnatural to consider $Y$ as fixed. Thus, in convergence complexity analysis the stochastic properties of $\bar{\rho}_n$ as a function of $Y$ are of primary interest. Our analysis leads to conditions that ensure $\limsup_{n\to\infty} \hat{\rho}_n < 1$ almost surely or $\mathsf{P}(\hat{\rho}_n < 1) \to 1$ as $n \to \infty$; when one of these holds we say that the geometric ergodicity is asymptotically stable almost surely or in probability. Such results have previously been established only for a few practically relevant MCMC algorithms [60].

The rest of the paper is organized as follows. In Section 4.2 we propose a collapsed Gibbs sampler. Conditions for geometric ergodicity for a fixed $n$ are presented in Section

4.3 and conditions for asymptotically stable geometric ergodicity are given in Section 4.4. Some concluding remarks are given in Section 4.5.

## 4.2 A collapsed Gibbs sampler

Let $Y = [Y_1, \ldots, Y_n]^\mathsf{T} \in \mathbb{R}^{n \times r}$, $X = [x_1, \ldots, x_n]^\mathsf{T} \in \mathbb{R}^{n \times p}$, $Z_t = [Y_{t-1}^\mathsf{T}, \ldots, Y_{t-q}^\mathsf{T}]^\mathsf{T} \in \mathbb{R}^{qr}$ $(t = 1, \ldots, n)$, and $Z = [Z_1, \ldots, Z_n]^\mathsf{T} \in \mathbb{R}^{n \times qr}$. The joint density for a sample path of length $n$ from the VAR($q$) is

$$f(Y \mid \mathcal{A}, \mathcal{B}, \Sigma) \propto |\Sigma|^{-n/2} \operatorname{etr}\left[-\frac{n}{2}\Sigma^{-1} S(\mathcal{A}, \mathcal{B})\right], \tag{4.3}$$

where $S = n^{-1}(Y - Z\mathcal{A} - X\mathcal{B})^\mathsf{T}(Y - Z\mathcal{A} - X\mathcal{B})$ [49]. This is the same density as that for $n$ observations in the classical multivariate linear regression with design matrix $[Z, X] \in \mathbb{R}^{n \times (qr+p)}$ and coefficient matrix $[\mathcal{A}^\mathsf{T}, \mathcal{B}^\mathsf{T}]^\mathsf{T}$. Thus, those of our results that hold for fixed $n$, with the data treated as observed, can be applied without change to a multivariate regression model where the design matrix is partitioned into two parts—one which has a flat prior for its coefficient, and one which has a proper prior for its coefficient. Partitioning the design matrix in this way leads to a proper posterior even when $Z$ is a wide matrix, or high-dimensional. This configuration is unlike those in other work on similar models which typically assume either that $[Z, X]$ has full rank or that the prior for $[\mathcal{A}^\mathsf{T}, \mathcal{B}^\mathsf{T}]^\mathsf{T}$ is proper [1, 3, 26, 33, 72].

Straightforward calculations show that, assuming it exists, the posterior distribution in our model has density

$$f(\mathcal{A}, \mathcal{B}, \Sigma \mid Y) \propto |\Sigma|^{-\frac{n+a}{2}} \operatorname{etr}\left(-\frac{1}{2}\Sigma^{-1}[D + nS] - \frac{1}{2}(\alpha - m)^\mathsf{T} C (\alpha - m)\right). \tag{4.4}$$

If $C = 0$, then one can show that this posterior is a normal-(inverse )Wishart for which no MCMC is necessary. For the case $C \in \mathbb{S}_{++}^{qr^2}$ but $\mathcal{B} = 0$, i.e. there are no predictors in the model, Karlsson [44] suggests using a two-component Gibbs sampler to explore the posterior. Our algorithm specializes to this two-component Gibbs sampler when $\mathcal{B} = 0$ and,

as a consequence, our results apply almost verbatim to that sampler.

The following lemma gives two different sets of conditions that lead to a proper posterior. For this result, the prior on $\alpha$ need not be normal or flat—any priors satisfying the conditions laid out in the lemma work.

**Lemma 4.2.1.** *The posterior distribution is proper if either*

1. $D \in \mathbb{S}^r_{++}$, $X$ *has full column rank,* $n + a > 2r + p$, *and* $f(\alpha)$ *is proper; or*

2. $[Y, Z, X] \in \mathbb{R}^{n\times(r+qr+p)}$ *has full column rank,* $n+a > (2+q)r+p$, *and* $f(\alpha)$ *is bounded.*

*Proof.* Appendix C. □

Notice that if $f(\Sigma)$ is a proper inverse Wishart density, then necessarily $a > 2r$ and $D \in \mathbb{S}^r_{++}$, and hence the first set of conditions in Lemma 4.2.1 holds if $n > p$ and $f(\alpha)$ is proper, assuming no superfluous predictors are included in $X$. In particular, $r$ or $q$ can be arbitrarily large in comparison to $n$. The second set of conditions allows for the use of improper priors also on $\alpha$ and $\Sigma$ when $n$ is large in comparison to all of $r$, $q$, and $p$. The full column rank of $[Y, Z, X]$ is natural in large $n$ settings. In practice, one expects it to hold unless $X$ includes superfluous predictors or the sample is so small that a least squares regression of $Y$ on $Z$ and $X$ gives residuals that are identically zero. In what follows we assume the posterior is proper unless otherwise noted.

There are many MCMC algorithms that can be used to explore the posterior in (4.4). For example, all full conditional distributions have familiar forms so it is straightforward to implement a three-component Gibbs sampler. Another sensible option mentioned in the introduction is to group $\mathcal{A}$ and $\mathcal{B}$ and update them together. Here, we will instead make use of the particular structure the partitioned matrix $[Z, X]$ offers and devise a collapsed Gibbs sampler. The results of Liu et al. [47] imply that the algorithm we propose generates a chain that converges to its stationary distribution at least as fast as both the three-component Gibbs sampler and the two-component sampler that groups $\mathcal{A}$ and $\mathcal{B}$. Moreover, as we will see in the next section, considering the collapsed sampler lets us work with a convenient transition kernel when establishing convergence rates. A formal description of our algorithm

is given in Algorithm 4.1. We next derive the conditional distributions necessary for its implementation.

**Algorithm 4.1** Collapsed Gibbs sampler

1: *Input:* Starting values $(\alpha^0, \mathcal{B}^0, \Sigma^0)$, $k = 0$, $K \geq 1$
2: **while** : $k < K$ **do**
3: Draw $\Sigma^{k+1}$ from the distribution of $\Sigma \mid \mathcal{A}^k, Y$
4: Draw $\alpha^{k+1}$ from the distribution of $\alpha \mid \Sigma^{k+1}, Y$
5: Draw $\mathcal{B}^{k+1}$ from the distribution of $\mathcal{B} \mid \mathcal{A}^{k+1}, \Sigma^{k+1}, Y$
6: Set $k = k + 1$
7: **end while**

Let $\mathcal{M}(M, U, V)$ denote the matrix normal distribution with mean $M$ and scale matrices $U$ and $V$ (see Definition C.1.1), and let $\mathcal{W}^{-1}(U, c)$ denote the inverse Wishart distribution with scale matrix $U$ and $c$ degrees of freedom. For any real matrix $F$, define $P_F$ to be the projection onto its column space and $Q_F$ the projection onto the orthogonal complement of its column space. Let also $\otimes$ denote the Kronecker product and define $B = B(\Sigma) = C + \Sigma^{-1} \otimes Z^\mathsf{T} Q_X Z$ and $u = u(\Sigma) = B^{-1}[Cm + (\Sigma^{-1} \otimes Z^\mathsf{T} Q_X) \operatorname{vec}(Q_X Y)]$.

**Lemma 4.2.2.** *If at least one of the two sets of conditions in Lemma 4.2.1 holds, then*

$$\Sigma \mid \mathcal{A}, Y \sim \mathcal{W}^{-1}\left(D + (Y - Z\mathcal{A})^\mathsf{T} Q_X (Y - Z\mathcal{A}), n + a - p - r - 1\right)$$
$$\alpha \mid \Sigma, Y \sim \mathcal{N}(u, B^{-1}), \text{ and}$$
$$\mathcal{B} \mid \mathcal{A}, \Sigma, Y \sim \mathcal{M}\left([X^\mathsf{T} X]^{-1} X^\mathsf{T} (Y - Z\mathcal{A}), [X^\mathsf{T} X]^{-1}, \Sigma\right).$$

*Proof.* Appendix C. □

It is useful for the coming discussion to notice from the proofs of Lemmas 4.2.1 and 4.2.2 that $f(\alpha, \Sigma \mid Y)$ is in the form of a posterior density in a VAR without predictors, but

with the data $Y$ and $Z$ replaced by $Q_X Y$ and $Q_X Z$:

$$f(\alpha, \Sigma \mid Y) \propto |\Sigma|^{-\frac{a+n-p}{2}} \operatorname{etr}\left(-\frac{1}{2}\Sigma^{-1}[D + (Y - Z\mathcal{A})^{\mathsf{T}} Q_X (Y - Z\mathcal{A})]\right) f(\alpha). \tag{4.5}$$

In the next section we prove that the proposed algorithm generates a geometrically ergodic Markov chain.

## 4.3 Geometric ergodicity

For the work in this section $n$ is fixed and the data $Y$ and $X$ observed, and hence treated as constant. Accordingly, we do not use a subscript for the sample size on the transition kernels. The one-step transition kernel for the collapsed Gibbs sampler in Algorithm 4.1 is, for any measurable $A \subseteq \Theta = \mathbb{R}^{qr^2} \times \mathbb{R}^{p\times r} \times \mathbb{S}^r_{++}$,

$$P_C(\theta', A) = \iiint I_A(\alpha, \mathcal{B}, \Sigma) f(\Sigma \mid \alpha', Y) f(\alpha \mid \Sigma, Y) f(\mathcal{B} \mid \alpha, \Sigma, Y)\, \mathrm{d}\Sigma\, \mathrm{d}\alpha\, \mathrm{d}\mathcal{B},$$

where the subscript $C$ is short for collapsed. By construction, the invariant distribution $\nu_C$ has density $f(\alpha, \mathcal{B}, \Sigma \mid Y)$. Instead of working directly with $P_C$ we will use its structure to reduce the problem in a convenient way. To that end, consider the sequence $\xi^k = (\alpha^k, \Sigma^k)$, $k = 1, 2, \ldots$ obtained by ignoring the component for $\mathcal{B}$ in the chain generated by Algorithm 4.1. The sequence $\xi^k$ is generated as a two-component Gibbs sampler exploring $f(\alpha, \Sigma \mid Y)$, and hence its transition kernel is, for any measurable $A \subseteq \mathbb{R}^{qr^2} \times \mathbb{S}^r_{++}$,

$$P_G((\alpha', \Sigma'), A) = \iint I_A(\alpha, \Sigma) f(\alpha \mid \Sigma, Y) f(\Sigma \mid \alpha', Y)\, \mathrm{d}\alpha\, \mathrm{d}\Sigma,$$

and its invariant distribution $\nu_G$ has density $f(\alpha, \Sigma \mid Y)$, where subscript $G$ stands for Gibbs.

The following lemma says that we can analyze $P_G$ in place of $P_C$, thereby reducing the complexity of the problem. Its proof, which for completeness can be found in Appendix C, uses only well known results about de-initializing Markov chains due to Roberts and

Rosenthal [62].

**Lemma 4.3.1.** *For any $\theta = (\alpha, \mathcal{B}, \Sigma) \in \Theta$, and $h \in \{1, 2, \dots\}$,*

$$\|P_C^h(\theta, \cdot) - \nu_C(\cdot)\|_{TV} = \|P_G^h((\alpha, \Sigma), \cdot) - \nu_G(\cdot)\|_{TV}$$

*Proof.* Appendix C. □

We next present some preliminary results that will lead geometric ergodicity of $P_C$ and $P_G$. The following well known result due to Rosenthal [65], here simplified and specialized to our setting using a calculation in Qin and Hobert [60], is instrumental.

**Lemma 4.3.2.** *If $P$ is a transition kernel with state space $\mathsf{X}$ and there exist $V : \mathsf{X} \to [0, \infty)$, $\lambda < 1$, $L \geq 0$, $T > 2L/(1-\lambda)$, $\varepsilon > 0$, and a measure $R$ on $\mathsf{X}$ such that for every $\mathsf{x}' \in \mathsf{X}$ and every $\mathsf{x}'' \in \mathsf{X} \cap \{\mathsf{x} \in \mathsf{X} : V(\mathsf{x}) \leq T\}$,*

$$\int V(\mathsf{x}) P(\mathsf{x}', \mathrm{d}\mathsf{x}) \leq \lambda V(\mathsf{x}') + L \tag{4.6}$$

*and*

$$P(\mathsf{x}'', \cdot) \geq \varepsilon R(\cdot), \tag{4.7}$$

*then there exists a $c \in (0, 1)$ such that the geometric convergence rate $\rho^\star$ of $P$ is upper bounded by*

$$\bar{\rho} := (1-\varepsilon)^c \vee \left(\frac{1 + 2L + \lambda T}{1 + T}\right)^{1-c} (1 + 2L + 2\lambda T)^c < 1.$$

For now it is enough to take from Lemma 4.3.2 that if the drift condition in (4.6) and the minorization condition in (4.7) are satisfied, then $P$ is geometrically ergodic—the specific form of the upper bound $\bar{\rho}$ will be useful in the next section. The next lemma verifies that the drift condition in (4.6) holds for $P_G$, which has state space $\mathsf{X} = \mathbb{R}^{qr^2} \times \mathbb{S}_{++}^r$. We use $\|\cdot\|$ for the Euclidean norm when applied to vectors and the spectral (induced) norm when applied to matrices, $\|\cdot\|_F$ denotes the Frobenius norm for matrices, and superscript $+$ denotes the Moore–Penrose pseudo-inverse.

**Lemma 4.3.3.** *If $C \in \mathbb{S}_{++}^{qr^2}$ and at least one of the two sets of conditions in Lemma 4.2.1 holds, then the transition kernel $P_G$ satisfies* (4.6) *with $V : \mathbb{R}^{qr^2} \times \mathbb{S}_{++}^{r} \to [0, \infty)$ defined by $V(\alpha, \Sigma) = \|\alpha\|^2$, $\lambda = 0$, and*

$$L = \left( \|C^{-1}\| \|Cm\| + \|C^{-1/2}\| \|C^{1/2}\| \|\hat{\mathcal{A}}\|_F \right)^2 + \operatorname{tr}(C^{-1}),$$

*where $\hat{\mathcal{A}} = (Z^\mathsf{T} Q_X Z)^+ Z^\mathsf{T} Q_X Y$.*

*Proof.* Assume without loss of generality that $Q_X = I_n$, where $I_n$ denotes the $n \times n$ identity matrix; the general case is recovered by replacing $Z$ and $Y$ by $Q_X Z$ and $Q_X Y$ everywhere. Let also $y = \operatorname{vec}(Y)$. By Lemma 4.2.2 and standard results for the multivariate normal distribution,

$$\int \|\alpha\|^2 f(\alpha \mid \Sigma, Y) \, \mathrm{d}\alpha = \|u\|^2 + \operatorname{tr}(B^{-1}).$$

Since $\Sigma^{-1} \otimes Z^\mathsf{T} Z$ is SPSD we have $\operatorname{tr}(B^{-1}) \leq \operatorname{tr}(C^{-1})$, so it remains only to deal with the first term. The triangle inequality gives $\|u\| \leq \|B^{-1} C m\| + \|B^{-1}(\Sigma^{-1} \otimes Z^\mathsf{T}) y\|$. Again using that $\Sigma^{-1} \otimes Z^\mathsf{T} Z$ is SPSD,

$$\|B^{-1} C m\| \leq \|C^{-1}\| \|Cm\|.$$

For the remaining term we have

$$\begin{aligned} \|B^{-1}(\Sigma^{-1} \otimes Z^\mathsf{T}) y\| &= \|C^{-1/2}(I_{qr^2} + C^{-1/2}(\Sigma^{-1} \otimes Z^\mathsf{T} Z) C^{-1/2})^{-1} C^{-1/2}(\Sigma^{-1} \otimes Z^\mathsf{T}) y\| \\ &\leq \|C^{-1/2}\| \|(I_{qr^2} + C^{-1/2}(\Sigma^{-1} \otimes Z^\mathsf{T} Z) C^{-1/2})^{-1} C^{-1/2}(\Sigma^{-1} \otimes Z^\mathsf{T}) y\| \end{aligned}$$

By Lemma C.1.2, with $(\Sigma^{-1/2} \otimes I_n) y$ and $(\Sigma^{-1/2} \otimes Z) C^{-1/2}$ taking the roles of what is there denoted $y$ and $X$, we have for any generalized inverse (denoted by superscript $g$) that

$$\|(I_{qr^2} + C^{-1/2}(\Sigma^{-1} \otimes Z^\mathsf{T} Z) C^{-1/2})^{-1} C^{-1/2}(\Sigma^{-1} \otimes Z^\mathsf{T}) y\|$$

is upper bounded by

$$\|(C^{-1/2}(\Sigma^{-1} \otimes Z^\mathsf{T} Z)C^{-1/2})^g C^{-1/2}(\Sigma^{-1} \otimes Z^\mathsf{T})y\|. \tag{4.8}$$

Lemma C.1.3 says that $C^{1/2}(\Sigma^{-1} \otimes Z^\mathsf{T} Z)^+ C^{1/2}$ is one such generalized inverse. Using that the Moore–Penrose pseudo-inverse distributes over the Kronecker product [50], the middle part of this generalized inverse can be written as $(\Sigma^{-1} \otimes Z^\mathsf{T} Z)^+ = \Sigma \otimes (Z^\mathsf{T} Z)^+$. Thus, for this particular choice of generalized inverse (4.8) is equal to

$$\|C^{1/2}(\Sigma \otimes [Z^\mathsf{T} Z]^+)(\Sigma^{-1} \otimes Z^\mathsf{T})y\| = \|C^{1/2}(I_r \otimes [Z^\mathsf{T} Z]^+ Z^\mathsf{T})y\|,$$

which is upper bounded by

$$\|C^{1/2}\|\|(I_r \otimes [Z^\mathsf{T} Z]^+ Z^\mathsf{T})y\|.$$

Finally, notice that $\|(Z^\mathsf{T} Z)^+ Z^\mathsf{T} Y\|_F = \|\operatorname{vec}((Z^\mathsf{T} Z)^+ Z^\mathsf{T} Y)\| = \|(I_r \otimes (Z^\mathsf{T} Z)^+ Z^\mathsf{T})y\|$. Thus, we have proven that, for every $\Sigma$,

$$\|u\|^2 + \operatorname{tr}(B^{-1}) \leq \left(\|C^{-1}\|\|Cm\| + \|C^{-1/2}\|\|C^{1/2}\|\|(Z^\mathsf{T} Z)^+ Z^\mathsf{T} Y\|_F\right)^2 + \operatorname{tr}(C^{-1}),$$

and the proof is completed upon integrating both sides with respect to $f(\Sigma \mid \alpha', Y)\,\mathrm{d}\Sigma$. □

**Lemma 4.3.4.** *For any $T > 0$, there exists a probability measure $R$ such that, whenever $V(\alpha, \Sigma) \leq T$,*

$$P_G((\alpha, \Sigma), \cdot) \geq \varepsilon R(\cdot)$$

*with*

$$\varepsilon = \frac{|D + Y^\mathsf{T} Q_{[Z,X]} Y|^{(n+a-p-r-1)/2}}{|D + I_r(\|Q_X Y\| + \|Q_X Z\|\sqrt{T})^2|^{(n+a-p-r-1)/2}}.$$

*Proof.* Let $c = n + a - p - r - 1$ and consider $f(\Sigma \mid \mathcal{A}, Y)$ on sets where $V(\alpha) = \|\mathcal{A}\|_F^2 \leq T$.

For such $\mathcal{A}$,

$$
\begin{aligned}
&f(\Sigma \mid \mathcal{A}, Y) \\
&= \frac{|D + (Y - Z\mathcal{A})Q_X(Y - Z\mathcal{A})|^{c/2}}{2^{cr/2}\Gamma_r(c/2)}|\Sigma|^{-\frac{n+a-p}{2}} \operatorname{etr}\left(-\frac{1}{2}\Sigma^{-1}[D + (Y - Z\mathcal{A})Q_X(Y - Z\mathcal{A})]\right) \\
&\geq \frac{|D + Y^{\mathsf{T}}Q_{[Z,X]}Y|}{2^{cr/2}\Gamma_r(c/2)}|\Sigma|^{-\frac{n+a-p}{2}} \operatorname{etr}\left(-\frac{1}{2}\Sigma^{-1}[D + I_r(\|Q_XY\| + \|Q_XZ\|\sqrt{T})^2]\right) \\
&=: g(\Sigma),
\end{aligned}
$$

where $\Gamma_r$ denotes the $r$-variate gamma function. The inequality uses two bounds: $\operatorname{tr}(\Sigma^{-1}(Y - Z\mathcal{A})^{\mathsf{T}}Q_X(Y - Z\mathcal{A})) = \operatorname{tr}(\Sigma^{-1/2}(Y - Z\mathcal{A})^{\mathsf{T}}Q_X(Y - Z\mathcal{A})\Sigma^{-1/2}) \leq \operatorname{tr}(\Sigma^{-1/2}I_r\|Q_X(Y - Z\mathcal{A})\|^2\Sigma^{-1/2})$ and $(Y - Z\mathcal{A})^{\mathsf{T}}Q_X(Y - Z\mathcal{A}) - (Y - Z\mathcal{A})^{\mathsf{T}}Q_{[Z,X]}(Y - Z\mathcal{A})$ is SPSD. Now, upon defining the measure $\tilde{R}$ by

$$\tilde{R}(A) = \iint I_A(\alpha, \Sigma)f(\alpha \mid \Sigma, Y)g(\Sigma)\,\mathrm{d}\alpha\,\mathrm{d}\Sigma$$

and letting $\varepsilon = \tilde{R}(\mathbb{R}^{qr^2} \times \mathbb{S}^r_{++}) = |D+Y^{\mathsf{T}}Q_{[Z,X]}Y|^{c/2}|D+I_r(\|Q_XY\|+\|Q_XZ\|\sqrt{T})^2|^{-c/2} > 0$ and $R = \tilde{R}/\varepsilon$, we are done. Indeed, $\varepsilon > 0$ since under the first set of conditions in Lemma 4.2.1 $D$ is SPD, and under the second set of conditions $Y^{\mathsf{T}}Q_{[Z,X]}Y$ is SPD by Lemma C.1.1. □

**Theorem 4.3.5.** *If $C \in \mathbb{S}^{qr^2}_{++}$ and at least one of the two sets of conditions in Lemma 4.2.1 holds, then the transition kernels $P_C$ and $P_G$ are both geometrically ergodic.*

*Proof.* By Lemma 4.3.1 it suffices to show it for $P_G$. Lemma 4.3.3 establishes that (4.6) holds for $P_G$ with $V(\alpha, \Sigma) = \|\alpha\|^2$ and $\lambda = 0$ and Lemma 4.3.4 verifies (4.7). The result now follows from Lemma 4.3.2. □

## 4.4 Asymptotic stability

The geometric ergodicity in Theorem 4.3.5 holds for any fixed $n$. In this section we establish asymptotically stable geometric ergodicity as $n \to \infty$. Formally, we say that the geometric

ergodicity of a sequence of transition kernels $P_1, P_2, \ldots$ with geometric convergence rates $\rho_1^\star, \rho_2^\star, \ldots$ is asymptotically stable almost surely if there exists a sequence of random variables $\bar{\rho}_1, \bar{\rho}_2, \ldots$ such that $\rho_n^\star \leq \bar{\rho}_n$ for every $n$ and $\limsup_{n\to\infty} \bar{\rho}_n < 1$ almost surely. If instead $\mathsf{P}(\bar{\rho}_n < 1) \to 1$, we say that the geometric ergodicity is asymptotically stable in probability. Here, the sequence of kernels under consideration is $P_{G,1}, P_{G,2}, \ldots$, where $P_{G,n}$ is the kernel $P_G$ from the previous section with the dependence on the sample size $n$ made explicit; we will continue to use $P_G$ when $n$ is arbitrary but fixed.

It is clear that as $n$ changes so do the data $Y$ and $X$. Hence, treating $Y$ as fixed (observed) is not appropriate unless we only want to discuss asymptotic properties holding pointwise, i.e. for particular paths of the stochastic process $Y_1, Y_2, \ldots$, which is unnecessarily restrictive. Let $(\Omega, \mathsf{P}, \mathcal{F})$ denote an underlying probability space on which $Y_1, Y_2, \ldots$ are defined. The joint distribution of the whole stochastic process $(Y_1, Y_2, \ldots)$, which is characterized by the true parameters $\mathcal{A}^\star$, $\mathcal{B}^\star$, and $\Sigma^\star$, and the non-stochastic sequence $x_1, x_2, \ldots$, is a push-forward measure of $\mathsf{P}$. In what follows, probabilistic statements made without specifying a measure are taken to be with respect $\mathsf{P}$ so that "almost surely" (a.s.) can often also be read as "for almost all sample paths of the VAR".

Like in the fixed $n$ setting, Lemma 4.3.2 is instrumental to our strategy: if $P_{G,n}$ satisfies Lemma 4.3.2 with some $V = V_n$, $\lambda = \lambda_n$, $L = L_n$, $\varepsilon = \varepsilon_n$, and $T = T_n$, then there exists a $\bar{\rho}_n < 1$ that upper bounds $\rho_n^\star$. Our exposition focuses on the properties of those $\bar{\rho}_n$, $n = 1, 2, \ldots$, as $n$ tends to infinity. Throughout the section we assume that the priors, and in particular the hyperparameters, are the same for every $n$.

Clearly, the choice of drift function $V_n$ is important for the upper bound $\hat{\rho}_n$ one obtains. The drift function used for the fixed $n$ regime is not well suited for the asymptotic analysis in this section. For example, as our next result shows, the sequence of $\varepsilon = \varepsilon_n$ given in the proof of Theorem 4.3.5 converges to zero almost surely as $n \to \infty$ for many reasonable configurations of the VAR, and hence the corresponding upper bounds satisfy $\lim_{n\to\infty} \bar{\rho}_n = 1$ almost surely.

**Proposition 4.4.1.** *If, almost surely as $n \to \infty$,*

$$\|Q_{[X,Z]}Y\|^2/\|Q_XZ\|^2 = o(n)$$

*then the $\varepsilon = \varepsilon_n$ in Theorem 4.3.5 tends to zero almost surely as $n \to \infty$. In particular, $\varepsilon_n \to 0$ a.s. if $n^{-1}Y^\mathsf{T}Q_{[Z,X]}Y$ and $n^{-1}Z^\mathsf{T}Q_XZ$ have positive definite limits a.s.*

*Proof.* It suffices to show that $\zeta_n := \varepsilon_n^{2n/c} \to 0$ a.s. since $2n/c \to 2$. We have

$$\zeta_n = \left[\frac{|D + Y^\mathsf{T}Q_{[Z,X]}Y|}{|D + I_r(\|Q_XY\| + \|Q_XZ\|\sqrt{T})^2|}\right]^n.$$

Let $\kappa_1 \geq \kappa_2 \geq \cdots \geq \kappa_r$ denote the eigenvalues of $D$. Expanding the square and using that $\|Y^\mathsf{T}Q_XY\| \geq \|Y^\mathsf{T}Q_{[Z,X]}Y\|$, we have

$$\begin{aligned}
\zeta_n &\leq \left[\frac{|D + I_r\|Q_{[Z,X]}Y\|^2|}{|D + I_r\|Q_XY\|^2 + I_r\|Q_XZ\|^2T|}\right]^n \\
&= \left[\prod_{i=1}^r \frac{\kappa_i + \|Q_{[Z,X]}Y\|^2}{\kappa_i + \|Q_XY\|^2 + \|Q_XZ\|^2T}\right]^n \\
&\leq \prod_{i=1}^r \left[1 + \|Q_XZ\|^2T/(\kappa_i + \|Y^\mathsf{T}Q_{[Z,X]}Y\|^2)\right]^{-n}.
\end{aligned}$$

Since $T > 2L > 2\|C^{-1}\|^2\|Cm\|^2 > 0$ and $\kappa_i < \infty$ for all $i$ and independently of $n$, the first and main conclusion follows from that $(1 + b/n)^{-n} \to e^{-b}$ for any $b \in \mathbb{R}$ as $n \to \infty$. The second conclusion follows from the first upon noticing that if $n^{-1}Y^\mathsf{T}Q_{[Z,X]}Y$ and $n^{-1}Z^\mathsf{T}Q_XZ$ have positive definite limits a.s., then $\|Q_{[X,Z]}Y\|^2/\|Q_XZ\|^2 = O(1)$. □

The intuition as to why the drift function that works in the fixed $n$ regime is not suitable for convergence complexity analysis is provided by Qin and Hobert [60]: they argue that the drift function should be centered (minimized) at a point the chain in question can be expected to visit often. The function defined by $V(\alpha, \Sigma) = \|\alpha\|^2$ is minimized in any point were $\alpha = 0$, but there is in general no reason to believe the $\alpha$-component of the chain will visit a neighborhood of the origin often. On the other hand, if the number of observations

grows fast enough in comparison to other quantities, then we expect the posterior density $f(\mathcal{A} \mid Y)$ to concentrate around $\mathcal{A}^\star$. We also expect that for large $n$ the least squares and maximum likelihood estimator $\hat{\mathcal{A}} = (Z^\mathsf{T} Q_X Z)^+ Z^\mathsf{T} Q_X Y$ is close to $\mathcal{A}^\star$. Thus, intuitively, the $\alpha$-component of the chain should visit the vicinity of $\hat{\alpha} = \operatorname{vec}(\hat{\mathcal{A}})$ often. Formalizing this intuition leads to the main result of the section.

Let us re-define $V : \mathbb{R}^{qr^2} \times \mathbb{S}^r_{++} \to [0, \infty)$ by $V(\alpha, \Sigma) = \|Q_X Z\mathcal{A} - Q_X Z\hat{\mathcal{A}}\|_F^2 = \|(I_r \otimes Q_X Z)(\alpha - \hat{\alpha})\|^2$. The following lemma establishes a result that will lead to verification of the drift condition in (4.9) for all large enough $n$ and almost all sample paths of the VAR under appropriate conditions. Notice, however, that the $\lambda$ given here need not be less than unity for a fixed $n$ or particular sample path of the VAR.

**Lemma 4.4.2.** *If $[Z, X]$ has full column rank, $C \in \mathbb{S}^{qr^2}_{++}$, and at least one of the two sets of conditions in Lemma 4.2.1 holds, then*

$$\int V(\alpha, \Sigma) P_G(\alpha', \mathrm{d}(\alpha, \Sigma)) = \iint V(\alpha, \Sigma) f(\alpha \mid \Sigma, Y) f(\Sigma \mid \alpha', Y) \, \mathrm{d}\alpha \, \mathrm{d}\Sigma \leq \lambda V(\alpha', \Sigma') + L,$$

*with*

$$\lambda = \frac{qr + \left( \|C\|^{1/2} \|\hat{\mathcal{A}}\|_F + \|C^{-1}\|^{1/2} \|Cm\| \right)^2}{n + a - 2r - p - 2} \quad \text{and} \quad L = \lambda \operatorname{tr}(D) + \lambda \|Q_{[Z,X]} Y\|_F^2$$

*Proof.* Suppose first that $Q_X = I_n$ and notice that by the assumptions $Z$ has full column rank, and hence $(Z^\mathsf{T} Z)^{-1}$ exists. Since $f(\alpha \mid \Sigma, Y)$ is a multivariate normal density,

$$\int V(\alpha, \Sigma) f(\alpha \mid \Sigma, Y) \, \mathrm{d}\alpha = \|(I_r \otimes Z)(u - \hat{\alpha})\|^2 + \operatorname{tr}((I_r \otimes Z) B^{-1} (I_r \otimes Z)^\mathsf{T}). \tag{4.9}$$

For the second term we have since $C$ is SPSD (even SPD) that

$$\begin{aligned}\operatorname{tr}\left((I_r \otimes Z)B^{-1}(I_r \otimes Z)^\mathsf{T}\right) &= \operatorname{tr}\left[(I_r \otimes Z)(C + \Sigma^{-1} \otimes Z^\mathsf{T}Z)^{-1}(I_r \otimes Z)^\mathsf{T}\right] \\ &\leq \operatorname{tr}\left[(I_r \otimes Z)(\Sigma^{-1} \otimes Z^\mathsf{T}Z)^{-1}(I_r \otimes Z)^\mathsf{T}\right] \\ &= \operatorname{tr}\left[\Sigma \otimes Z(Z^\mathsf{T}Z)^{-1}Z^\mathsf{T}\right] \\ &= \operatorname{tr}(\Sigma)\operatorname{tr}[Z(Z^\mathsf{T}Z)^{-1}Z^\mathsf{T}] \\ &= \operatorname{tr}(\Sigma)qr,\end{aligned}$$

where the last line uses that the trace of a projection matrix is the dimension of the space onto which it is projecting. Focusing now on the first term on the right hand side in (4.9) we have, defining $H = \Sigma^{-1} \otimes Z^\mathsf{T}Z$ and using $\hat{\alpha} = H^{-1}(\Sigma^{-1} \otimes Z^\mathsf{T})y$, that

$$\|(I_r \otimes Z)(u - \hat{\alpha})\| = \|(I_r \otimes Z)(\hat{\alpha} - B^{-1}(Cm + [\Sigma^{-1} \otimes Z^\mathsf{T}Z]y))\|$$

is upper bounded by

$$\|(I_r \otimes Z)(H^{-1} - B^{-1})(\Sigma^{-1} \otimes Z^\mathsf{T})y\| + \|(I_r \otimes Z)B^{-1}Cm\|. \tag{4.10}$$

Moreover, since $B = C + H$ the Woodbury identity gives $H^{-1} - B^{-1} = H^{-1}(C^{-1} + H^{-1})^{-1}H^{-1}$ so that the first term in (4.10) can be upper bounded as follows:

$$\begin{aligned}\|(I_r \otimes Z)(H^{-1} - B^{-1})(\Sigma^{-1} \otimes Z^\mathsf{T})y\| &= \|(I_r \otimes Z)H^{-1}(H^{-1} + C^{-1})^{-1}H^{-1}(\Sigma^{-1} \otimes Z^\mathsf{T})y\| \\ &= \|(I_r \otimes Z)H^{-1}(H^{-1} + C^{-1})^{-1}\hat{\alpha}\| \\ &\leq \|(I_r \otimes Z)H^{-1/2}\|\|H^{-1/2}(H^{-1} + C^{-1})^{-1}\|\|\hat{\alpha}\|.\end{aligned}$$

Here, the power $G^t$, $t \in \mathbb{R}$, for a SPD matrix $G$ is defined by taking the spectral decomposition $G = U_G \operatorname{diag}(\lambda_{\max}(G), \ldots, \lambda_{\min}(G))U_G^\mathsf{T}$, where $\lambda_{\max}(\cdot)$ and $\lambda_{\min}(\cdot)$ denote the largest

and smallest eigenvalues, respectively, and setting

$$G^t = U_G \operatorname{diag}(\lambda_{\max}^t(G), \ldots, \lambda_{\min}^t(G)) U_G^{\mathsf{T}}.$$

Now by standard properties of eigenvalues and eigenvectors of Kronecker products we get

$$\|(I_r \otimes Z)H^{-1/2}\| = \|(I_r \otimes Z)(\Sigma^{1/2} \otimes [Z^{\mathsf{T}}Z]^{-1/2})\| = \|\Sigma^{1/2}\|\|Z(Z^{\mathsf{T}}Z)^{-1/2}\| = \|\Sigma^{1/2}\|.$$

In addition,

$$\begin{aligned}
\|H^{-1/2}(H^{-1} + C^{-1})^{-1}\| &= \lambda_{\max}^{1/2}\left((H^{-1} + C^{-1})^{-1})H^{-1}(H^{-1} + C^{-1})^{-1}\right) \\
&\leq \lambda_{\max}^{1/2}\left((H^{-1} + C^{-1})^{-1})(H^{-1} + C^{-1})(H^{-1} + C^{-1})^{-1}\right) \\
&= \lambda_{\max}^{1/2}\left((H^{-1} + C^{-1})^{-1}\right) \\
&\leq \lambda_{\max}^{1/2}(C) \\
&= \|C\|^{1/2}.
\end{aligned}$$

It remains to deal with the second term in (4.10). Using a similar technique as with the previous term, we have

$$\begin{aligned}
\|(I_r \otimes Z)B^{-1}Cm\| &= \|(\Sigma^{1/2} \otimes I_n)(\Sigma^{-1/2} \otimes I_n)(I_r \otimes Z)B^{-1}Cm\| \\
&\leq \|\Sigma^{1/2}\|\|(\Sigma^{-1/2} \otimes Z)(C + \Sigma^{-1} \otimes Z^{\mathsf{T}}Z)^{-1}\|\|Cm\| \\
&= \|\Sigma^{1/2}\|\lambda_{\max}^{1/2}\left([C + \Sigma^{-1} \otimes Z^{\mathsf{T}}Z]^{-1}[\Sigma^{-1} \otimes Z^{\mathsf{T}}Z][C + \Sigma^{-1} \otimes Z^{\mathsf{T}}Z]^{-1}\right)\|Cm\| \\
&\leq \|\Sigma^{1/2}\|\lambda_{\max}^{1/2}([C + \Sigma^{-1} \otimes Z^{\mathsf{T}}Z]^{-1})\|Cm\| \\
&\leq \|\Sigma^{1/2}\|\|C^{-1}\|^{1/2}\|Cm\|
\end{aligned}$$

Putting things together we have shown that, for any $\Sigma$,

$$\|(I_r \otimes Z)(\hat{\alpha} - u)\| \leq \|\Sigma^{1/2}\|\left(\|C\|^{1/2}\|\hat{\alpha}\| + \|C^{-1}\|^{1/2}\|Cm\|\right),$$

and hence we get from (4.9)

$$\int V(\alpha,\Sigma) f(\alpha \mid \Sigma, Y)\,\mathrm{d}\alpha \le \|\Sigma\| \left(\|C\|^{1/2}\|\hat{\alpha}\| + \|C^{-1}\|^{1/2}\|Cm\|\right)^2 + qr\,\mathrm{tr}(\Sigma).$$

The proof for the case $Q_X = I_n$ is completed by upper bounding $\|\Sigma\| \le \mathrm{tr}(\Sigma)$, integrating both sides with respect to $f(\Sigma \mid \alpha', Y)\,\mathrm{d}\Sigma$, and noting that

$$\begin{aligned}\int \mathrm{tr}(\Sigma) f(\Sigma \mid \alpha', Y)\,\mathrm{d}\Sigma &= \frac{1}{n+a-2r-p-2}\,\mathrm{tr}\left(D + (Y - Z\mathcal{A}')^{\mathsf{T}}(Y - Z\mathcal{A}')\right) \\ &= \frac{1}{n+a-2r-p-2}\left(\mathrm{tr}(D) + \|Q_Z Y\|_F^2 + \|Z\hat{\mathcal{A}} - Z\mathcal{A}'\|_F^2\right),\end{aligned}$$

where we have used that $(Y - Z\mathcal{A}')^{\mathsf{T}}(Y - Z\mathcal{A}') = (Y - Z\mathcal{A}')^{\mathsf{T}} P_Z (Y - Z\mathcal{A}') + (Y - Z\mathcal{A}')^{\mathsf{T}} Q_Z (Y - Z\mathcal{A}')$, and that $P_Z Y = Z\hat{\mathcal{A}}$. The general case is recovered by replacing $Z$ and $Y$ by $Q_X Z$ and $Q_X Z$ everywhere and invoking Lemma C.1.1. That $Z^{\mathsf{T}} Q_X Z$ is invertible also in the general case follows from the same lemma. □

**Lemma 4.4.3.** *For any $T > 0$, there exists a probability measure $R$ such that, whenever $V(\alpha,\Sigma) \le T$,*

$$P_G((\alpha,\Sigma),\cdot) \ge \varepsilon R(\cdot),$$

*with*

$$\varepsilon = \left[1 + T\lambda_{\min}^{-1}\left(D + Y^{\mathsf{T}} Q_{[Z,X]} Y\right)\right]^{-r(n+a-p-r-1)/2}.$$

*Proof.* Assume first that $Q_X = I_n$ and let $c = n+a-r-p-1$ be the degrees of freedom in the full conditional distribution for $\Sigma$. On sets where $V(\alpha,\Sigma) = \|Z\hat{\mathcal{A}} - Z\mathcal{A}\|_F^2 \le T$ we also have $\|Z\hat{\mathcal{A}} - Z\mathcal{A}\|^2 \le T$. Thus, using that $P_Z Y = Z\hat{\mathcal{A}}$, and hence that $(Y - Z\mathcal{A})^{\mathsf{T}}(Y - Z\mathcal{A}) =$

$Y^\mathsf{T} Q_Z Y + (Z\hat{\mathcal{A}} - Z\mathcal{A})^\mathsf{T}(Z\hat{\mathcal{A}} - Z\mathcal{A})$, we get

$$
\begin{aligned}
&f(\Sigma \mid \mathcal{A}, Y) \\
&= \frac{|D + (Y - Z\mathcal{A})^\mathsf{T}(Y - Z\mathcal{A})|^{c/2}}{2^{cr/2}\Gamma_r(c/2)} |\Sigma|^{-\frac{n+a-p}{2}} \operatorname{etr}\left(-\frac{1}{2}\Sigma^{-1}[D + (Y - Z\mathcal{A})^\mathsf{T}(Y - Z\mathcal{A})]\right) \\
&\geq \frac{|D + Y^\mathsf{T} Q_Z Y|^{c/2}}{2^{cr/2}\Gamma_r(c/2)} |\Sigma|^{-\frac{n+a-p}{2}} \operatorname{etr}\left(-\frac{1}{2}\Sigma^{-1}[D + Y^\mathsf{T} Q_Z Y + I_r T]\right) \\
&=: g(\Sigma),
\end{aligned}
$$

where $\Gamma_r$ denotes the $r$-variate gamma function. Take now

$$
\tilde{\varepsilon} = \int g(\Sigma)\,\mathrm{d}\Sigma = \left(\frac{|D + Y^\mathsf{T} Q_Z Y|}{|D + Y^\mathsf{T} Q_Z Y + I_r T|}\right)^{c/2}
$$

and $R$ defined by

$$
R(A) = \iint I_A(\alpha, \Sigma) f(\alpha \mid \Sigma, Y) g(\Sigma)\,\mathrm{d}\alpha\,\mathrm{d}\Sigma/\tilde{\varepsilon}.
$$

By construction, $\tilde{\varepsilon}$ and $R$ satisfy the minorization condition. Let $\kappa_1 \geq \kappa_2 \geq \cdots \geq \kappa_r > 0$ denote the eigenvalues of $D + Y^\mathsf{T} Q_Z Y$ so that

$$
\begin{aligned}
\tilde{\varepsilon} &= \left(\prod_{i=1}^{r} \frac{\kappa_i}{\kappa_i + T}\right)^{c/2} \\
&\geq \left(\frac{1}{1 + T/\kappa_r}\right)^{rc/2} \\
&= \varepsilon.
\end{aligned}
$$

Since the minorization condition is satisfied with $\tilde{\varepsilon}$, it is satisfied with its lower bound $\varepsilon$ and that completes the proof for the case $Q_X = I_n$. The general case is recovered upon replacing $Z$ and $Y$ by $Q_X Z$ and $Q_Z Y$ everywhere and invoking Lemma C.1.1. □

We are ready to state the main result of the section.

**Theorem 4.4.4.** *If*

*(a)* $C \in \mathbb{S}_{++}^{qr^2}$,

*(b)* $[Y, Z, X]$ *has full column rank for all large enough* $n$ *almost surely,*

*(c)* $\|\hat{\alpha}\|^2 = O(1)$ *almost surely as* $n \to \infty$*, and*

*(d) there exists a random variable* $K : \Omega \to (0, \infty)$ *such that, almost surely,*

$$K^{-1} \leq \liminf_{n\to\infty} n^{-1}\lambda_{\min}(Y^{\mathsf{T}}Q_{[Z,X]}Y) \leq \limsup_{n\to\infty} n^{-1}\lambda_{\max}(Y^{\mathsf{T}}Q_{[Z,X]}Y) \leq K,$$

*then* $\limsup_{n\to\infty} \bar{\rho}_n < 1$ *almost surely.*

*Proof.* Inspecting the definition of $\bar{\rho}_n$ one sees that it suffices to show that Lemmas 4.4.2 and 4.4.3 apply and that the $\lambda = \lambda_n$, $L = L_n$, $T = T_n$, and $\varepsilon = \varepsilon_n$ they give almost surely satisfy, respectively: (i) $\limsup_{n\to\infty} \lambda_n < 1$, (ii) $\limsup_{n\to\infty} L_n < \infty$, (iii) $\limsup_{n\to\infty} T_n < \infty$, and (iv) $\liminf_{n\to\infty} \varepsilon_n > 0$. Assumption (a) and (b) ensure Lemma 4.4.2 applies and assumption (c) gives $\lambda_n = O(1/n)$, so (i) holds. That $\lambda_n = O(1/n)$ and assumption (d) give $L_n = O(1)$, i.e. (ii) holds, and hence we can pick a sequence $T_n > 2L_n(1 - \lambda_n)$, $n = 1, 2, \ldots$, such that (iii) holds. For (iv), notice that (iii), assumption (d), and that $[r(n+a-p-r-1)/2]/n \to r/2$ imply that we can find random variables $K_1, K_2 > 0$ such that, almost surely,

$$\varepsilon_n \geq (1 + K_1/n)^{-K_2 n} \to e^{-K_1 K_2} > 0,\ n \to \infty,$$

which gives (iv). □

Assumption (b) is weak in the large $n$ setting we are currently considering. In fact, if the non-stochastic design matrix $X$ has full column rank then $[Z, X]$ has full column rank almost surely for all large enough $n$ if the true covariance matrix $\Sigma^\star$ has full rank. Assumption (c) holds if, for example, the least squares estimator $\hat{\alpha}$ is strongly consistent, which it is under many common assumptions on the true parameters in the VAR [58]. Assumptions (d) holds if, for example, the MLE $n^{-1}Y^{\mathsf{T}}Q_{[Z,X]}Y$ of $\Sigma^\star$ is strongly consistent, or more generally if it converges to a positive definite matrix almost surely.

If some of the assumptions in Theorem 4.4.4 are relaxed to hold in probability, or with probability tending to one, instead of almost surely, then the conclusion can be weakened

accordingly to give the following corollary.

**Corollary 4.4.1.** *If*

*(a)* $C \in \mathbb{S}_{++}^{qr^2}$

*and, as* $n \to \infty$*,*

*(b)* $[Y, Z, X]$ *has full column rank with probability tending to one,*

*(c)* $\|\hat{\alpha}\|^2 = O_{\mathsf{P}}(1)$*, and*

*(d) there exists a constant* $K > 0$ *such that, with probability tending to one,*

$$K^{-1} \leq n^{-1}\lambda_{\min}(Y^{\mathsf{T}}Q_{[Z,X]}Y) \leq n^{-1}\lambda_{\max}(Y^{\mathsf{T}}Q_{[Z,X]}Y) \leq K,$$

*then* $\bar{\rho}_n < 1$ *with probability tending to one.*

## 4.5 Discussion

Markov chain Monte Carlo is used in a wide range of problems, including but not limited to the Bayesian settings considered here. However, the theoretical properties of algorithms used by practitioners are not always well understood. Here we have focused on the case of Bayesian vector autoregressions. This is one of the most common models in time series, and in particular in the analysis and forecasting of macroeconomic time series. The Gibbs sampler has been suggested to explore the posterior distribution of the parameters $\mathcal{A}$ and $\Sigma$ when there are no predictors [44], but there has been a lack of theoretical support for its appropriateness. We have addressed this by proposing a collapsed Gibbs sampler that handles predictors and deriving theoretical guarantees for that sampler. Since our algorithm simplifies to the usual Gibbs sampler when there are no predictors, our results also offer a peace of mind to practitioners using that sampler.

We have proven that our algorithm generates a geometrically ergodic Markov chain under reasonable assumptions (Theorem 4.3.5). This result is applicable both in classical

settings where the sample size is large (but fixed) in comparison to the number of parameters, and in large VARs where the dimension of the process or the lag length is (much) larger than the number of observations. Thus, with the algorithm we propose, Bayesian analysis of VARs can be carried out with confidence, knowing that characteristics of the posterior distribution can be reasonably estimated and that asymptotically valid confidence intervals can be constructed for those estimates, among other things [19, 42, 74]. Our asymptotic analysis, or convergence complexity analysis, indicates our algorithm should perform well in large samples; we have proven that, as the sample size tends to infinity, the geometric ergodicity of the sequence of transition kernels corresponding to our algorithm is asymptotically stable. This result is one of the first of its kind for practically relevant MCMC algorithms [60].

Avenues for future research includes convergence complexity analysis of cases where the dimension of the process or the lag length tends to infinity, either together with the sample size or for a fixed sample size. By inspecting the proof of Theorem 4.4.4 one sees that the same proof idea can work also if the dimension of the process or the lag length changes, as long as the sample size grows fast enough. However, the proof relies on formalizing the intuition that as the sample size increases, the posterior mode of the $\alpha$-chain and the least squares estimator of $\alpha$ are close—if the sample size is fixed or grows slowly in comparison to other quantities, then we do not expect this to be the case. For such settings one would likely have to use a different drift function than the one used in Theorem 4.4.4.

# References


[1] T. Abrahamsen and J. P. Hobert. Convergence analysis of block Gibbs samplers for Bayesian linear mixed models with $p > N$. *Bernoulli*, 23(1):459–478, feb 2017.

[2] T. W. Anderson. *An Introduction to Multivariate Statistical Analysis*. John Wiley & Sons Inc, 2003.

[3] G. Backlund and J. P. Hobert. A note on the convergence rate of MCMC for robust Bayesian multivariate linear regression with proper priors, 2018. `http://users.stat.ufl.edu/~jhobert/papers/robust_mult_proper.pdf`.

[4] M. Bańbura, D. Giannone, and L. Reichlin. Large Bayesian vector auto regressions. *Journal of Applied Econometrics*, 25(1):71–92, 2009.

[5] Y. Bar-Shalom. On the asymptotic properties of the maximum-likelihood estimate obtained from dependent observations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 33(1):72–77, 1971.

[6] R. Bhatia. *Matrix Analysis*. Springer New York, 2012.

[7] E. M. Bronshteyn and L. D. Ivanov. The approximation of convex sets by polyhedra. *Siberian Mathematical Journal*, 16(5):852–853, 1975.

[8] S. Brooks, A. Gelman, G. L. Jones, and X.-L. Meng, editors. *Handbook of Markov chain Monte Carlo*. Chapman & Hall / CRC, 2011.

[9] J. Burg, D. Luenberger, and D. Wenger. Estimation of structured covariance matrices. *Proceedings of the IEEE*, 70(9):963–974, 1982.

[10] H. Chen and T. E. Wehrly. Assessing correlation of clustered mixed outcomes from a multivariate generalized linear mixed model. *Statistics in Medicine*, 34(4):704–720, 2014.

[11] B. A. Coull and A. Agresti. Random effects modeling of multiple binomial responses using the multivariate binomial logit-normal distribution. *Biometrics*, 56(1):73–80, 2000.

[12] H. Cramér. *Mathematical Methods of Statistics*. Princeton Mathematical Series, vol. 9. Princeton University Press, Princeton, N. J., 1946.

[13] N. Cressie and C. K. Wikle. *Statistics for Spatio-Temporal Data*. Wiley, 2011.

[14] M. J. Crowder. Maximum likelihood estimation for dependent observations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 38(1):45–53, 1976.

[15] E. Demidenko. *Mixed models: Theory and Applications with R*. John Wiley & Sons, 2013.

[16] J. L. Doob. Probability and statistics. *Transactions of the American Mathematical Society*, 36(4):759–775, 1934.

[17] P. Dutilleul and B. Pinel-Alloul. A doubly multivariate model for statistical analysis of spatio-temporal environmental data. *Environmetrics*, 7(6):551–565, 1996.

[18] T. S. Ferguson. *A Course in Large Sample Theory*. Taylor & Francis Ltd, 1996.

[19] J. M. Flegal, M. Haran, and G. L. Jones. Markov chain Monte Carlo: Can we trust the third significant figure? *Statistical Science*, 23(2):250–260, may 2008.

[20] J. M. Flegal and G. L. Jones. Batch means and spectral variance estimators in Markov chain Monte Carlo. *The Annals of Statistics*, 38(2):1034–1070, apr 2010.

[21] G. B. Folland. *Real Analysis*. John Wiley & Sons, 1999.

[22] A. E. Gelfand and A. F. M. Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410):398–409, 1990.

[23] C. J. Geyer. Practical Markov chain Monte Carlo. *Statistical Science*, 7(4):473–483, nov 1992.

[24] C. J. Geyer. On the convergence of Monte Carlo maximum likelihood calculations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 56(1):261–274, 1994.

[25] C. J. Geyer. Introduction to Markov chain Monte Carlo. In S. Brooks, A. Gelman, G. L. Jones, and X.-L. Meng, editors, *Handbook of Markov chain Monte Carlo*. Chapman & Hall / CRC, 2011.

[26] S. Ghosh, K. Khare, and G. Michailidis. High-dimensional posterior consistency in Bayesian vector autoregressive models. *Journal of the American Statistical Association*, pages 1–14, 2018.

[27] T. Gneiting, M. G. Genton, and P. Guttorp. Geostatistical space-time models, stationarity, separability, and full symmetry. *Monographs On Statistics and Applied Probability*, 107:151, 2006.

[28] R. V. Gueorguieva and A. Agresti. A correlated probit model for joint modeling of clustered binary and continuous responses. *Journal of the American Statistical Association*, 96(455):1102–1112, 2001.

[29] B. Güven. Asymptotic properties of maximum likelihood estimation in the mixed analysis of variance model. *Statistical Papers*, 36(1):175–182, 1995.

[30] P. Hall and C. C. Heyde. *Martingale Limit Theory and Its Application*. Academic Press, 1980.

[31] D. A. Harville. *Matrix Algebra From a Statistician's Perspective*. Springer New York, 1997.

[32] R. D. Heijmans and J. R. Magnus. Consistent maximum-likelihood estimation with dependent observations. *Journal of Econometrics*, 32(2):253–285, 1986.

[33] J. P. Hobert, Y. J. Jung, K. Khare, and Q. Qin. Convergence analysis of MCMC algorithms for Bayesian multivariate linear regression with non-Gaussian errors. *Scandinavian Journal of Statistics*, jan 2018.

[34] R. D. James G. MacKinnon. *Econometric Theory and Methods*. Oxford University Press, 2003.

[35] J. Jiang. *Linear and Generalized Linear Mixed Models and Their Applications*. Springer, 2007.

[36] J. Jiang. The subset argument and consistency of MLE in GLMM: Answer to an open problem and beyond. *The Annals of Statistics*, 41(1):177–195, 2013.

[37] J. Jiang. *Asymptotic Analysis of Mixed Effects Models: Theory, Applications, and Open Problems*. Chapman and Hall/CRC, 2017.

[38] A. A. Johnson and G. L. Jones. Gibbs sampling for a Bayesian hierarchical general linear model. *Electronic Journal of Statistics*, 4(0):313–333, 2010.

[39] B. L. Johnson and K. H. Hagerty. Status and trends of selected resources in the Upper Mississippi River system. techreport, U.S. Geological Survey, Upper Midwest Environmental Sciences Center, 2008. `https://pubs.usgs.gov/mis/LTRMP2008-T002/pdf/LTRMP2008-T002_web.pdf` (accessed 18 April 2018).

[40] G. L. Jones. On the Markov chain central limit theorem. *Probability Surveys*, 1(0):299–320, 2004.

[41] G. L. Jones, M. Haran, B. S. Caffo, and R. Neath. Fixed-width output analysis for Markov chain Monte Carlo. *Journal of the American Statistical Association*, 101(476):1537–1547, 2006.

[42] G. L. Jones and J. P. Hobert. Honest exploration of intractable probability distributions via Markov chain Monte Carlo. *Statistical Science*, 16(4):312–334, nov 2001.

[43] K. R. Kadiyala and S. Karlsson. Numerical methods for estimation and inference in Bayesian VAR-models. *Journal of Applied Econometrics*, 12(2):99–132, 1997.

[44] S. Karlsson. Forecasting with Bayesian vector autoregression. In *Handbook of Economic Forecasting*, pages 791–897. Elsevier, 2013.

[45] G. M. Koop. Forecasting with medium and large Bayesian VARs. *Journal of Applied Econometrics*, 28(2):177–203, 2013.

[46] D. Korobilis. Forecasting in vector autoregressions with many predictors. In *Bayesian Econometrics*, pages 403–431. Emerald Group Publishing Limited, 2008.

[47] J. S. Liu, W. H. Wong, and A. Kong. Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika*, 81(1):27–40, mar 1994.

[48] N. Lu and D. L. Zimmerman. The likelihood ratio test for a separable covariance matrix. *Statistics & Probability Letters*, 73(4):449–457, 2005.

[49] H. Lütkepohl. *New Introduction to Multiple Time Series Analysis*. Springer Berlin Heidelberg, 2005.

[50] J. R. Magnus and H. Neudecker. *Matrix Differential Calculus with Applications in Statistics and Econometrics*. Wiley John & Sons, 2002.

[51] P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman and Hall/CRC, 1989.

[52] C. E. McCulloch. Joint modelling of mixed outcome types using latent variables. *Statistical Methods in Medical Research*, 17(1):53–73, 2008.

[53] C. E. McCulloch, S. R. Searle, and J. M. Neuhaus. *Generalized, Linear, and Mixed Models*. Wiley-Interscience, 2008.

[54] S. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability*. Cambridge University Press, 2011.

[55] J. J. Miller. Asymptotic properties of maximum likelihood estimates in the mixed model of the analysis of variance. *The Annals of Statistics*, 5(4):746–762, 1977.

[56] M. Min. *Asymptotic Normality in Generalized Linear Mixed Models*. PhD thesis, University of Maryland, College Park, 2007.

[57] L. Nie. Strong consistency of the maximum likelihood estimator in generalized linear and nonlinear mixed-effects models. *Metrika*, 63(2):123–143, 2006.

[58] B. Nielsen. Strong consistency results for least squares estimators in general vector autoregressions with deterministic trends. *Econometric Theory*, 21(03), 2005.

[59] D. Pollard. *Empirical Processes: Theory and Applications*. Conference Board of the Mathematical Science: NSF-CBMS regional conference series in probability and statistics. Institute of Mathematical Statistics, 1990.

[60] Q. Qin and J. P. Hobert. Convergence complexity analysis of Albert and Chib's algorithm for Bayesian probit regression. *The Annals of Statistics*, 2019. To appear.

[61] C. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer New York, 2013.

[62] G. O. Roberts and J. S. Rosenthal. Markov chains and de-initializing processes. *Scandinavian Journal of Statistics*, 28(3):489–504, 2001.

[63] G. O. Roberts and J. S. Rosenthal. General state space Markov chains and MCMC algorithms. *Probability Surveys*, 1(0):20–71, 2004.

[64] B. Roś, F. Bijma, J. C. de Munck, and M. C. de Gunst. Existence and uniqueness of the maximum likelihood estimator for models with a Kronecker product covariance structure. *J. Multivar. Anal.*, 143(C):345–361, 2016.

[65] J. S. Rosenthal. Minorization conditions and convergence rates for Markov chain Monte Carlo. *Journal of the American Statistical Association*, 90(430):558–566, jun 1995.

[66] J. S. Rosenthal. Optimal proposal distributions and adaptive MCMC. In S. Brooks, A. Gelman, G. L. Jones, and X.-L. Meng, editors, *Handbook of Markov chain Monte Carlo*. Chapman & Hall / CRC, 2011.

[67] H. Royden and P. Fitzpatrick. *Real Analysis*. Pearson, 2010.

[68] S. D. Silvey. A note on maximum-likelihood in the case of dependent random variables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 23(2):444–452, 1961.

[69] I. Soloveychik and D. Trushin. Gaussian and robust Kronecker product covariance estimation: Existence and uniqueness. *Journal of Multivariate Analysis*, 149:92–113, 2016.

[70] Y. J. Sung and C. J. Geyer. Monte Carlo likelihood inference for missing data models. *The Annals of Statistics*, 35(3):990–1011, 2007.

[71] A. Szczepańska-Álvarez, C. Hao, Y. Liang, and D. von Rosen. Estimation equations for multivariate linear models with Kronecker structured covariance matrices. *Communications in Statistics – Theory and Methods*, 46(16):7902–7915, 2017.

[72] G. C. Tiao and A. Zellner. On the Bayesian estimation of multivariate regression. *Journal of the Royal Statistical Society. Series B (Methodological)*, 26(2):277–285, 1964.

[73] D. Vats. Geometric ergodicity of Gibbs samplers in Bayesian penalized regression models. *Electronic Journal of Statistics*, 11(2):4033–4064, 2017.

[74] D. Vats, J. M. Flegal, and G. L. Jones. Multivariate output analysis for markov chain monte carlo. *Biometrika*, 2018+.

[75] D. Vats, J. M. Flegal, and G. L. Jones. Strong consistency of multivariate spectral variance estimators in Markov chain monte carlo. *Bernoulli*, 24(3):1860–1909, 2018.

[76] A. Wald. Asymptotic properties of the maximum likelihood estimate of an unknown parameter of a discrete stochastic process. *The Annals of Mathematical Statistics*, 19(1):40–46, 1948.

[77] A. Wald. Note on the consistency of the maximum likelihood estimate. *The Annals of Mathematical Statistics*, 20(4):595–601, 1949.

[78] D. I. Warton, F. G. Blanchet, R. B. O'Hara, O. Ovaskainen, S. Taskinen, S. C. Walker, and F. K. Hui. So many variables: Joint modeling in community ecology. *Trends in Ecology & Evolution*, 30(12):766–779, 2015.

[79] L. Weiss. Asymptotic properties of maximum likelihood estimators in some nonstandard cases. *Journal of the American Statistical Association*, 66(334):345–350, 1971.

[80] H. White. Maximum likelihood estimation of misspecified models. *Econometrica: Journal of the Econometric Society*, pages 1–25, 1982.

[81] A. Wiesel. Geodesic convexity and covariance estimation. *IEEE transactions on signal processing*, 60(12):6182–6189, 2012.

[82] J. Yang and J. S. Rosenthal. Complexity Results for MCMC derived from Quantitative Bounds. *arXiv e-prints*, page arXiv:1708.00829, 2017.

# Appendix A

# Consistent maximum likelihood estimation using subsets

## A.1 Theory

### A.1.1 Preliminary results

We first present some lemmas that will be useful when proving the main results.

**Lemma A.1.1.** *For any positive random variables $X$, $Y$, $Z$, defined on the same probability space, and $c > 0$, $\mathsf{P}(XY \geq Z) \leq \mathsf{P}(X \geq c) + \mathsf{P}(Y \geq Z/c)$.*

*Proof.* If for positive constants $x, y, z, c$ it holds that $xy \geq z$, then either $x \geq c$ or $y \geq z/c$, since otherwise $xy < c(z/c) = z$. Thus, $\{\omega : X(\omega)Y(\omega) \geq Z(\omega)\} \subseteq \{\omega : X(\omega) \geq c\} \cup \{\omega : Y(\omega) \geq Z(\omega)/c\}$. By sub-additivity of measures, $\mathsf{P}(XY \geq Z) \leq \mathsf{P}(\{X \geq c\} \cup \{Y \geq Z/c\}) \leq \mathsf{P}(X \geq c) + \mathsf{P}(Y \geq Z/c)$. □

**Lemma A.1.2.** *Suppose $A_i, i = 1, \ldots, n$ are compact subsets of some metric space $(\mathcal{T}, d_{\mathcal{T}})$ such that $\cap_{i=1}^{n} A_i = \emptyset$, then the open covers $C_i = \cup_{x \in A_i} B_\delta(x)$, $i = 1, \ldots, s$, also have an empty intersection for all small enough $\delta > 0$.*

*Proof.* Consider the covers $C_{k,i} = \cup_{x \in A_i} B_{1/k}(x)$, $k = 1, 2, \ldots$, $i = 1, \ldots, n$. If $C_k = \cap_i C_{k,i} = \emptyset$ for some $k < \infty$, then we are done. Suppose for contradiction $C_k$ is non-empty for every $k < \infty$. By construction, every point $x_k \in C_k$ is within $1/k$ of at least one point in every $A_i$. That is, we can pick, for every $k \geq 1$ and $i = 1, \ldots, n$, an $x_k \in C_k$ and $y_{k,i} \in A_i$ such

that $d(x_k, y_{k,i}) \leq 1/k$. Thus, by the triangle inequality, for every $k$, $d(y_{k,i}, y_{k,j}) \leq 2/k$. By compactness of $A_1$, say, $y_{k,1}$ has a convergent subsequence $y_{k_m,1} \to y_1$ as $m \to \infty$, for some $y_1 \in A_1$ by the fact that $A_1$ is closed as a compact subset of a metric space. But then, for every $i$, by the triangle inequality, $d(y_{k_m,i}, y_1) \leq d(y_{k_m,i}, y_{k_m,1}) + d(y_{k_m,1}, y_1) \leq 2/k_m + d(y_{k_m,1}, y_1) \to 0$ as $m \to \infty$. Thus, since every $A_i$ is closed, $y_1 \in A_i$ for every $i$, which is the desired contradiction. □

**Lemma A.1.3.** *Suppose $\Theta$ is a compact subset of some metric space and, for every $\theta \in \Theta$, $f_\theta$ is a probability density against some dominating measure $\nu$ which does not depend on $\theta$. Suppose also that $f_\theta(x)$ is continuous in $\theta$ for every $x$ and define the measures $\nu_\theta$ by $\nu_\theta(A) = \int_A f_\theta(x)\mathrm{d}\nu(x)$ for any $\nu$-measurable $A$. Then for any $\theta^0 \in \Theta$, the set $\Theta^0 = \{\theta \in \Theta : \ \nu_\theta = \nu_{\theta^0}\}$ is compact.*

*Proof.* Because $\Theta$ is a compact subset of a metric space, it suffices to show that $\Theta^0$ is closed. Note that $\Theta^0$ always includes the point $\theta^0$ and is thus non-empty. Pick an arbitrary converging sequence $\theta_n \in \Theta^0$, call the limit point $\theta^\star$. By continuity of $\theta \mapsto f_\theta(x)$ for every $x$, $f_{\theta_n} \to f_{\theta^\star}$ pointwise. Now for any $\nu$-measurable $A$, $|\nu_{\theta^\star}(A) - \nu_{\theta^0}(A)| \leq |\nu_{\theta^\star}(A) - \nu_{\theta_n}(A)| + |\nu_{\theta^0}(A) - \nu_{\theta_n}(A)| = |\nu_{\theta^\star}(A) - \nu_{\theta_n}(A)|$, which vanishes as $n \to \infty$ by a generalized dominated convergence theorem [67, Theorem 19] – the dominating sequence of functions for which the integrals converge can be $f_{\theta_n}(x) \geq f_{\theta_n}(x) I_A(x)$ – so indeed $\theta^\star \in \Theta^0$. □

### A.1.2 Main results

For economical notation in the proofs we write $f_\theta(y) = f^n_\theta(y)$, $f_\theta(y_i) = f_{\theta,i}(y_i)$, $f_\theta(w) = g_\theta(w)$, $f_\theta(u) = \phi^r_\theta(u)$, and so on. That is, the letter $f$ is overloaded and the argument indicates which density we are referring to.

*Proof Lemma 2.2.1.* Let $Y = (W, Z)$, where $Z$ consists of the components of $Y$ that are not in the subcollection $W$. Then $f_\theta(y) = f_\theta(w, z)$ and by (conditional) Markov's inequality,

for any $k > 0$,

$$\mathsf{P}\left(L_n(\theta; Y) \geq c \mid W\right) \leq c^{-1}\mathbb{E}\left(L_n(\theta; Y) \mid W\right) = c^{-1}\mathbb{E}\left(\frac{f_\theta(W, Z)}{f_{\theta^0}(W, Z)} \mid W\right).$$

Now the following calculation shows the random variable $L_m(\theta; W) = f_\theta(W)/f_{\theta^0}(W)$ is a version of $\mathbb{E}\left(f_\theta(W, Z)/f_{\theta^0}(W, Z) \mid W\right)$:

$$\begin{aligned}\int_{\mathcal{Z}} \frac{f_\theta(w, z)}{f_{\theta^0}(w, z)} f_{\theta^0}(z \mid w)\nu_Z(\mathrm{d}z) &= \int_{\mathcal{Z}} \frac{f_\theta(w, z)}{f_{\theta^0}(w, z)} \frac{f_{\theta^0}(w, z)}{f_{\theta^0}(w)}\nu_Z(\mathrm{d}z) \\ &= \int_{\mathcal{Z}} \frac{f_\theta(w, z)}{f_{\theta^0}(w)}\nu_Z(\mathrm{d}z) \\ &= \frac{f_\theta(w)}{f_{\theta^0}(w)},\end{aligned}$$

where $\nu_Z$ is the measure against which the components in $Z$ have joint density $f_\theta(z)$ and $\mathcal{Z}$ is the range space of $Z$. Since the conditional expectation is unique up to $\mathsf{P}$-null sets, this finishes the proof. □

*Proof of Lemma 2.2.2.* Fix some arbitrary $\varepsilon > 0$. If $\sup_{\theta \in A_i} L_n(\theta; Y) < 1$ for $i = 1, \ldots, s$, then, since $L_n(\theta^0; Y) = 1$, there are no global maximizers in $\cup_{i=1}^s A_i \supseteq \Theta \cap B_\varepsilon(\theta^0)^c$. Thus, it suffices to prove

$$\mathsf{P}\left(\bigcup_{i=1}^s \left\{\sup_{\theta \in A_i} L_n(\theta; Y) \geq 1\right\}\right) \leq \sum_{i=1}^s \mathsf{P}\left(\sup_{\theta \in A_i} L_n(\theta; Y) \geq 1\right) \to 0.$$

Since $s$ is fixed it is enough that $\mathsf{P}\left(\sup_{\theta \in A_i} L_n(\theta; Y) \geq 1\right) \to 0$ for every $i = 1, \ldots, s$. Without loss of generality, consider $i = 1$. Pick a cover of $A_1$ as given by Assumption 3 and, for every ball in the cover, pick a $\theta^j$ in the intersection of that ball with $A_1$. If there are some balls that do not intersect $A_1$, they may be discarded from the cover, so we assume without loss of generality that all balls do intersect $A_1$. We then get $M_{n,1}$ points such that every point in $A_1$ is within $\delta_{n,1}$ of at least one of them. For any $\theta \in A_1$, let $\theta^j(\theta)$ denote the $\theta^j$ closest to it (pick an arbitrary one if there are many). Using the Lipschitz continuity

given by Assumption 2 and that $x \mapsto e^x$ is increasing we have,

$$\begin{aligned}
\mathsf{P}\left(\sup_{\theta\in A_1} L_n(\theta;Y) \geq 1\right) &= \mathsf{P}\left(\sup_{\theta\in A_1} \Lambda_n(\theta;Y) \geq 0\right) \\
&= \mathsf{P}\left(\sup_{\theta\in A_1} \ell_n(\theta;Y) \geq \ell_n(\theta^0;Y)\right) \\
&\leq \mathsf{P}\left(\sup_{\theta\in A_1} \left[\ell_n(\theta^j(\theta);Y) + K_{n,1} d_{\mathcal{T}}(\theta,\theta^j(\theta))\right] \geq \ell_n(\theta^0;Y)\right).
\end{aligned}$$

Because there are only $M_{n,1}$ points $\theta^j$, and $d_{\mathcal{T}}(\theta^j(\theta),\theta) \leq \delta_{n,1}$ since $\theta^j(\theta)$ is the one closest to $\theta$, we get that the last line is upper bounded by

$$\mathsf{P}\left(\max_{j\leq M_{n,1}} \left[\ell_n(\theta^j;Y) + K_{n,1}\delta_{n,1}\right] \geq \ell_n(\theta^0;Y)\right) = \mathsf{P}\left(\max_{j\leq M_{n,1}} f_{\theta^j}(Y) e^{K_{n,1}\delta_{n,1}} \geq f_{\theta^0}(Y)\right).$$

But by applying Lemma A.1.1 with $c = 2$,

$$\begin{aligned}
\mathsf{P}\left(\max_{j\leq M_{n,1}} f_{\theta^j}(Y) e^{K_{n,1}\delta_{n,1}} \geq f_{\theta^0}(Y)\right) &\leq \mathsf{P}\left(2 \max_{j\leq M_{n,1}} f_{\theta^j}(Y) \geq f_{\theta^0}(Y)\right) + \mathsf{P}\left(e^{K_{n,1}\delta_{n,1}} \geq 2\right) \\
&= \mathsf{P}\left(2 \max_{j\leq M_{n,1}} f_{\theta^j}(Y) \geq f_{\theta^0}(Y)\right) + o(1)
\end{aligned}$$

where the last line uses Assumption 3. The choice of the constant 2 in the application of Lemma A.1.1 is arbitrary – any number with positive logarithm works. The remaining term,

$$\mathsf{P}\left(2 \max_{j\leq M_{n,1}} f_{\theta^j}(Y) \geq f_{\theta^0}(Y)\right) = \mathsf{P}\left(\max_{j\leq M_{n,1}} L_n(\theta^j;Y) \geq 1/2\right),$$

we will deal with using Lemma 2.2.1 and dominated convergence. After conditioning on $W^{(1)}$ we have

$$\begin{aligned}
\mathsf{P}\left(\max_{j\leq M_{n,1}} L_n(\theta^j;Y) \geq 1/2 \mid W^{(1)}\right) &\leq \sum_{i=1}^{M_{n,1}} 2L_{m_1}(\theta^j;W^{(1)}) \\
&\leq 2M_{n,1} \sup_{\theta\in A_1} L_{m_1}(\theta,W^{(1)}),
\end{aligned}$$

$\mathsf{P}$-almost surely, where the first inequality is by subadditivity and Lemma 2.2.1, and the

second uses that $L_n(\theta^j; W^{(1)}) \leq \sup_{\theta \in A_1} L_{m_1}(\theta; W^{(1)})$ by definition. The expression in the last line vanishes as $n \to \infty$ by Assumption 3. Thus,

$$\mathsf{P}\left( \max_{j \leq M_{n,1}} L_n(\theta^j; Y) \geq 1/2 \right) \to 0$$

by dominated convergence. The dominating function can be the constant 1. This finishes the proof. □

## A.2 Applications

Let $\lambda_{\max}(\cdot)$ and $\lambda_{\min}(\cdot)$ denote the maximum and minimum eigenvalue of its matrix argument, respectively. For matrices, $\|\cdot\|$ denotes the spectral norm and $\|\cdot\|_F$ the Frobenius norm. Differentiation with respect to $\theta_i$ is denoted $\nabla_i$.

We will use the following well known fact repeatedly. It is stated as a lemma for easy reference.

**Lemma A.2.1.** *If $h$ is a continuous function from some metric space $\mathcal{X}$ to $\mathbb{R}$ and $A$ is a compact subset of $\mathcal{X}$, then $\sup_{x \in A} h(x) = h(x^\star)$ for some $x^\star \in A$. In particular, if $h(x) < c$ for some constant $c$ and every $x \in A$, then $\sup_{x \in A} h(x) < c$.*

Of course, the same holds if the supremum is replaced by an infimum or if less than is replaced by greater than.

**Lemma A.2.2.** *Let $X_{n,1}, \ldots, X_{n,n}$ be a triangular array with rows of i.i.d. multivariate normal $q$-vectors with mean $\mathbb{E}(X_{n,i}) = \mu = \mu(\theta)$ and covariance matrix $\operatorname{cov}(X_{n,i}) = \Sigma = \Sigma(\theta)$, $\theta \in \Theta$. Suppose that*

$$0 < 1/c_1 \leq \inf_{\theta \in \Theta} \lambda_{\min}(\Sigma(\theta)) \leq \sup_{\theta \in \Theta} \lambda_{\max}(\Sigma(\theta)) \leq c_1 < \infty$$

*and $\sup_{\theta \in \Theta} \|\mu(\theta)\| \leq c_2$ for some $c_1, c_2 \in (0, \infty)$, then*

$$\sup_{\theta \in \Theta} \left| n^{-1} \sum_{i=1}^{n} \{\Lambda_i(\theta; X_{n,i}) - \mathbb{E}[\Lambda_1(\theta; X_{n,1})]\} \right| \to 0,$$

$\mathsf{P}$-*almost surely, where* $\Lambda_i(\theta; X_{n,i}) = \log f_\theta(X_{n,i})/f_{\theta^0}(X_{n,i})$*, and* $f_\theta(X_{n,i})$ *means the density for* $X_{n,i}$ *evaluated at* $X_{n,i}$*.*

*Proof.* Theorem 16 in Ferguson [18] applies almost verbatim to triangular arrays in place of i.i.d. sequences. The only necessary modification to its proof is that the pointwise strong law of large numbers needs to be motivated. Write

$$\begin{aligned}
\sup_{\theta\in\Theta}|\Lambda_1(\theta;x)| &\le \sup_{\theta\in\Theta}|\log f_\theta(x)| + |\log f_{\theta^0}(x)| \\
&\le \sup_{\theta\in\Theta}|\log\det\Sigma| + \sup_{\theta\in\Theta}\|x-\mu\|^2\|\Sigma^{-1}\| + |\log f_{\theta^0}(x)| \\
&\le \sup_{\theta\in\Theta}|\log\det\Sigma| + \sup_{\theta\in\Theta}(\|x\|+\|\mu\|)^2 \sup_{\theta\in\Theta}\|\Sigma^{-1}\| + |\log f_{\theta^0}(x)| \\
&\le |\log(qc_1)| + (\|x\|+c_2)^2c_1 + |\log f_{\theta^0}(x)| \\
&=: K(x),
\end{aligned}$$

which is a quadratic function of $x$, not depending on $\theta$. Thus, since the $X_{n,i}$s are i.i.d. and normal random variables have all finite moments, $\Lambda_i(\theta; X_{n,i})$ has bounded fourth moment, uniformly in $i$, $n$, and $\theta$. Classical proofs for a strong law with finite fourth moment applies without change to triangular arrays. The other conditions of Ferguson's Theorem 16 are easy to verify, using $K(x)$ as the dominating function. □

*Proof Proposition 2.3.1.* Lemma A.1.3 gives that $\{\theta \in \Theta : \nu^i_\theta = \nu^i_{\theta^0}\}$ is a closed set, $i = 1,\dots,s$. Thus, the sets $D_i = \{\theta \in \Theta : \nu^i_\theta = \nu^i_{\theta^0}\} \cap B_\varepsilon(\theta^0)^c$, $i = 1,\dots,s$, are closed as intersections of closed sets and compact as a closed subsets of a compact set, $\Theta$. By Lemma A.1.2 we can pick $\delta$ small enough that the open covers $B_i = \cup_{\theta\in D_i} B_\delta(\theta) \supseteq D_i$ have an empty intersection, $\cap_{i=1}^s B_i = \emptyset$. Let $A_i = \Theta \cap B_\varepsilon(\theta^0)^c \cap B_i^c$ and note $\cup_{i=1}^s A_i = \Theta \cap B_\varepsilon(\theta^0)^c \cap (\cup_{i=1}^s B_i^c) = \Theta \cap B_\varepsilon(\theta^0)^c \cap (\cap_{i=1}^s B_i)^c = \Theta \cap B_\varepsilon(\theta^0)^c$. Each $A_i$ closed as the intersection of closed sets, and compact as a closed subset of a compact set, $\Theta$. By construction, for any $\theta \in A_i$ it must be that $\theta \in B_i^c \subseteq D_i^c$. Since $A_i$ is a subset of $B_\varepsilon(\theta^0)^c$ by construction this implies $\theta \in \{\theta \in \Theta : \nu^i_\theta = \nu^i_{\theta^0}\}^c$, which finishes the proof. □

### A.2.1 Longitudinal linear mixed model

**Lemma A.2.3.** *The log-likelihood $\ell_n(\theta; y)$ is differentiable in $\theta$ at any interior point of $\Theta$, for every $n \geq 1$ and every $y$ in the support of $Y$.*

*Proof.* The multivariate normal log-likelihood $\ell_n(\theta; Y)$ is differentiable in its mean $m$ and covariance matrix $C$ everywhere $C = C(\theta)$ is positive definite [50]. It is easy to see that $C$ is positive definite on all interior point since $\Psi$ is (c.f. Lemma A.2.4). Now $\ell_n(\theta; Y)$ is differentiable on all interior points by the chain rule since the elements of $m$ and $C$ are differentiable in $\theta$. □

*Proof Lemma 2.3.2.* Fix an $\varepsilon > 0$ small enough that all points of $\bar{B}_\varepsilon(\theta^0)$ are interior. By construction of the subcollections, the assumptions of Proposition 2.3.1 are satisfied with what is there denoted $\Theta$ replaced by $\bar{B}_\varepsilon(\theta^0)$. Take $A_1$ and $A_2$ to be the compact sets given by that proposition. The proof of point 1 is standard [18, p. 115] and hence omitted. Point 2 is proven by checking the conditions of Lemma A.2.2 with what is there denoted $\Theta$ replaced by the compact $A_i, i = 1, 2$. The following argument works for either subcollection. First note $\lambda_{\max}(C_i) = \|C_i(\theta)\| \leq \|C_i(\theta)\|_F$ [6]. Since the Frobenius norm is the square root of the sum of squared entries and the entries are continuous functions of $\theta$, $\theta \mapsto \|C_i(\theta)\|_F$ is continuous and attains its supremum on the compact set $A_i$, so $\|C(\theta)\|$ is bounded above on $\bar{B}_\varepsilon(\theta^0)$. By spectral decomposition of $C_i$ it is immediate that $\lambda_{\min}(C_i) = 1/\lambda_{\max}(C_i^{-1})$. Thus, since $C_i(\theta)$ is clearly positive definite on all interior points and the inverse is a continuous mapping at points where $C_i$ is positive definite [50], we get by the same arguments that $\lambda_{\max}(C_i^{-1}(\theta))$ is bounded on $A_i$. It is obvious that $\theta \mapsto m_i(\theta)$ is continuous and hence attains its supremum on $A_i$. This concludes the proof of point 2.

It remains to prove point 3. By point 1 we may pick an $\epsilon > 0$ such that, for either subcollection,

$$\sup_{\theta \in A_i} N^{-1}\mathbb{E}[\Lambda_{N/2}(\theta; W^{(i)})] = \sup_{\theta \in A_i} \mathbb{E}[\Lambda_1(\theta; W_1^{(i)})]/2 < -3\epsilon.$$

By point 2 we have, P-almost surely and for all large enough $N$,

$$\sup_{\theta \in A_i} N^{-1} |\Lambda_{N/2}(\theta; W^{(i)}) - \mathbb{E}[\Lambda_{N/2}(\theta; W^{(i)})]| \leq \epsilon.$$

Thus we have that $\sup_{\theta \in A_i} \Lambda_{N/2}(\theta; W^{(i)}) < -2N\epsilon$, and hence that

$$\sup_{\theta \in A_i} L_{N/2}(\theta; X^{(i)}) \leq e^{-2N\varepsilon},$$

for all large enough $N$, P-almost surely. The last right hand side is clearly $o(e^{-\epsilon N})$ as $N \to \infty$. □

We will use the following results in the proof of Lemma 2.3.6.

**Lemma A.2.4.** *The following hold when all points in $\bar{B}_\varepsilon(\theta^0)$ are interior (the first inequality in 1 holds always):*

1. $\|\Psi\|_F \leq T$ *and* $\sup_{\theta \in \bar{B}_\varepsilon(\theta^0)} \|\Psi^{-1}\|_F \leq c\sqrt{T}$ *for some* $c > 0$,
2. $\sup_{\theta \in \bar{B}_\varepsilon(\theta^0)} \|C(\theta)\| \leq c_1 NT + c_2 T + c_3$ *for some* $c_1, c_2, c_3 > 0$,
3. $\sup_{\theta \in \bar{B}_\varepsilon(\theta^0)} \|C(\theta)^{-1}\| \leq c$ *for some* $c > 0$,
4. $\sup_{\theta \in \bar{B}_\varepsilon(\theta^0)} \|\nabla_i \Sigma\| \leq NT + cT^2$ *for some* $c > 0$ *and every* $i \geq 3$.
5. $\sup_{\theta \in \bar{B}_\varepsilon(\theta^0)} \|Y - m(\theta)\| = o_{\mathsf{P}}(n)$, *and*

*Proof.* 1. The Frobenius norm is the square root of the sum of squared elements, and all elements of $\Psi$ are in the form $\theta_7^k$ for some integer $k$ – this establishes the first inequality. The inverse of $\Psi$ can be written as $(1 - \theta_7^2)^{-1}$ times a tri-diagonal matrix where the diagonal entries are 1 or $1 + \theta_7^2$, and the leading off-diagonals have entries $-\theta_7$. Thus, $\|\Psi^{-1}\|_F$ is the square root of the sum of $3T$ possibly non-zero elements, each a continuous function of $\theta$. The inequality now follows from Lemma A.2.1.

2. Using that eigenvalues of the sum of two positive, semi-definite matrices must be at least as large as those of either summand and that the eigenvalues of Kronecker

products are the products of the multiplicands' eigenvalues [50], we get

$$\begin{aligned}\lambda_{\max}(C) &\leq \lambda_{\max}(\theta_3 I_n) + \lambda_{\max}(\theta_4 I_N \otimes J_{NT}) + \lambda_{\max}(\theta_5 J_N \otimes I_N \otimes J_T) \\ &\quad + \lambda_{\max}(\theta_6 I_{N^2} \otimes \Psi) \\ &\leq \theta_3 + \theta_4 NT + \theta_5 NT + \theta_6 T,\end{aligned}$$

where in the last step we also used $\lambda_{\max}(\Psi) \leq \|\Psi\|_F \leq T$ by 1. The existence of the constants $c_1, c_2, c_3$ now follows from Lemma A.2.1.

3. Since $Z\Sigma Z^\mathsf{T}$ is positive definite, we get $\lambda_{\min}(C) = \lambda_{\min}(\theta_3 I_n + Z\Sigma Z^\mathsf{T}) \geq \theta_3$. Since all points in $\bar{B}_\varepsilon(\theta^0)$ are interior, $\theta_3$ is lower bounded by some $c^{-1} > 0$ on it (Lemma A.2.1). Thus, using that the eigenvalues of $C^{-1}$ are the reciprocals of the eigenvalues of $C$, we get $\|C^{-1}\|_F \leq (nc^2)^{1/2} = N\sqrt{T}c$.

4. Clearly, $\nabla_3 C(\theta) = I_n$ which has eigenvalue 1 with multiplicity $n$. If $i = 4$ or $i = 5$, then the derivative is either $I_N \otimes J_N \otimes J_T$ or $J_N \otimes I_N \otimes J_T$, which both have maximal eigenvalue $NT$. If $i = 6$, then the derivative is $\Psi \otimes I_{N^2}$, which has maximal eigenvalue less than $T$ by 1. If $i = 7$, then the derivative is $\theta_6 \nabla_7 \Psi$. We have $\nabla_7 \Psi_{i,j} = |i - j|\theta_7^{|i-j|-1}$ if $|i-j| \geq 1$ and $\nabla_7 \Psi_{i,j} = 0$ otherwise. Thus, $\nabla_7 \Psi_{i,j} \leq T$ and, consequently, $\|\nabla_7 \Psi\|_F \leq T^2$. We conclude, by Lemma A.2.1, $\nabla_7 C(\theta) \leq cT^2$ for some $c > 0$.

5. Let $U\Lambda U^\mathsf{T}$ be the spectral decomposition of $C$. Then $\|Y - m(\theta)\| = \|U^\mathsf{T}(Y - m(\theta))\|$. The vector $U^\mathsf{T}(Y - m(\theta))$ is multivariate normal with mean 0 and covariance matrix $\Lambda$. Thus, since a Gaussian process is determined by its finite dimensional distributions, the stochastic process $\|Y - m(\theta)\|^2$, $\theta \in \bar{B}_\varepsilon(\theta^0)$, has the same distribution as the process $\sum_{i=1}^n \Lambda_{i,i}(\theta)\xi_i^2$, where $\xi_1, \ldots, \xi_n$ are i.i.d. standard normal. By point 2, the supremum of the latter process satisfies $\sup_{\theta \in B_\varepsilon(\theta^0)} \sum_{i=1}^n \Lambda_{i,i}(\theta)\xi_i^2 \leq (c_1 NT + c_2 T + c_3) \sum_{i=1}^n \xi_i^2 = o_\mathsf{P}(n^2)$, which follows from that the last sum is a positive random variable with mean $n$, and hence it converges to zero in $L_1$ when divided by anything of higher order than $n$.

□

*Proof Proposition 2.3.3.* Let $e = e(\theta) = Y - m(\theta)$ and let $\nabla_e$ and $\nabla_C$ denote differentiation with respect to $e$ and $C$. Since $e$ is linear in $\theta_1$ and $\theta_2$, and $C(\theta)$ is differentiable in each $\theta_i$, $i \geq 3$, bounding the gradient for $\theta$ is easily done after establishing bounds for $\nabla_C \ell_n(\theta; Y)$ and $\nabla_e \ell_n(\theta)$. These derivatives exist for every $n$ because the covariance matrix $C(\theta)$ is positive-definite on $\bar{B}_\varepsilon(\theta^0)$ by Lemma A.2.4 and the multivariate normal log-likelihood is differentiable wherever the covariance matrix is non-singular [50]. We have

$$\nabla_C \ell_n(\theta; Y) = -\frac{1}{2}\left[C^{-1} + C^{-1} e e^\mathsf{T} C^{-1}\right] \text{ and } \nabla_e \ell_n(\theta) = -C^{-1} e.$$

Thus,

$$\begin{aligned}
|\nabla_1 \ell_n(\theta)| &= |\nabla_e \ell_n(\theta)^\mathsf{T} \nabla_1 e(\theta)| = |e^\mathsf{T} C^{-1} 1_n| \leq \|e\| \|C^{-1}\| N^2 T, \\
|\nabla_2 \ell_n(\theta)| &= |\nabla_e \ell_n(\theta)^\mathsf{T} \nabla_2 e(\theta)| = |e^\mathsf{T} C^{-1} h_n| \leq \|e\| \|C^{-1}\| N^2 T/2,
\end{aligned}$$

and, for $i \geq 3$,

$$\begin{aligned}
|\nabla_i \ell_n(\theta)| &= |\operatorname{vec}[\nabla_C \ell_n(\theta)]^\mathsf{T} \operatorname{vec}[\nabla_i C]| = \frac{1}{2} \operatorname{vec}\left[C^{-1} + C^{-1} e e^\mathsf{T} C^{-1}\right]^\mathsf{T} \operatorname{vec}\left[\nabla_i C\right]| \\
&= \operatorname{tr}\left[(C^{-1} + C^{-1} e e^\mathsf{T} C^{-1}) \nabla_i C\right] \\
&\leq \|C^{-1}\|_F \|\nabla_i C\|_F + |e^\mathsf{T} C^{-1} \nabla_i C^{-1} e| \\
&\leq \|C^{-1}\|_F \|\nabla_i C\|_F + \|e\|^2 \|C^{-1}\|^2 \|\nabla_i C\|,
\end{aligned}$$

where $\operatorname{vec}(\cdot)$ denotes the vectorization operator stacking the columns of its matrix argument. Thus, by Lemma A.2.4,

$$\sup_{\theta \in \bar{B}_\varepsilon(\theta)} |\nabla_1 \ell_n(\theta)| \leq \sup_{\theta \in \bar{B}_\varepsilon(\theta)} \|e\| \|C^{-1}\| N^2 T \leq o_\mathsf{P}(n) O(NT + T) T N^2 = o_\mathsf{P}(T^3 N^5),$$

$$\sup_{\theta \in \bar{B}_\varepsilon(\theta)} |\nabla_2 \ell_n(\theta)| \leq \sup_{\theta \in \bar{B}_\varepsilon(\theta)} \|e\| \|C^{-1}\| N^2 T/2 \leq o_\mathsf{P}(n) O(NT + T) T N^2 = o_\mathsf{P}(T^3 N^5),$$

and, for $i \geq 3$,

$$\sup_{\theta \in \bar{B}_\varepsilon(\theta)} |\nabla_i \ell_n(\theta)| \leq \sup_{\theta \in \bar{B}_\varepsilon(\theta)} (\|C^{-1}\|_F \|\nabla_i C\|_F + \|e\|^2 \|C^{-1}\|^2 \|\nabla_i C\|).$$

By Lemma A.2.4 the supremum of each of the terms in the last line are of at most polynomial order, which finishes the proof. □

### A.2.2 Logit–normal MGLMM

**Lemma A.2.5.** *The log-likelihood $\ell_n(\theta; y)$ is differentiable in $\theta$ on $\bar{B}_\varepsilon(\theta^0)$, for every $n \geq 1$ and every $y$ in the support of $Y$.*

*Proof.* To prove differentiability of $f_\theta(y)$ in $\theta$ on $\bar{B}_\varepsilon(\theta^0)$, checking the usual conditions for differentiation under the integral are sufficient [21, Theorem 2.27]. It's obvious that $f_\theta(y \mid u) f_\theta(u)$ is differentiable in $\theta$ on every interior point of $\Theta$, so it suffices to find, for $i = 1, \ldots, d$, functions $K_i : \mathbb{R}^{2n} \times \mathbb{R}^{2N} \to [0, \infty)$, not depending on $\theta$, such that $|\nabla_i f_\theta(y \mid u) f_\theta(u)| \leq K_i(y, u)$ and $\int K_i(y, u) \mathrm{d}u < \infty$. Clearly, $|\nabla_i f_\theta(y \mid u) f_\theta(u)| \leq \|\nabla_{\beta_1} f_\theta(y \mid u) f_\theta(u)\|$, for any $i$ such that $\theta_i$ is a component of $\beta_1$, and similarly for the components of $\beta_2$. Thus, it suffices to find bounds for $\|\nabla_{\beta_i} f_\theta(y \mid u) f_\theta(u)\|$, $i = 1, 2$, and $|\nabla_{\theta_d} f_\theta(y \mid u) f_\theta(u)|$. For the purposes of this integration, the responses $y_{i,j,k}$ are constant and the sample size $n$ is fixed. We prove the existence of integrable bounds in the following forms, where $c_1, \ldots, c_4 > 0$,

1. $K_1(y, u) = c_1 \exp\left(-\frac{1}{2c_2} u^\mathsf{T} u\right) \sum_{i,j} \left(|y_{i,j,1}| + 1 + |u_i^{(1)}| + |u_j^{(2)}|\right) \geq \|\nabla_{\beta_1} f_\theta(y \mid u) f_\theta(u)\|$,
2. $K_2(y, u) = c_3 \exp\left(-\frac{1}{2c_2} u^\mathsf{T} u\right) \geq \|\nabla_{\beta_2} f_\theta(y \mid u) f_\theta(u)\|$, and
3. $K_3(y, u) = c_4 \exp\left(-\frac{1}{2c_2} u^\mathsf{T} u\right) (u^\mathsf{T} u + 1) \geq |\nabla_{\theta_d} f_\theta(y \mid u) f_\theta(u)|$.

It is clear that $K_1, K_2, K_3$ so defined are integrable because they are, up to scaling, moments of multivariate normal distributions. Thus, it remains only to prove the stated inequalities indeed hold.

By the triangle inequality, that $f_\theta(y \mid u) \leq (2\pi)^{-n/2}$, and the fact that $f_\theta(u)$ does not

depend on $\beta_1$, we have

$$\|\nabla_{\beta_1} f_\theta(y \mid u) f_\theta(u)\| = \left\| f_\theta(y \mid u) f_\theta(u) \sum_{i,j} (y_{i,j,1} - l_{i,j,1}) x_{i,j} \right\|$$
$$\leq (2\pi)^{-n/2} (2\pi\theta_d)^{-N} \exp\left(-\frac{1}{2\theta_d} u^\mathsf{T} u\right) \sum_{i,j} (|y_{i,j,1}| + \|\beta_1\| \|x_{i,j}\| + |u_i^{(1)}| + |u_j^{(2)}|) \|x_{i,j}\|$$

The inequality in the definition of $K_1$ follows from Lemma A.2.1 upon noting that $\theta_d$ is bounded both away from zero and above on interior points, that $\|\beta_1\|$ is similarly upper bounded on such points, and that $\|x_{i,j}\| \leq 1$ by assumption.

For the inequality in the definition of $K_2$ we use that $f_\theta(y \mid u) \leq (2\pi)^{-n/2}$ and that $|y_{i,j,2} - 1/(1 + e^{-l_{i,j,2}})| \leq 1$. The latter assertion follows from that $y_{i,j,2} \in \{0, 1\}$ and that $1/(1 + e^t) \in (0, 1)$ for all $t \in \mathbb{R}$. Thus, since $f_\theta(u)$ does not depend on $\beta_2$,

$$\|\nabla_{\beta_2} f_\theta(y \mid u) f_\theta(u)\| = \left\| f_\theta(y \mid u) f_\theta(u) \sum_{i,j} (y_{i,j,2} - 1/(1 + e^{-l_{i,j,2}})) x_{i,j} \right\|$$
$$\leq n(2\pi)^{-n/2} (2\pi\theta_d)^{-N} \exp\left(-\frac{1}{2\theta_d} u^\mathsf{T} u\right) \|x_{i,j}\|.$$

Now the desired inequality follows from again noting the bounds from below and above of $\theta_d$ and that $\|x_{i,j}\| \leq 1$.

The inequality in the definition of $K_3$ follows similarly. First, $f_\theta(y \mid u)$ does not depend on $\theta_d$ so we get

$$|\nabla_{\theta_d} f_\theta(y \mid u) f_\theta(u)| = \left| f_\theta(y \mid u) f_\theta(u) \left(-\frac{N}{\theta_d} + \frac{u^\mathsf{T} u}{2\theta_d^2}\right) \right|$$
$$\leq (2\pi)^{-n/2} (2\pi\theta_d)^{-N} \exp\left(-\frac{u^\mathsf{T} u}{2\theta_d}\right) \left(\frac{N}{\theta_d} + \frac{u^\mathsf{T} u}{2\theta_d^2}\right).$$

Now we are done upon again appealing to the lower and upper bounds of $\theta_d$ on $\bar{B}_\varepsilon(\theta^0)$. □

For the proof of Lemma 2.3.5 we will need the following lemmas.

**Lemma A.2.6.** *Let $\mathcal{X}$ be a metric space and $f : \mathcal{X} \times \mathbb{R}^d \to \mathbb{R}$, for some $d > 0$, be continuous*

*under the product metric. If $\mathcal{X}$ is compact, then $h(y) = \sup_{x\in\mathcal{X}} f(x,y)$ is continuous.*

*Proof.* Fix some $y$ and consider the compact set $A = \mathcal{X} \times \bar{B}_1(y)$. Since $A$ is compact, $f$ is uniformly continuous on $A$. Thus, for any $\epsilon > 0$ we can pick $\delta$ such that for every $(x', y'), (x'', y'') \in A$, it holds that if $d((x', y'), (x'', y'')) < \delta$, then $|f(x', y') - f(d'', y'')| < \epsilon$. Thus, for any $y' \in B_\delta(y) \subseteq B_1(y)$, we have $|h(y) - h(y')| = |\sup_{x\in\mathcal{X}} f(x,y) - \sup_{x\in\mathcal{X}} f(x,y')| \leq \sup_{x\in\mathcal{X}} |f(x,y) - f(x,y')| = |f(x^\star(y,y'),y) - f(x^\star(y,y'),y')| < \epsilon$, where

$$x^\star(y,y') = \arg\max_{x\in\mathcal{X}} |f(x,y) - f(x,y')|.$$

The arg max exists by Lemma A.2.1 since continuity of $f$ implies continuity in $x$ for every $y$. □

**Lemma A.2.7.** *The K–L divergence from a Bernoulli distribution with parameter $p$ to one with parameter $q$ is lower bounded by $2(p-q)^2$.*

*Proof.* By direct computation, the K–L divergence is $p\log(p/q) + (1-p)\log([1-p]/[1-q])$. Now using that $t(1-t) \leq 1/4$ for all $t \in \mathbb{R}$ and assuming $p > q$ we get

$$\begin{aligned} p\log(p/q) + (1-p)\log([1-p]/[1-q]) &= \int_q^p \left(\frac{p}{t} - \frac{1-p}{1-t}\right) \mathrm{d}t \\ &= \int_q^p \left(\frac{p-t}{t(1-t)}\right) \mathrm{d}t \\ &\geq 4\int_q^p (p-t)\mathrm{d}t \\ &= 2(p-q)^2 \end{aligned}$$

If instead $q > p$, then the same inequality results from letting $1-p$ and $1-q$ take the roles of $p$ and $q$. If $p = q$, then the inequality is an equality. □

Let $\mathrm{C}(\delta, G, \|\cdot\|)$ denote the $\delta$-covering number of the set $G$ under the distance associated with the norm $\|\cdot\|$, that is, the least number of open balls of radius $\delta$ needed to cover $G$.

**Lemma A.2.8** (Theorem 8.2 [59])**.** *Let $h_1(\omega,\theta), h_2(\omega,\theta), \ldots$, $\theta \in A \subseteq \Theta$, be independent processes with integrable envelopes $H_1(\omega), H_2(\omega), \ldots$, meaning $|h_i(\omega,\theta)| \leq H_1(\omega)$, for all $i$*

*and* $\theta \in A$. *Let* $H = (H_1, \ldots, H_N)$ *and* $\mathcal{H}_{N,\omega} = \{[h_1(\omega,\theta), \ldots, h_N(\omega,\theta)] \in \mathbb{R}^N : \theta \in A\}$. *If for every* $\epsilon > 0$ *there exists a* $K > 0$ *such that*

1. $N^{-1}\sum_{i=1}^{N} \mathbb{E}[H_i I(H_i > K)] < \epsilon$ *for all* $N$, *and*

2. $\log \mathrm{C}(\epsilon \|H\|_1, \mathcal{H}_{N,\omega}, \|\cdot\|_1) = o_{\mathsf{P}}(N)$ *as* $N \to \infty$,

*then*

$$\sup_{\theta \in A} N^{-1} \left| \sum_{i=1}^{N} h_i(\omega,\theta) - \mathbb{E}(h_i(\omega,\theta)) \right| \xrightarrow{\mathsf{P}} 0.$$

*Proof.* Pollard [59] proves this result with packing numbers replaced by covering numbers. Since [59, p. 10]

$$\mathrm{C}(\epsilon, \mathcal{H}_{N,\omega}, \|\cdot\|_1) \leq \mathrm{D}(\epsilon, \mathcal{H}_{N,\omega}, \|\cdot\|_1) \leq \mathrm{C}(\epsilon/2, \mathcal{H}_{N,\omega}, \|\cdot\|_1),$$

where D denotes packing numbers, there is nothing more to prove. □

*Proof Lemma 2.3.5.* Let us first prove that, given $\varepsilon > 0$, there exists an $\eta > 0$, and hence $A_i = A_i(\varepsilon, \eta)$, $i = 1, 2$, such that point 1 in the lemma holds. The definition of $A_i(\varepsilon, \eta)$ is as in the main text. Let $c(t) = \log(1 + e^t)$ denote the cumulant function in the conditional distribution of $Y_{i,i,2}$ given the random effects and define

$$p_i(\beta_2, \theta_d) = \mathbb{E}\left[ c'\left( x_{i,i}^{\mathsf{T}}\beta_2 + \sqrt{\theta_d/\theta_d^0}\left(U_i^{(1)} + U_j^{(2)}\right)\right)\right].$$

Recall, $\mathbb{E}$ denotes expectation with respect to the distributions indexed by $\theta^0$, so $p_i(\beta_2, \theta_d)$ is the success probability of $Y_{i,i,2}$ when $\beta_2$ and $\theta_d$ are the true parameters.

Note that because the components in $W^{(2)}$ are independent, $\mathbb{E}[\Lambda_N(\theta; W^{(2)})]$ is a sum of $N$ terms, each summand being the negative K–L divergence between two Bernoulli variables with parameters $p_i(\beta_2, \theta_d)$ and $p_i(\beta_2^0, \theta_d^0)$. Thus, by Lemma A.2.7, Jensen's inequality, the

reverse triangle inequality, and the triangle inequality, respectively,

$$
\begin{aligned}
&N^{-1}\mathbb{E}[\Lambda_N(\theta; W^{(2)})] \\
&\leq -2N^{-1}\sum_{i=1}^{N}[p_i(\beta_2,\theta_d) - p_i(\beta_2^0,\theta_d^0)]^2 \\
&\leq -2\left[N^{-1}\sum_{i=1}^{N}|p_i(\beta_2,\theta_d) - p_i(\beta_2^0,\theta_d^0)|\right]^2 \\
&\leq -2\left[N^{-1}\sum_{i=1}^{N}\big||p_i(\beta_2,\theta_d) - p_i(\beta_2,\theta_d^0)| - |p_i(\beta_2^0,\theta_d^0) - p_i(\beta_2,\theta_d^0)|\big|\right]^2 \\
&\leq -2\left[N^{-1}\sum_{i=1}^{N}|p_i(\beta_2,\theta_d) - p_i(\beta_2,\theta_d^0)| - N^{-1}\sum_{i=1}^{N}|p_i(\beta_2^0,\theta_d^0) - p_i(\beta_2,\theta_d^0)|\right]^2. \qquad \text{(A.1)}
\end{aligned}
$$

Let us work separately with the averages in the last line. We will show that the second can be made arbitrarily small on $A_2$ by selecting $\eta$ small enough, and that the first is bounded away from zero on the same $A_2$, leading to an asymptotic upper bound on $\sup_{\theta\in A_2} N^{-1}\mathbb{E}[\Lambda_N(\theta; W^{(2)})]$ away from zero. We start with the first average.

Let $H$ be a compact subset of $\mathbb{R}$ such that $x_{i,i}^{\mathsf{T}}\beta_2 \in H$ for all $i$ and $\theta \in \bar{B}_\varepsilon(\theta^0)$. Such $H$ exists because the predictors are bounded and $\beta_2$ is bounded on $\bar{B}_\varepsilon(\theta^0)$. Then, defining $\tilde{p}_i(\gamma,\theta_d)$ as $p_i(\beta_2,\theta_d)$ but with $x_{i,i}^{\mathsf{T}}\beta_2$ replaced by $\gamma$, we get

$$
\sup_{\theta\in A_2}|p_i(\beta_2,\theta_d) - p_i(\beta_2,\theta_d^0)| \leq \sup_{\theta\in A_2}\sup_{\gamma\in H}|\tilde{p}_i(\gamma,\theta_d) - \tilde{p}_i(\gamma,\theta_d^0)|.
$$

Since the random variable in the expectation defining $\tilde{p}_i$ is bounded by 1 (it is the mean of a Bernoulli random variable), $\tilde{p}_i$ is continuous by dominated convergence. Thus, since $H$ is compact, $\sup_{\gamma\in H}|\tilde{p}_i(\gamma,\theta_d) - \tilde{p}_i(\gamma,\theta_d^0)|$ is continuous in $\theta_d$ by Lemma A.2.6. That is, we can make $\sup_{\gamma\in H}|\tilde{p}_i(\gamma,\theta_d) - \tilde{p}_i(\gamma,\theta_d^0)|$ arbitrarily small on $A_2 = A_2(\eta,\varepsilon)$ by picking $\eta$ small enough, which is what we wanted to show. We next work with the second average in (A.1).

By the mean value theorem, for some $\tilde{\beta}_{2,i}$ between $\beta_2$ and $\beta_2^0$, $|p_i(\beta_2^0,\theta_d^0) - p_i(\beta_2,\theta_d^0)| = |\mathbb{E}(c''(x_{i,i}^{\mathsf{T}}\tilde{\beta}_{2,i} + U_i^{(2)} + U_j^{(2)}))x_{i,i}^{\mathsf{T}}(\beta_2 - \beta_2^0)|$. Here, differentiation under the expectation is permissible since $c''$ is the variance of a Bernoulli random variable, hence bounded by $1/4$,

and $|x_{ii}^{\mathsf{T}}(\beta_2 - \beta_2^0)| \leq \|x_{i,i}\|\|\beta_2 - \beta_2^0\|^2 \leq \varepsilon$ on $\bar{B}_\varepsilon(\theta^0)$. By the same bound on $c''$ we get that $\mathbb{E}(c''(\gamma + U_i^{(1)} + U_j^{(2)}))$ is continuous in $\gamma$. Thus, by Lemma A.2.1, $\inf_{\gamma \in H} \mathbb{E}(c''(\gamma + U_i^{(1)} + U_j^{(2)})) \geq c_1 > 0$. That $c_1$ must be positive follows from that $c''$ is strictly positive on all of $\mathbb{R}$. We have thus proven that $|p_i(\beta_2^0, \theta_d^0) - p_i(\beta_2, \theta_d^0)| \geq c_1 |x_i^{\mathsf{T}}(\beta_2 - \beta_2^0)|$, uniformly on $\bar{B}_\varepsilon(\theta^0)$. Using this and that $|x_{i,i}^{\mathsf{T}}(\beta_2 - \beta_2^0)| \leq \|x_{i,i}\|\|\beta_2 - \beta_2^0\| \leq \varepsilon \leq 1$ so that squaring it makes it smaller,

$$\begin{aligned} N^{-1} \sum_{i=1}^N |p_i(\beta_2^0, \theta_d^0) - p_i(\beta_2, \theta_d^0)| &\geq c_1 N^{-1} \sum_{i=1}^N |x_{i,i}^{\mathsf{T}}(\beta_2 - \beta_2^0)| \\ &\geq c_1 N^{-1} (\beta_2 - \beta_2^0)^{\mathsf{T}} \left( \sum_{i=1}^N x_{i,i} x_{i,i}^{\mathsf{T}} \right) (\beta_2 - \beta_2^0) \\ &\geq c_1 \|\beta_2 - \beta_2^0\|^2 N^{-1} \lambda_{\min} \left( \sum_{i=1}^N x_{i,i} x_{i,i}^{\mathsf{T}} \right) \end{aligned}$$

which lower limit as $N \to \infty$ is bounded below by some strictly positive constant, say $c_2$, since $\liminf_{N\to\infty} N^{-1} \lambda_{\min} \left( \sum_{i=1}^N x_{i,i} x_{i,i}^{\mathsf{T}} \right) \geq c_3 > 0$, for some $c_3$, and $\|\beta_2 - \beta_2^0\| \geq \varepsilon/2 > 0$ on $A_2$. To summarize, we may pick $\eta$ so small that the second average in (A.1) is less than $c_2/2$, say, and hence get $\sup_{\theta \in A_2} N^{-1} \mathbb{E}[\Lambda_N(\theta; W^{(2)})] \leq -2(c_2 - c_2/2)^2 < 0$, for all but at most finitely many $N$. This proves point 1 as it pertains to $A_2$.

Consider next

$$A_1 = \partial B_\varepsilon(\theta^0) \cap \big( \{\theta : |\theta_d - \theta_d^0| \geq \eta\} \cup \{\theta : \|\beta_2 - \beta_2^0\| \leq \varepsilon/2\} \big)$$

and $W^{(1)}$. Similarly to for $W^{(2)}$, $\mathbb{E}[\Lambda_N(\theta; W^{(1)})]$ can due to independence be written as a sum of $N$ terms in the form

$$\mathbb{E}\{\log[f_\theta(Y_{i,i,1})/f_{\theta^0}(Y_{i,i,1})]\} = -\frac{1}{2} \left[ \log \left( \frac{1 + 2\theta_d}{1 + 2\theta_d^0} \right) + \frac{1 + 2\theta_d^0 + [x_i^{\mathsf{T}}(\beta_2 - \beta_2^0)]^2}{1 + 2\theta_d} - 1 \right], \quad \text{(A.2)}$$

which is the negative K–L divergence between two univariate normal distributions. Let us consider the possible values this can take for $\theta \in A_1$. If $|\theta_d - \theta_d^0| \geq \eta$, then (A.2) is upper bounded by what is obtained when $\beta_1 = \beta_1^0$. This in turn is a continuous function in $\theta_d$

and hence attains its supremum on the compact set $\{\theta_d : \eta \leq |\theta_d - \theta_d^0| \leq \varepsilon\}$, and hence on $A_1$. This supremum is strictly positive because the divergence can be zero only if $\theta_d = \theta_d^0$. If instead $\|\beta_2 - \beta_2^0\| \leq \varepsilon/2$. Then either $|\theta_d - \theta_d^0| \geq \varepsilon/4$ or $\|\beta_1 - \beta_1^0\| \geq \varepsilon/4$, for otherwise it cannot be that $\|\theta - \theta^0\| = \varepsilon$. If $|\theta_d - \theta_d^0| \geq \varepsilon/4$ the divergence in (A.2) has a lower bound away from zero by the same argument as for the cases $|\theta_d - \theta_d^0| \geq \eta$. It remains to deal with the case $\|\beta_1 - \beta_1^0\| \geq \varepsilon/4$.

Write $[x_{i,i}^{\mathsf{T}}(\beta_1^0 - \beta_1)]^2 = (\beta_1^0 - \beta_1)^{\mathsf{T}} x_i x_i^{\mathsf{T}} (\beta_1^0 - \beta_1)$ to see that $-2N^{-1}\Lambda_N(\theta; W^{(1)})$ is equal to

$$\log\left(\frac{1+2\theta_d}{1+2\theta_d^0}\right) + \frac{1 + 2\theta_d^0 + N^{-1}\sum_{i=1}^N (\beta_1^0 - \beta_1)^{\mathsf{T}} x_i x_i^{\mathsf{T}} (\beta_1^0 - \beta_1)}{1+2\theta_d} - 1,$$

which has a lower limit that is greater than

$$\log\left(\frac{1+2\theta_d}{1+2\theta_d^0}\right) + \frac{1 + 2\theta_d^0 + c_3(\varepsilon/4)^2}{1+2\theta_d} - 1.$$

This expression is in turn maximized in $\theta_d$ at $\theta_d = \theta_d^0 + c_3(\varepsilon/16)^2$; this follows from a straightforward optimization in $1 + 2\theta_d$. The corresponding maximum evaluates to $\log(1 + 2\theta_d^0 + c_3(\varepsilon/4)^2) - \log(1 + 2\theta_d^0) > 0$. This finishes the proof of point 1.

The proof of point 2 consists of checking the conditions of Lemma A.2.8. We first work with $A_1$ and $W^{(1)}$. Let $h_i(\omega, \theta) = \log[f_\theta(Y_{i,i,1}(\omega))/f_{\theta^0}(Y_{i,i,1}(\omega))]$ be the log-likelihood ratio for the $i$th observation in the first subcollection, $i = 1, \ldots, N$. We equip $\mathcal{H}_{N,\omega}$ with the $L_1$ norm $\|\cdot\|_1$, and $\Theta$ is equipped with the $L_2$ norm as before. To facilitate checking the two conditions we will first derive envelopes with the following properties: $\sup_{-\infty<i<\infty} \mathbb{E}H_i^k < \infty$ for every $k \geq 0$, $\sup_{-\infty<i<\infty} \mathsf{P}(H_i \geq K) \to 0$ as $K \to 0$, and each $h_i(\omega, \theta)$ is $H_i$-Lipschitz in $\theta$ on $\bar{B}_\varepsilon(\theta^0)$, and hence on $A_1$, for every $\omega$. We start with the Lipschitz property.

Let us use the slight abuse of notation that $y_{i,i,1} = Y_{i,i,1}(\omega)$. Since the distribution of $W^{(1)}$ does not depend on $\beta_2$ we have $\nabla_{\beta_2} h_i(\omega, \theta) = 0$, and for some $c_1, c_2, c_3, c_4, c_5 > 0$

(depending on $\varepsilon$), and every $\theta \in \bar{B}_\varepsilon(\theta^0)$,

$$\|\nabla_{\beta_1} h_i(\omega,\theta)\| = \|(y_{i,i,1} - x_{i,i}^\mathsf{T}\beta_1)x_{i,i}/(1+2\theta_d)\| \le c_1|y_{i,i,1}| + c_2$$
$$|\nabla_{\theta_d} h_i(\omega,\theta)| = \frac{1}{2}\left|\frac{1}{1+2\theta_d} - (y_{i,i,1} - x_{i,i}^\mathsf{T}\beta_1)^2/(1+2\theta_d)^2\right| \le c_3 + c_4(|y_{i,i,1}| + c_5)^2.$$

The existence of these constants follow from Lemma A.2.1. Let $H_i$ be the sum of the bounds, i.e.

$$H_i(\omega) = c_1|y_{i,i,1}| + c_2 + c_3 + c_4(|y_{i,i,1}| + c_5)^2.$$

By the mean value theorem, $|h_i(\omega,\theta) - h_i(\theta',\omega)| = |(\theta-\theta')^\mathsf{T}\nabla h_i(\omega,\tilde{\theta})| \le \|\theta-\theta'\|H_i$ for some $\tilde{\theta}$ between $\theta$ and $\theta'$. That is, $h_i$ is $H_i$-Lipschitz on $\bar{B}_\varepsilon(\theta^0)$. That $H_i$ is an envelope for $h_i$ follows from noting that $h_i(\omega,\theta^0) = 0$ so by taking $\theta' = \theta^0$ in the previous calculation, $|h_i(\omega,\theta)| \le H_i\|\theta-\theta^0\| \le H_i$ on $\bar{B}_\varepsilon(\theta^0)$. That $\sup_i \mathbb{E}(H_i^k) < \infty$ for every $k > 0$ and $\sup_i \mathsf{P}(H_i > K) \to 0$ as $K \to \infty$ follow from that $Y_{i,i,1}$ is normally distributed with variance $1+2\theta_d^0$, not depending on $i$, and mean satisfying $-\|\beta_1^0\| \le x_{i,i}^\mathsf{T}\beta_1^0 \le \|\beta_1^0\|$. We are now ready to check the conditions of Lemma A.2.8.

By the Cauchy–Schwartz inequality and the properties just derived, we have for every fixed $N$ that

$$N^{-1}\sum_{i=1}^N \mathbb{E}[H_i I(H_i > K)] \le \sup_i \mathbb{E}[H_i^2] \sup_i \mathsf{P}(H_i \ge K) \to 0,\ K \to \infty,$$

which verifies the first condition.

For the second condition, note that the derived Lipschitz property gives, for arbitrary

$h = (h_1(\omega, \theta), \ldots, h_N(\omega, \theta))$ and $h' = (h_1(\omega, \theta'), \ldots, h_N(\omega, \theta'))$ in $\mathcal{H}_{N,\omega}$:

$$\begin{aligned}\|h - h'\|_1 &= \sum_{i=1}^{N} |h_i(\omega, \theta) - h_i(\omega, \theta')| \\ &\leq \sum_{i=1}^{N} \|\theta - \theta'\| H_i(\omega) \\ &= \|\theta - \theta'\| \|H\|_1.\end{aligned}$$

Thus, if we cover $\partial B_\varepsilon(\theta^0)$ with $\epsilon$-balls with centers $\theta^j$, $j = 1, \ldots, M$, then the corresponding $L_1$ balls in $\mathbb{R}^N$ of radius $\epsilon\|H\|_1$ with centers $h^j = (h_1(\omega, \theta^j), \ldots, h_N(\omega, \theta^j))$ cover $\mathcal{H}_{N,\omega}$. This is so because for every $\theta \in \partial B_\varepsilon(\theta^0)$ there is a $\theta^j$ such that $\|\theta - \theta^j\| \leq \epsilon$, and hence by the Lipschitz property $\|h(\omega, \theta) - h(\omega, \theta^j)\|_1 \leq \|H\|_1\epsilon$. Thus, $\mathrm{C}(\epsilon\|H\|_1, \mathcal{H}_{N,\omega}, \|\cdot\|_1) \leq \mathrm{C}(\epsilon, \partial B_\varepsilon(\theta^0), \|\cdot\|)$. Since $\mathrm{C}(\epsilon, \partial B_\varepsilon(\theta^0), \|\cdot\|)$ is constant in $N$, the second condition of Lemma A.2.8 is verified for $A_1$ and $W^{(1)}$.

The arguments for $A_2$ and $W^{(2)}$ are similar, redefining $h_i(\omega, \theta)$ with $Y_{i,i,1}$ replaced by $Y_{1,1,2}$, taking $A_2$ in place of $A_1$, and so on. We need only prove the existence of envelopes $H_1, \ldots, H_N$ with the desired properties. Using that $|y_{i,j,2} - c'(l_{i,2,1})]| \leq 1$ and that $f_\theta(y_{i,i,2} \mid u) f_\theta(u)/f_\theta(y_{i,i,2}) = f_\theta(u \mid y_{i,i,2})$ one gets,

$$\begin{aligned}\|\nabla_{\beta_2} h_i(\omega, \theta)\| &= \left\| \nabla_{\beta_2} \log \int f_\theta(y_{i,i,2} \mid u) f_\theta(u) \mathrm{d}u \right\| \\ &= \left\| \frac{1}{f_\theta(y_{i,i,2})} \int f_\theta(y_{i,i,2} \mid u) f_\theta(u) [y_{i,i,2} - c'(l_{i,j,2})] x_{i,i} \mathrm{d}u \right\| \\ &\leq \|x_{i,i}\| \leq 1.\end{aligned}$$

Using that $U_i^{(1)}$ and $U_j^{(2)}$ are the only random effects entering the linear predictor $l_{i,j,2}$, and

that $f_\theta(y_{i,j,2} \mid u) \leq 1$,

$$\begin{aligned}
|\nabla_{\theta_d} h_i(\omega, \theta)| &= \left| \frac{1}{f_\theta(y_{i,i,2})} \int f_\theta(y_{i,i,2} \mid u) f_\theta(u_i^{(1)}, u_j^{(2)}) \left( \frac{(u_i^{(1)})^2 + (u_j^{(2)})^2}{2\theta_d^2} - \frac{1}{\theta_d} \right) \mathrm{d}u \right| \\
&\leq \frac{1}{2\theta_d f_\theta(y_{i,i,2})} \int f_\theta(u_i^{(1)}, u_j^{(2)}) \left( \frac{(u_i^{(1)})^2 + (u_j^{(2)})^2}{\theta_d} \right) \mathrm{d}u + \frac{1}{\theta_d} \\
&= \frac{1}{\theta_d f_\theta(y_{i,j,2})} + \frac{1}{\theta_d}.
\end{aligned}$$

By Lemma A.2.1 the quantity in the last line attains its supremum on $\bar{B}_\varepsilon(\theta^0)$. This maximum is finite for both $y_{i,i,2} = 1$ and $y_{i,i,2} = 0$ since the marginal success probability cannot be one or zero on interior points of $\Theta$. Thus, on $\bar{B}_\varepsilon(\theta^0)$, $\|\nabla h_i(\omega, \theta)\|$ is bounded by a constant, say $H$, the largest needed for the two cases $y_{i,i,2} = 0$ and $y_{i,i,2} = 1$. By setting $H_i = H, i = 1, \ldots, N$, we have envelopes with the right properties and this completes the proof of point 2.

Finally, we prove point 3. Consider without loss of generality the first subset and subcollection. For economical notation we write $L_N(\theta) = L_N(\theta; W^{(1)})$ and $\Lambda_N(\theta) = \Lambda_N(\theta; W^{(1)})$. Point 1 gives that $\sup_{\theta \in A_1} \mathbb{E}[\Lambda_N(\theta)] < -3\epsilon$ for some $\epsilon > 0$ and all large enough $N$. Assuming that $N$ is large enough that this holds, we get

$$\begin{aligned}
\mathsf{P}\left( e^{\epsilon N} \sup_{\theta \in A_1} L_N(\theta) > e^{-\epsilon N} \right) &= \mathsf{P}\left( N^{-1} \sup_{\theta \in A_1} \Lambda_N(\theta) > -2\epsilon \right) \\
&\leq \mathsf{P}\left( N^{-1} \sup_{\theta \in A_1} \Lambda_N(\theta) > \epsilon + \sup_{\theta \in A_1} \mathbb{E}[\Lambda_N(\theta)] \right) \\
&= \mathsf{P}\left( N^{-1} \sup_{\theta \in A_1} \Lambda_N(\theta) - \sup_{\theta \in A_1} \mathbb{E}[\Lambda_N(\theta)] > \epsilon \right) \\
&\leq \mathsf{P}\left( N^{-1} \sup_{\theta \in A_1} |\Lambda_N(\theta) - \mathbb{E}[\Lambda_N(\theta)]| > \epsilon \right),
\end{aligned}$$

which vanishes as $N \to \infty$ by point 2. Thus, since $e^{-\epsilon N} \to 0$, $e^{\epsilon N} \sup_{\theta \in A_1} L_n(\theta) \xrightarrow{\mathsf{P}} 0$. □

*Proof Proposition 2.3.6.* We will find a Lipschitz constant (random variable) with the desired properties by bounding $\|\nabla \log f_\theta(y)\|$. We first consider derivatives with respect to $\theta_d$.

Define

$$\mathrm{J}^n(\theta) = (2\pi\theta_d)^N f_\theta(y) = \int f_\theta(y \mid u) \exp\left(-\frac{u^\mathsf{T} u}{2\theta_d}\right) \mathrm{d}u$$

and

$$\mathrm{K}^n(\theta) = \int f_\theta(y \mid u) \exp\left(-\frac{u^\mathsf{T} u}{2\theta_d}\right) \frac{u^\mathsf{T} u}{2\theta_d^2} \mathrm{d}u.$$

Then $\nabla_{\theta_d} \mathrm{J}^n(\theta) = \mathrm{K}^n(\theta)$, and hence

$$\nabla_{\theta_d} \log f_\theta(y) = \nabla_{\theta_d} \log[(2\pi\theta_d)^{-N} J^n(\theta)] = -\frac{N}{\theta_d} + \frac{\mathrm{K}^n(\theta)}{\mathrm{J}^n(\theta)}.$$

We focus on the second term first. Let $A_n = \{u \in \mathbb{R}^{2N} : u^\mathsf{T} u \leq a_n\}$ for some constant $a_n$ (depending on the total sample size $n$). Let $\mathrm{K}_1^n(\theta)$ be the integral defining $\mathrm{K}^n(\theta)$ restricted to $A_n$, and let $\mathrm{K}_2^n(\theta)$ be the same integral but instead restricted to $A_n^c$ so that $\mathrm{K}^n(\theta) = \mathrm{K}_1^n(\theta) + \mathrm{K}_2^n(\theta)$. Then, since the integrands are non-negative,

$$\mathrm{K}_1^n(\theta)/\mathrm{J}^n(\theta) = \frac{\int_{A_n} f_\theta(y \mid u) \exp\left(-\frac{u^\mathsf{T} u}{2\theta_d}\right) \frac{u^\mathsf{T} u}{2\theta_d^2} \mathrm{d}u}{\int f_\theta(y \mid u) \exp\left(-\frac{u^\mathsf{T} u}{2\theta_d}\right) \mathrm{d}u} \leq \frac{a_n}{2\theta_d^2}$$

and, hence,

$$|\nabla_{\theta_d} \log f_\theta(y)| \leq \frac{N}{\theta_d} + \frac{a_n}{2\theta_d^2} + \frac{\mathrm{K}_2^n(\theta)}{\mathrm{J}^n(\theta)}.$$

On $A_n^c$ we have by definition that $u^\mathsf{T} u \geq u^\mathsf{T} u/2 + a_n/2$. Thus, using that $f_\theta(y \mid u) \leq$

$(2\pi)^{-n/2}$,

$$\begin{aligned}
\mathrm{K}_2^n(\theta) &\leq \int_{A_n^c} f_\theta(y \mid u) \exp\left(-\frac{1}{2\theta_d}(u^\mathsf{T}u/2 + a_n/2)\right) \frac{u^\mathsf{T}u}{2\theta_d^2}\, \mathrm{d}u \\
&\leq \frac{1}{2\theta_d^2} e^{-\frac{a_n}{4\theta_d}} \int f_\theta(y \mid u) \exp\left(-\frac{u^\mathsf{T}u}{4\theta_d}\right) u^\mathsf{T}u\, \mathrm{d}u \\
&\leq \frac{1}{2\theta_d^2} e^{-\frac{a_n}{4\theta_d}} (2\pi)^{-n/2} \int \exp\left(-\frac{u^\mathsf{T}u}{4\theta_d}\right) u^\mathsf{T}u\, \mathrm{d}u \\
&= \frac{1}{2\theta_d^2} e^{-\frac{a_n}{4\theta_d}} (2\pi)^{-n/2} (4\pi\theta_d)^N \int (4\pi\theta_d)^{-N} \exp\left(-\frac{u^\mathsf{T}u}{4\theta_d}\right) u^\mathsf{T}u\, \mathrm{d}u \\
&= \frac{4N\theta_d}{2\theta_d^2} e^{-\frac{a_n}{4\theta_d}} (2\pi)^{-n/2} (4\pi\theta_d)^N. \qquad \text{(A.3)}
\end{aligned}$$

Using Lemma A.2.1, (A.3) can be upper bounded on $\bar{B}_\varepsilon(\theta^0)$ by $h_1^n = \exp(c_1 a_n + c_2 n + c_3 N + c_4 \log N + c_5)$ for some constants $c_1, \ldots, c_5$. It will be important later to note that the constant $c_1$ is negative in this expression.

We next derive a lower bound on $\mathrm{J}^n(\theta)$. To that end, let $B_n = \{u \in \mathbb{R}^{2N} : |u_i| \leq 1, i = 1, \ldots, N\}$. Since the integrand in $\mathrm{J}^n(\theta)$ is positive, we may lower bound it by the same integral restricted to $B_n$. We then get, using that $\exp(-u^\mathsf{T}u/(2\theta_d)) \geq \exp(-N/\theta_d))$ on $B_n$ and that Lebesgue measure of $B_n$ is $4^N$,

$$\begin{aligned}
\mathrm{J}^n(\theta) &\geq \exp\left(-\frac{N}{\theta_d}\right) \int_{B_n} f_\theta(y \mid u) \mathrm{d}u \\
&\geq e^{-\frac{N}{\theta_d}} (2\pi)^{-n/2} \exp\left(-\sum_{i,j} y_{i,j,1}^2/2 + |y_{i,j,1}|(|x_{i,j}^\mathsf{T}\beta_1| + 2) + (|x_{i,j}^\mathsf{T}\beta_1| + 2)^2\right) \\
&\quad \times \exp\left(-\sum_{i,j} |y_{i,j,2}|(|x_{i,j}^\mathsf{T}\beta_2| + 2) + \log(1 + e^{|x_{i,j}^\mathsf{T}\beta_2|} + 2)\right) 4^N. \qquad \text{(A.4)}
\end{aligned}$$

Here, the last inequality lower bounds all terms in the exponent by minus their absolute values. Again using Lemma A.2.1, that the predictors are bounded, and that $|y_{i,j,2}| \leq 1$, we thus see that $\mathrm{J}^n(\theta)$ can be lower bounded on $\bar{B}_\varepsilon(\theta^0)$ by $h_2^n(y) = \exp(c_6 N + c_7 n + c_8 \sum_{i,j} y_{i,j,1}^2 + c_9 \sum_{i,j} |y_{i,j,1}| + c_{10})$, for some constants $c_6, \ldots, c_{10}$. Thus, by lower bounding

$\theta_d > c_{11}^{-1}$ on $\bar{B}_\varepsilon(\theta^0)$ for some $c_{11} > 0$ we get

$$\sup_{\theta \in \bar{B}_\varepsilon(\theta^0)} |\nabla_{\theta_d} \log f_\theta(y)| \leq c_{11} N + c_{11}^2 a_n/2 + \frac{h_1^n}{h_2^n(y)}.$$

Now, take $a_n = n^{1+\epsilon/2}$ for some $\epsilon > 0$. Then the first two terms are $O(a_n)$ as $n \to \infty$. Moreover, since $\sum_{i,j} \mathbb{E} Y_{i,j,1}^2 \leq n(1+2\theta_d^0) + n\|\beta_1^0\| = O(n)$ by boundedness of the predictors, both sums in the exponent of $h_1^n/h_2^n(y)$ converges to zero in $L_1$ if divided by $a_n$, and hence also in probability. It follows from the continuous mapping theorem that $h_1/h_2^n(y) = O_{\mathsf{P}}(1)$ since, as remarked above, $c_1 < 0$. We have thus proven that $\sup_{\theta \in \bar{B}_\varepsilon(\theta^0)} |\nabla_{\theta_d} \log f_\theta(y)| = O_{\mathsf{P}}(a_n) = o_{\mathsf{P}}(n^{1+\epsilon})$, for every $\epsilon > 0$.

For $\beta_1$ we get by using the triangle inequality, boundedness of the predictors, $t(1-t) \leq 1/4$, $t \in \mathbb{R}$, and $f_\theta(y \mid u) f_\theta(u)/f_\theta(y) = f_\theta(u \mid y)$,

$$\begin{aligned}
\|\nabla_{\beta_1} \log f_\theta(y)\| &= \left\| \frac{1}{f_\theta(y)} \int f_\theta(y \mid u) f_\theta(u) \sum_{i,j} [y_{i,j,1} - l_{i,j,1}] x_{i,j} \mathrm{d}u \right\| \\
&\leq \left| \sum_{i,j} (y_{i,j,1} - x_{i,j}^\mathsf{T} \beta_1) \right| + \left| \frac{1}{f_\theta(y)} \int f_\theta(y \mid u) f_\theta(u) \sum_{i,j} |u_i^{(1)} + u_j^{(2)}| \mathrm{d}u \right| \\
&\leq \left| \sum_{i,j} (y_{i,j,1} - x_{i,j}^\mathsf{T} \beta_1) \right| + \left| \frac{1}{f_\theta(y)} \int f_\theta(y \mid u) f_\theta(u) \sum_{i,j} [1/2 + (u_i^{(1)})^2 + (u_j^{(2)})^2] \mathrm{d}u \right| \\
&= \left| \sum_{i,j} (y_{i,j,1} - x_{i,j}^\mathsf{T} \beta_1) \right| + n/2 + \frac{1}{f_\theta(y)} \int f_\theta(y \mid u) f_\theta(u) u^\mathsf{T} u \mathrm{d}u \\
&= \left| \sum_{i,j} (y_{i,j,1} - x_{i,j}^\mathsf{T} \beta_1) \right| + n/2 + 2\theta_d^2 \frac{\mathrm{K}^n(\theta)}{\mathrm{J}^n(\theta)}
\end{aligned}$$

Thus, by Lemma A.2.1 and the same arguments as for $\nabla_{\theta_d} \log f_\theta(y)$ we get that

$$\sup_{\theta \in B_\varepsilon(\theta^0)} \|\nabla_{\beta_1} \log f_\theta(Y)\| = o_{\mathsf{P}}(n^{1+\epsilon})$$

for any $\epsilon > 0$.

Finally, by the triangle inequality and that $|y_{i,j,2} - c'(l_{i,j,2})| \leq 1$ for all $i$ and $j$,

$$\begin{aligned}\|\nabla_{\beta_2} \log f_\theta(y)\| &= \left\| \frac{1}{f_\theta(y)} \int f_\theta(y \mid u) \sum_{i,j} [y_{i,j,2} - c'(l_{i,j,2})] x_{i,j} f_\theta(u) \mathrm{d}u \right\| \\ &\leq n\end{aligned}$$

□

# Appendix B

# Maximum likelihood estimation of covariance matrices with separable correlation

## B.1 A model for sample means

The decision to model data from different years as independent reflects inspection of sample means; that dissolved oxygen is driven by flow, photosynthesis, respiration and other variables that themselves vary considerably within and among years; and that sampling sites are re-selected annually, thereby precluding dependence among years at the sampling unit scale. Strata are defined geomorphically, while reaches denote navigation pools that are routinely sampled by the U.S. Army Corps of Engineers' Long Term Resource Monitoring Element.

Let $Y_{i,j,k,l}$ denote the $l$th measurement in year $i$, area $j$, and season $k$. Assume that correlation between different measurements only depend on in which area and season those measurements are taken. For example, any two distinct measurements from the same area

and season are assumed equally correlated. Mathematically,

$$\operatorname{cov}(Y_{i,j,k,l}, Y_{i',j',k',l'}) = \begin{cases} 0, & \text{if } i \neq i' \\ \sigma^2_{j,k}, & \text{if } i = i', j = j', k = k', l = l' \\ \rho_{j,k;j',k'}\sqrt{\sigma^2_{j,k}\sigma^2_{j',k'}}, & \text{otherwise,} \end{cases} \tag{B.1}$$

where $\sigma^2_{j,k}$ is a variance parameter depending on season and location, and $\rho_{j,k;j',k'}$ is the correlation between measurements from location-area combinations $(j,k)$ and $(j',k')$.

The sample means we model are $Y_{i,j,k} = n^{-1}_{i,j,k}\sum_{l=1}^{n_{i,j,k}} Y_{i,j,k,l}$. Every year, each season-area combination is sampled between 8 – 82 times, so $8 \leq n_{i,j,k} \leq 82$ for all $i,j,k$. The average number of measurements on which the sample means are based, $n_{i,j,k}$, is 39.5. The response vector in our model is taken to be all the measurements from the same year, $Y_i = [Y_{i,1,1}, \ldots, Y_{i,3,16}]^T \in \mathbb{R}^{48}$ $(i = 1, \ldots, n)$, since #seasons×#areas $= 3 \times 16 = 48$. We have 21 independent observations (years) of that response vector.

The dependence structure assumed in (B.1) implies:

$$\operatorname{cov}(Y_{i,j,k}, Y_{i',j',k'}) = \begin{cases} 0, & \text{if } i \neq i' \\ \sigma^2_{j,k}[n^{-1}_{i,j,k} + (1 - n^{-1}_{i,j,k})\rho_{j,k}], & \text{if } i = i', j = j', k = k' \\ \sigma_{j,k,j',k'}, & \text{otherwise.} \end{cases}$$

The second term in the expression for variances of sample means, $\kappa_{i,j,k} = [n^{-1}_{i,j,k} + (1 - n^{-1}_{i,j,k})\rho_{j,k}]$, changes by years, which complicates modeling. This complication does not arise in applications where one is modeling raw data rather than sample means, or where $n_{i,j,k}$ in fact does not depend on $i$. Treating the covariance matrix as constant in years in our data example is an approximation, which error depends on how much $\kappa_{i,j,k}$ differs among years. Exploratory analysis of the full data indicates $\rho_{j,k} \approx 0.5$ for all $j,k$. Using this estimate, we examined the ratio $[\max_i \kappa_{i,j,k} - \min_i \kappa_{i,j,k}]/\bar{\kappa}_{j,k}$, where $\bar{\kappa}_{j,k}$ is the average $\kappa$ for season $k$ and location $j$ over the $n = 21$ years. The maximum such ratio was around 0.08, with only two being grater than 0.05. For 42/48 season-location combinations, the ratio was less

Table B.1: Estimation Error and Test Size II

Data generated with separable correlation and compound symmetry

| $n$ | $r$ | $c$ | $err_{cor}$ | $err_{cov}$ | $err_{ur}$ | $rej_{cov}$ | $rej_{cov,b}$ | $rej_{cor}$ | $rej_{cor,b}$ |
|---|---|---|---|---|---|---|---|---|---|
| 10 | 5 | 5 | 20.16 | 17.68 | - | 0.97 | 0.77 | - | - |
| 10 | 5 | 15 | 54.78 | 37.51 | - | 0.98 | 0.13 | - | - |
| 10 | 15 | 15 | 107.76 | 75.05 | - | 1.00 | 0.33 | - | - |
| 20 | 5 | 5 | 12.54 | 12.05 | - | 1.00 | 1.00 | - | - |
| 20 | 5 | 15 | 29.79 | 25.25 | - | 0.96 | 0.68 | - | - |
| 20 | 15 | 15 | 64.23 | 50.64 | - | 1.00 | 0.91 | - | - |
| 160 | 5 | 5 | 4.25 | 6.45 | 7.85 | 1.00 | 1.00 | 0.24 | 0.05 |
| 160 | 5 | 15 | 9.82 | 9.96 | 21.27 | 1.00 | 1.00 | 1.00 | 0.05 |
| 160 | 15 | 15 | 21.12 | 20.47 | - | 1.00 | 1.00 | - | - |
| 320 | 5 | 5 | 3.03 | 5.83 | 5.53 | 1.00 | 1.00 | 0.12 | 0.06 |
| 320 | 5 | 15 | 6.93 | 7.96 | 15.15 | 1.00 | 1.00 | 0.98 | 0.05 |
| 320 | 15 | 15 | 14.91 | 16.89 | 42.98 | 1.00 | 1.00 | 1.00 | 0.06 |

Columns labeled *err* show average spectral norm errors. Subscripts indicate separable correlation, separable covariance, and unrestricted estimators. Columns with label *rej* show empirical rejection rates. The subscripts indicate the null hypotheses covariance separability and correlation separability. A second subscript $b$ indicates the parametric bootstrap was used.

than 0.02.

## B.2 Additional simulations

Table B.1 shows simulation results similar to those of the lower panel of Table 3.1. In these simulations, $\mathcal{B} = 0$, $x_i = 1$ $(i = 1, \ldots, n)$, $U$ and $V$ have all off-diagonal entries set to $1/2$, and $W$ is diagonal with entries evenly spaced between 0.1 and 10. Qualitatively the findings are similar to those of Table 3.1 in that i) the separable covariance model does better in terms of spectral error when the sample size is small, ii) the separable correlation model does better when sample size is larger, and iii) the bootstrap based likelihood ratio test with $H_0$ : separable correlation has near nominal size in all settings where it is applicable.

### B.2.1 Convergence diagnostics

For each setting in Table B.2, 10000 datasets were generated from the separable correlation model, using the same configuration as in our other simulations. The model was then fit,

Table B.2: Convergence proportions

| $r$ | $c$ | $n$ | Conv. | Max | $V$ | $U$ |
|---|---|---|---|---|---|---|
| 2 | 2 | 2 | 0.18 | 0.00 | 0.82 | 0.01 |
| 2 | 2 | 4 | 0.76 | 0.24 | 0.00 | 0.00 |
| 2 | 2 | 10 | 1.00 | 0.00 | 0.00 | 0.00 |
| 2 | 4 | 2 | 0.01 | 0.00 | 0.99 | 0.00 |
| 2 | 4 | 4 | 0.07 | 0.92 | 0.01 | 0.00 |
| 2 | 4 | 10 | 1.00 | 0.00 | 0.00 | 0.00 |
| 15 | 15 | 2 | 0.01 | 0.00 | 0.91 | 0.07 |
| 15 | 15 | 4 | 0.00 | 1.00 | 0.00 | 0.00 |
| 15 | 15 | 10 | 1.00 | 0.00 | 0.00 | 0.00 |

Columns 4 – 7 are proportions out of 10,000 simulated datasets in which the algorithm terminated because it converged, reached the maximum number of iterations, an update of $V$ was singular, or an update of $U$ was singular, respectively.

using our algorithm, to each of these datasets. The maximum number of iterations was set to 1000 and the tolerance parameter $\epsilon$ in our algorithm to $10^{-10}$.

# Appendix C

# Convergence complexity analysis of a collapsed Gibbs sampler for Bayesian vector autoregressions

## C.1 Preliminary results

**Definition C.1.1.** We say that $X \in \mathbb{R}^{n\times m}$ has a matrix normal distribution with mean $M \in \mathbb{R}^{n\times m}$ and scale matrices $U \in \mathbb{S}_+^n$ and $V \in \mathbb{S}_+^m$ if $\text{vec}(X) \sim \mathcal{N}(\text{vec}(M), V \otimes U)$. We write $X \sim \mathcal{M}(M, U, V)$.

**Lemma C.1.1.** *If $X_i \in \mathbb{R}^{n\times m_i}$, $m_i \in \{1, 2, \dots\}$, $i = 1, 2, 3$, and $X = [X_1, X_2, X_3] \in \mathbb{R}^{n\times(m_1+m_2+m_3)}$ has full column rank, then with $\tilde{X}_i = Q_{X_2} X_i$, $i = 1, 2, 3$,*

1. *$X_1^\mathsf{T} Q_{X_2} X_1$ is invertible,*
2. *$X_1^\mathsf{T} Q_{[X_2,X_3]} X_1$ is invertible, and*
3. *$X_1^\mathsf{T} Q_{[X_2,X_3]} X_1 = \tilde{X}_1^\mathsf{T} Q_{\tilde{X}_3} \tilde{X}_1$.*

*Proof.* We start with 1. Suppose for contradiction that there exists $v \in \mathbb{R}^{m_1} \setminus \{0\}$ such that $X_1^\mathsf{T} Q_{X_2} X_1 v = 0$, which is equivalent to $Q_{X_2} X_1 v = 0$. This can happen either if $X_1 v = 0$, which contradicts the full column rank of $X$, or if $w = X_1 v$ is a non-zero vector in the column space of $X_1$ that also lies in the column space of $X_2$, which again contradicts the full column rank of $X$. The proof for 2 is exactly the same as that of 1 but with $[X_2, X_3]$

in place of $X_2$. Point 3 is an immediate consequence the Frisch–Waugh–Lovell theorem [34, Section 2.4], which says among other things that $Q_{\tilde{X}_3}\tilde{X}_1 = Q_{[X_2,X_3]}X_1$. □

**Lemma C.1.2.** *For any $X \in \mathbb{R}^{n\times p}$, $y \in \mathbb{R}^n$, and $c > 0$,*

$$\|(I_p c + X^\mathsf{T}X)^{-1}X^\mathsf{T}y\| \leq \|(X^\mathsf{T}X)^g X^\mathsf{T}y\|,$$

where superscript $g$ denotes an arbitrary generalized inverse.

*Proof.* Consider the optimization problem of minimizing $g_c : \mathbb{R}^p \to [0,\infty)$ defined by

$$g_c(b) := \|y - Xb\|^2 + c\|b\|^2.$$

If $c = 0$, then any $b$ such that $X^\mathsf{T}Xb = X^\mathsf{T}y$ is a solution. Thus, for any generalized inverse, $b_1 = (X^\mathsf{T}X)^g X^\mathsf{T}y$ solves the problem [31, Theorem 9.1.2]. On the other hand, if $c > 0$ then since $Ic + X^\mathsf{T}X$ has full rank, the unique solution is $b_2 = (cI + X^\mathsf{T}X)^{-1}X^\mathsf{T}y$. Now a contradiction arises if for some $c > 0$, $\|b_1\| < \|b_2\|$, which finishes the proof. □

**Lemma C.1.3.** *For $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{S}^n_{++}$, we have that $B^{-1}A^gB^{-1}$ is a generalized inverse of $BAB$, where superscript $g$ indicates a generalized inverse.*

*Proof.* We check the definition, namely that $BABB^{-1}A^gB^{-1}BAB = BAB$. Indeed, using that $AA^gA = A$, $BABB^{-1}A^gB^{-1}BAB = BAA^gAB = BAB$. □

## C.2 Main results

*Proof 4.2.1.* Under either of the two sets of conditions, $X$ has full column rank so $X^\mathsf{T}X$ is invertible and we may define $H_X = (X^\mathsf{T}X)^{-1}X^\mathsf{T}$, $P_X = XH_X$, and $Q_X = I_n - P_X$. Let

also $W = Y - Z\mathcal{A}$ and use $Q_X + P_X = I_n$ to write

$$\begin{aligned} f(Y \mid \mathcal{A}, \mathcal{B}, \Sigma) &\propto |\Sigma|^{-\frac{n}{2}} \operatorname{etr}\left(-\frac{1}{2}[X\mathcal{B} - W]^\mathsf{T}[X\mathcal{B} - W]\Sigma^{-1}\right) \\ &= |\Sigma|^{-\frac{n}{2}} \operatorname{etr}\left(-\frac{1}{2}[X\mathcal{B} - W]^\mathsf{T}(Q_X + P_X)[X\mathcal{B} - W]\Sigma^{-1}\right) \\ &= |\Sigma|^{-\frac{n}{2}} \operatorname{etr}\left(-\frac{1}{2}W^\mathsf{T}Q_X W\Sigma^{-1}\right) \operatorname{etr}\left(-\frac{1}{2}[\mathcal{B} - H_X W]^\mathsf{T}X^\mathsf{T}X[\mathcal{B} - H_X W]\Sigma^{-1}\right). \end{aligned} \tag{C.1}$$

The right-most term is the kernel of a matrix normal density for $\mathcal{B}$ with mean $H_X W$ and scale matrices $(X^\mathsf{T}X)^{-1}$ and $\Sigma$. Thus, integrating with respect to $\mathcal{B}$ gives,

$$\begin{aligned} \int f(Y \mid \mathcal{A}, \mathcal{B}, \Sigma)\,\mathrm{d}\mathcal{B} &\propto |\Sigma|^{-\frac{n}{2}} \operatorname{etr}\left(-\frac{1}{2}W^\mathsf{T}Q_X W\Sigma^{-1}\right)(2\pi)^{rp/2}|X^\mathsf{T}X|^{-r}|\Sigma|^{p} \\ &\propto |\Sigma|^{-\frac{n-p}{2}} \operatorname{etr}\left(-\frac{1}{2}W^\mathsf{T}Q_X W\Sigma^{-1}\right). \end{aligned}$$

Thus, to show that $f(Y \mid \mathcal{A}, \mathcal{B}, \Sigma)f(\alpha)f(\Sigma)$ can be normalized to a proper posterior, we need only show that

$$\iint |\Sigma|^{-\frac{n-p}{2}} \operatorname{etr}\left(-\frac{1}{2}W^\mathsf{T}Q_X W\Sigma^{-1}\right) f(\alpha)f(\Sigma)\,\mathrm{d}\alpha\,\mathrm{d}\Sigma < \infty. \tag{C.2}$$

Let us consider the two sets of conditions separately, starting with the first. Since

$$\operatorname{tr}(W^\mathsf{T}Q_X W\Sigma^{-1}) = \operatorname{tr}(\Sigma^{-1/2}W^\mathsf{T}Q_X W\Sigma^{-1/2}) \geq 0,$$

we can upper bound the integrand in (C.2) by

$$|\Sigma|^{-\frac{n-p}{2}} f(\alpha)f(\Sigma) = |\Sigma|^{-\frac{n+a-p}{2}} \operatorname{etr}\left(-\frac{1}{2}\Sigma^{-1}D\right) f(\alpha),$$

which since we are assuming that $n - p + a - r - 1 > r - 1$, i.e. that $n + a > 2r + p$ and that $D$ is SPD, is the product of a proper inverse Wishart and a proper density for $\alpha$. This finishes the proof for the first set of conditions.

For the second set of conditions, notice that for (C.2) it suffices, since $D$ is SPSD, and hence $f(\Sigma)$ and $f(\alpha)$ both bounded, to show that

$$\iint |\Sigma|^{-\frac{n+a-p}{2}} \operatorname{etr}\left(-\frac{1}{2}W^\intercal Q_X W\Sigma^{-1}\right) \mathrm{d}\alpha\, \mathrm{d}\Sigma < \infty.$$

Let $\tilde{Y} = Q_X Y$ and $\tilde{Z} = Q_X Z$ so that $Q_X W = \tilde{Y} - \tilde{Z}\mathcal{A}$. Using the same decomposition as before we have for the last integrand

$$\begin{aligned}
&|\Sigma|^{-\frac{n+a-p}{2}} \operatorname{etr}\left(-\frac{1}{2}W^\intercal Q_X W\Sigma^{-1}\right) \\
&= |\Sigma|^{-\frac{n+a-p-qr}{2}} \operatorname{etr}\left(-\frac{1}{2}\tilde{Y}^\intercal Q_{\tilde{Z}}\tilde{Y}\Sigma^{-1}\right) |\Sigma|^{-\frac{qr}{2}} \operatorname{etr}\left(-\frac{1}{2}[\mathcal{A} - H_{\tilde{Z}}\tilde{Y}]^\intercal \tilde{Z}^\intercal\tilde{Z}[\mathcal{A} - H_{\tilde{Z}}\tilde{Y}]\Sigma^{-1}\right).
\end{aligned}$$

Under the second set of assumptions, the last line is proportional to the product of an inverse Wishart density for $\Sigma$ with scale matrix $\tilde{Y}^\intercal Q_{\tilde{Z}}\tilde{Y}$ and $n+a-p-qr-r-1$ degrees of freedom and a matrix normal density for $\mathcal{A}$ with mean $H_{\tilde{Z}}\tilde{Y}$ and scale matrices $(\tilde{Z}^\intercal\tilde{Z})^{-1}$ and $\Sigma$, and hence integrable. The assumption that $[Y, X, Z]$ has full column ensures that, by Lemma C.1.1, $\tilde{Y}^\intercal Q_{\tilde{Z}}\tilde{Y}$ and $\tilde{Z}^\intercal\tilde{Z}$ are positive definite matrices. □

*Proof of Lemma 4.2.2.* The full conditional distribution of $\mathcal{B}$ is immediate from dropping terms not depending on $\mathcal{B}$ in (C.1). Consider next the integrand in (C.2). The first term in the exponential is $\operatorname{tr}([Y - Z\mathcal{A}]^\intercal Q_X[Y - Z\mathcal{A}]\Sigma^{-1}) = \|Q_X(Y - Z\mathcal{A})\Sigma^{-1/2}\|_F^2 = \|(\Sigma^{-1/2} \otimes I_n)(\operatorname{vec}(Q_X Y) - \operatorname{vec}(Q_X Z\mathcal{A}))\|^2 = \|(\Sigma^{-1/2} \otimes I_n)(\operatorname{vec}(Q_X Y) - [I_r \otimes Q_X Z]\alpha)\|^2$. Thus, the log of the integrand is quadratic as a function of $\alpha$, with Hessian $-B = -\Sigma^{-1} \otimes Z^\intercal Q_X Z - C$ and gradient $-(\Sigma^{-1} \otimes Z^\intercal Q_X)\operatorname{vec}(Q_X Y) - Cm$, which implies the desired distribution for $\alpha \mid \Sigma, Y$. Finally, the distribution of $\Sigma \mid \alpha, Y$ is immediate from dropping terms in the integrand in (C.2) not depending on $\Sigma$. □

*Proof Lemma 4.3.1.* Assume $\alpha^k, \mathcal{B}^k, \Sigma^k$, $k = 1, 2, \ldots$ are generated by the collapsed Gibbs sampler in Algorithm 4.1 started at some point $\theta^0 \in \Theta$. The equality follows from showing that $\xi^k = (\alpha^k, \Sigma^k)$ and $\theta^k$ are co-de-initializing Markov chains [62, Corollary 1]. That they are both Markov chains is clear from the construction of the updates in Algorithm 4.1. That

$\theta^k$ is de-initializing for $\xi^k$, i.e. that the distribution of $\xi^k \mid \theta^k, \xi^0$ does not depend on $\xi^0$, is immediate from that $\xi^k$ is a function (coordinate projection) of $\theta^k$. The other direction, that $\xi^k$ is de-initializing for $\theta^k$, is by construction of the algorithm: since $\xi^k$ is a coordinate projection of $\theta^k$, the distribution of $\theta^k \mid \xi^k, \theta^0$ is determined by that of $\mathcal{B}^k \mid \xi^k, \theta^0$, and the distribution from which this value is drawn (line 5, Algorithm 4.1) does not depend on $\theta^0$. □