
THE UNIVERSITY OF QUEENSLAND

AUSTRALIA

# Vector Semiparametric Generalized Linear Models

Kyle Macaskill
BMath/BCompSci

*A thesis submitted for the degree of BMath (Honours) at The University of Queensland in Year*
School of Mathematics and Physics

## Acknowledgments

To my supervisor, Dr Alan Huang, thank you for the guidance and care you have shown me throughout the year. I greatly appreciate the time you have put aside through regular consultations to help me grow, and for always answering my questions. It's because of your guidance and vast statistical knowledge that I was able to accomplish what I have this year.

To Sara, thank you for your continual love and support throughout the year. I really appreciate having you by my side while I tackled the challenges this year and for motivating me to do the best I can. Thank you for also being there to listen to my rambles about my assignments and thesis, and for all the little adventures we went on throughout the year.

To my roommates, Ben and Zander, thank you for the great laughs and good times. Living with my closest high school friends has been amazing and I'm gonna miss the times we spent together when I move away. Wherever I go, I'll always still be playing games and chat till the late hours of the night.

To Jacob, Adrian and Qingyuan, it's been a lot of fun doing honours together, cramming assignments and coursework while also being able to chat about our projects together. I wish you the best in your future academic or working endeavours.

To Dennis, Josh and Jasmine, thank you for being an amazing friendship group and for supporting me throughout the honours. It's been really nice catching up for ramen, the markets, bouldering, golf and lunch at Pantina. I hope we can keep in touch as we all go forward into our working careers.

Thank you to all the students and teaching team I worked with throughout the year in COSC2500, STAT2003, STAT2004, STAT3001, STAT3006 and STAT3500. I've had a blast tutoring and it's entirely because of the passionate people I teach and work with. Best of luck to everyone for the future.

Thank you also to Kenko and Market Kart for the lunches and coffee. I probably spent way too much money throughout the last few years on food at uni, but it was always something I looked forward to having every day.

Finally, to my Mum and Dad, thank you for the love and care throughout my uni life. It's tough living away from home for both of us but I'm really grateful every time we meet each other and for the time we spend together. You've been the best parents anyone could ask for, and I really appreciate that you both support me in chasing my dreams and working towards my future.

## Abstract

Generalized Linear Models (GLMs) can be extended to handle a vector of responses, but it can often be challenging to correctly specify the response's joint distribution conditional on the covariates, especially when there are a large number of components or mixed response types. The thesis aims to overcome misspecification of the response distribution by introducing a Vector Semiparametric Generalized Linear Model (VSPGLM), which is a multivariate extension of the Semiparametric Generalized Linear Model (SPGLM) introduced by Huang (2014). The model utilises an exponential tilt representation of the underlying distribution, which is then left unspecified and jointly estimated with the usual mean model coefficients. It has been shown that mean model coefficients and the distribution parameter are orthogonal and their joint estimation is asymptotically independent. The advantage of this approach is that it enforces minimal distributional assumptions while maintaining useful asymptotic properties, allowing it to be applied to a wide range of vector response regression problems. We show how to construct consistent estimators for both parameters using an empirical likelihood approach and establish joint asymptotic normality of the estimators. Methods for performing inference and hypothesis testing using both Wald tests and the Empirical Likelihood Ratio Tests (ELRT) via the empirical profile likelihood have also been shown. Finally, the VSPGLM has been implemented in `MATLAB` and applications of VSPGLM to a variety of vector response problems have been explored.

Firstly, Chapter 1 explores Generalized Linear Models, Vector Generalized Linear Models (VGLM) and their semiparametric extensions, as well as a review of empirical processes and semiparametric theory. Chapter 2 outlines the model and derivations of the score equations, along with how estimation is performed using maximum empirical likelihood estimation (MELE) and inference using either Wald tests or empirical likelihood ratio tests. Furthermore, generalizations of the asymptotic properties of the SPGLM model are given for VSPGLM. Chapter 3 provides an outline of the computational implementation of VSPGLM in `MATLAB`, along with syntax and numerical optimisations. Chapter 4 explores a simulation study to verify VSPGLM's asymptotic properties in a correctly specified example. Then, we explore fitting the model to a wide range of applications with a variety of different types of data structures and problems that are common in vector regression literature, comparing with other methods in the literature where appropriate. Chapter 5 then provides the technical details and derivations of all of the properties of VSPGLM given in Chapter 2. Finally, Chapter 6 concludes the thesis with a discussion of the advantages and current limitations of the model, with mentions of future directions where research could be conducted and for improvements to be made to the model.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

Regression analysis is used to model the relationship between a set of covariates and response variables, seeing application across a variety of domains such as natural sciences, economics, sociology and bio-medicine. The most common regression model is Linear Regression for continuous responses which are normally distributed, conditional on the covariates. However, for different conditional distributions of the response variable, there are different types of regression such as Logistic Regression for binary responses or Poisson Regression for count data which is conditionally Poisson. Generalized Linear Models (GLMs) introduced by Nelder and Wedderburn (1972) are a popular and invaluable tool in applied statistics, providing a unified modelling framework for handling these various response types. The widespread use of GLMs comes from the flexibility of being able to model a linear relationship on a link scale between covariates and response in an interpretable manner.

A constraint on the classical GLM framework is the requirement of a complete parametric specification of the response distribution. The correct specification of the distribution or a mean-variance relationship can be challenging, where misspecification leads to a loss of efficiency in parameter estimation as well as biased standard errors and inference of the parameters. This has led to the exploration of semiparametric methods (Wedderburn 1974, Liang and Zeger 1986) for fitting GLMs which do not require the specification of the underlying response distribution, allowing the data to be generated from some arbitrary distribution. Thus, semiparametric GLMs reduce the chance of misspecification while offering a flexible and parsimonious framework, allowing them to be applied broadly to a wide range of problems.

The classic GLM framework has also been extended to handle multivariate response data (Fahrmeir and Gerhard 2001, Yee 2015) through Vector Generalized Linear Models (VGLMs). The VGLM framework suffers from the same problem of misspecification, where correct parametric expression of the distribution is more complex as components may have mixed data types and now a covariance structure between the components needs to be considered. Thus, it is of interest to consider unified semiparametric frameworks for handling multivariate response data which includes mixed data and repeated measures, as data of this form is common in domains such as natural sciences and bio-medicine.

In this chapter, we introduce the classical GLM framework and how estimation is performed before exploring various semiparametric extensions in the GLM literature. We will then explore the generalisation of the framework to the multivariate setting by introducing VGLMs, and discuss the current semiparametric techniques for handling misspecification in VGLMs to motivate an alternative approach to semiparametric VGLMs. The model proposed in this thesis extends on the work by

Huang (2014) by utilising an exponential tilt formulation of the density and viewing the underlying distribution $F$ as an infinite dimensional parameter which is jointly estimated with the usual mean model parameters denoted by $\boldsymbol{\beta}$. Following the literature review, in section 1.2 we will review some key background concepts and definitions required for understanding the derivations and technical details of the model proposed in Chapter 2.

## 1.1 Literature Review

### 1.1.1 Generalized Linear Models (GLMs)

Let $(\boldsymbol{X}, Y) \in \mathcal{X} \times \mathcal{Y} \subset \mathbb{R}^q \times \mathbb{R}$, where the response variable $Y$ is independently sampled, conditional on the covariates $\boldsymbol{X}$ according to some design measure $G_{\boldsymbol{X}}$ on $\mathcal{X}$. Generalized Linear Models (McCullagh and Nelder 1989) have two key defining components, the first is that the conditional distribution of $Y$ given covariates $\boldsymbol{X}$ has a density with respect to a $\sigma$-finite measure of the form

$$dF(y; \theta, \phi) = \exp\left\{ \frac{b(\theta) + \theta y}{a(\phi)} + d(y, \phi) \right\}, \tag{1.1}$$

where $b(\theta)$ is the cumulant generating function, $\phi$ is a scale/dispersion parameter, $d(y, \phi) \geq 0$ is a normalising term and $\theta$ is the canonical parameter. Jørgensen (1987) refers to distributions with densities of the form (1.1) as coming from an exponential dispersion family. In the case where $\phi$ is known, the density is a part of the one-parameter exponential family (Definition 1.1). In cases where $\phi$ is unknown, it can be difficult to estimate so it's usually estimated by the method of moments estimator. The conditional distribution depends on $\boldsymbol{X}$ only via the linear predictor $\boldsymbol{X}^T\boldsymbol{\beta}$ where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_q)^T$ is a vector of unknown mean model parameters. The second component is the conditional mean model for the responses,

$$\mathbb{E}\left[Y|\boldsymbol{X}\right] = \mu(\boldsymbol{X}^T\boldsymbol{\beta}), \tag{1.2}$$

where $\mu(\cdot)$ is denoted as the mean function and $\mu^{-1}(\cdot)$ is the user-specified link function. For $n$ independent and identically distributed (i.i.d) observations $Y_1, \ldots, Y_n$, we estimate $\boldsymbol{\beta}$ via Maximum Likelihood (ML) estimation, where the ML estimate $\hat{\boldsymbol{\beta}}$ is the solution to the set of score equations

$$0 = \sum_{i=1}^{n} \boldsymbol{X}_i \, \mu'(\boldsymbol{X}_i^T\boldsymbol{\beta}) \left( \frac{Y_i - \mu(\boldsymbol{X}_i^T\boldsymbol{\beta})}{\mathrm{Var}[Y_i|\boldsymbol{X}_i]} \right), \tag{1.3}$$

typically solved using an iterative numerical method such as Iteriavely Reweighted Least Squares (IRLS). The ML estimate $\hat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ is unbiased, efficient and consistent, while also having asymptotically normality which allows for parameter inference using Wald tests.

**Theorem 1.1 (Asymptotic Normality of $\boldsymbol{\beta}$)** *As $n \to \infty$,*

$$\sqrt{n}\,(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \rightsquigarrow \mathcal{N}(\mathbf{0}, \Sigma) \tag{1.4}$$

*in $\mathbb{R}^q$, where $\mathcal{N}$ is a multivariate normal distribution with mean $\mathbf{0}$ and asymptotic covariance matrix $\Sigma$ is given by*

$$\Sigma = \left( \mathbb{E}^{\mathcal{X}} \left[ \frac{\mu'(\boldsymbol{X}^T\boldsymbol{\beta})^2 \boldsymbol{X}\boldsymbol{X}^T}{Var(Y|\boldsymbol{X})} \right] \right)^{-1}. \tag{1.5}$$

### 1.1.2 Semiparametric GLMs

A key constraint on the GLM framework is that the response variable's conditional distribution has to be correctly specified prior to fitting the model. Misspecification of the distribution leads to a loss of efficiency in parameter estimation and asymptotically biased inference on $\boldsymbol{\beta}$. Semiparametric extensions of the GLM framework loosen this constraint by not requiring the specification of the underlying distribution while maintaining asymptotically consistent estimates and correct inference of $\boldsymbol{\beta}$. As a result, the advantage of semiparametric methods is that they are more robust and can be applied to a broader range of problems, being useful in situations where the form of the data is complex. However, the disadvantage is that there is a loss in efficiency in parameter estimation relative to correct specified parametric GLMs, resulting in poor performance for small sample sizes.

An early example of a semiparametric method for fitting GLMs was explored by Wedderburn (1974), using Quasi-Likelihood (QL) methods. Quasi-likelihood methods are concerned with specifying the first two moments of the data, fitting regression models to problems where the variance of an observation $\text{Var}(Y|\boldsymbol{X}) = \phi V(\mu)$ is a function of its expectation $\mathbb{E}(Y) = \mu(\boldsymbol{X}^T\boldsymbol{\beta})$. For $n$ i.i.d observations $Y_1, Y_2, \ldots, Y_n$, parameter estimation is done by defining a nonparametric quasi-likelihood function through the formulation

$$Q(\mu_i; Y_i) = \int_{Y_i}^{\mu_i} \frac{Y_i - t}{\phi V(t)} \, \mathrm{d}t + g(Y_i) \tag{1.6}$$

or equivalently

$$U(\mu_i, Y_i) := \frac{\partial Q(\mu_i; Y_i)}{\partial \mu_i} = \frac{Y_i - \mu_i}{\phi V(\mu)} \tag{1.7}$$

where $\phi > 0$ is a dispersion parameter, $V(\mu) > 0$ is a given variance function and $g$ is some function of $\boldsymbol{Y}$. The full quasi-likelihood across the $n$ observations is formed by the summation

$$Q(\boldsymbol{\mu}, \boldsymbol{Y}) = \sum_{i=1}^{n} Q(\mu_i; Y_i), \tag{1.8}$$

without needing to assume any particular distribution of $Y$. Wedderburn (1974) showed that $Q$ has the following properties in common with log-likelihood functions, and is in fact equivalent if and only if $Y$ comes from a one-parameter exponential family.

$$\mathbb{E}\left[\frac{\partial Q(\mu; Y)}{\partial \mu}\right] = 0, \quad \text{Var}\left[\frac{\partial Q(\mu; Y)}{\partial \mu}\right] = \frac{1}{\phi \text{Var}(\mu)}, \quad -\mathbb{E}\left[\frac{\partial^2 Q(\mu; Y)}{\partial \mu^2}\right] = \frac{1}{\phi \text{Var}(\mu)} \, .$$

For $n$ i.i.d observations $Y_1, \ldots, Y_n$, estimation of $\boldsymbol{\beta}$ is then done by finding the quasi-score functions with respect to $\boldsymbol{\beta}$, given by

$$U(\boldsymbol{\beta}) := \sum_{i=1}^{n} \frac{\partial Q(\boldsymbol{\mu}_i, \boldsymbol{Y}_i)}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \frac{\partial \mu}{\partial \boldsymbol{\beta}} \frac{\partial Q(\mu_i; Y)_i}{\partial \mu_i} = \sum_{i=1}^{n} \frac{\partial \mu_i}{\partial \boldsymbol{\beta}} \frac{Y_i - \mu_i}{\phi V(\mu_i)} \, . \tag{1.9}$$

Then our estimate $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is the Quasi-Maximum Likelihood estimate, defined as the solution to $U(\hat{\boldsymbol{\beta}}) = 0$. The dispersion parameter however similar to classical GLMs is estimated using a moment estimator, namely

$$\hat{\phi} = \frac{1}{n - q} \sum_{i=1}^{n} \frac{(Y_i - \hat{\mu_i})^2}{V(\hat{\mu_i})} \tag{1.10}$$

where $q$ is the number of parameters in $\boldsymbol{\beta}$.

QL models are a more general approach to GLMs and provide consistent estimates for $\boldsymbol{\beta}$ as long as the mean model is correctly specified for the distribution, without needing to correctly specify the variance function (Crowder 1986). However, similar to parametric GLMs, the correct specification of the variance function is still a challenge and if this is misspecified there is a loss in asymptotic efficiency. This can be corrected via empirical (sandwich) variance estimators that provide asymptotically correct standard errors, although this can perform poorly even when the working variance function is correctly specified and the sample size is moderate to large (Kauermann and Carroll 2001). Adaptive quasi-likelihood methods have been explored, leaving the variance function unspecified and estimating it non-parametrically (Chiou and Müller 1999, Dewanji and Zhao 2002). However, QL models don't always correspond to actual probability models and thus there is an inability to make inferences about the cumulative distribution of the response.

To address the limitations of Quasi-likelihood and provide a flexible alternative when a full distribution model is desired, Rathouz and Gao (2009) introduced a new class of semiparametric generalized linear models (SPGLMs) which estimates nonparametrically a reference distribution for the response, and uses exponential tilting (Definition 1.2) of this reference distribution to yield a response model. More explicitly, the conditional distribution of the response $F_{\boldsymbol{\theta}}(y)$ is supposed to have a probability density

$$dF_\theta(y) = \exp\{b(\theta) + \theta y\}\, dF(y)\ , \tag{1.11}$$

where $\theta$ is the tilting parameter, $dF(y)$ is the baseline density and

$$b(\theta) = -\log\left\{\int_{\mathcal{Y}} \exp(\theta y) dF(y)\, dy\right\}\ . \tag{1.12}$$

The tilting parameter $\boldsymbol{\theta}$ is defined as the implicit solution to the mean constraint

$$\mathbb{E}\left[Y|\boldsymbol{X}\right] = \mu(\boldsymbol{X}^T\boldsymbol{\beta}) = \int_{\mathcal{Y}} y \exp(b\{\theta) + \theta y\} dF(y)\ dy\ . \tag{1.13}$$

Rathouz and Gao (2009) estimates the mean-model parameters $\boldsymbol{\beta}$ and the reference density $dF(y)$ using Maximum Likelihood Estimation for a univariate response $Y$, where the response space $\mathcal{Y}$ contains a finite number of values. Interestingly, in the event that $F$ is misspecified, poorly estimated or the tilting representation doesn't hold, the ML estimate $\hat{\boldsymbol{\beta}}$ is still generally consistent and asymptotically normal (Crowder 1986).

The exponential tilt model was then extended by Huang (2014) to consider a univariate response $Y$ where the response space $\mathcal{Y}$ is arbitrary, and the reference distribution $F$ is treated as an infinite dimensional parameter which is orthogonal to the mean model parameters $\boldsymbol{\beta}$. The joint estimation of $(\boldsymbol{\beta}, F)$ is instead done by Maximum Empirical Likelihood Estimation (MELE), by constructing an empirical log-likelihood and a coinciding profile empirical likelihood. The advantage of these frameworks over QL-based methods is that they remain within a full probability framework, where insights can be made about the underlying probabilistic generating mechanism that generated the data. Furthermore, these models have an explicit mean model which allows for a more immediate interpretation of $\boldsymbol{\beta}$ as mean contrasts, or treatment effects. We will explore the form of semiparametric GLMs based on exponential tilting in more detail in Chapter 2, as this is the motivating basis for the proposed model in the context of regression with vector responses.

Huang (2014) highlights the interesting connection between adaptive QL models with unspecified variance functions and GLMs with unspecified error distribution based on a result shown by Hiejima (1997). For any mean-variance relationship, there exists an exponential family whose score equations for $\boldsymbol{\beta}$ admit estimates that asymptotically are arbitrarily close to the estimates from corresponding QL

score equations. Thus, asymptotically any non-parametric QL model can be approximately arbitrarily well by a GLM with an unspecified error distribution $F$, where the mean-model parameter $\boldsymbol{\beta}$ and reference distribution $F$ are orthogonal.

### 1.1.3 Vector Generalized Linear Models (VGLMs)

A vector generalized linear model (VGLM) is an extension of a standard GLM to handle data of the form $\{(\boldsymbol{X}_i, \boldsymbol{Y}_i) \in \mathcal{X} \times \mathcal{Y} \subset \mathbb{R}^q \times \mathbb{R}^K : i = 1, 2, \ldots, n\}$. Therefore, the response variable $\boldsymbol{Y}$ is a $K$-dimensional vector and across the components we have a vector $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_q)^T$ of mean model parameters. Therefore, similar to a GLM there are conditional mean models for each component of the responses,

$$\mathbb{E}\left[Y_{(k)}|\boldsymbol{X}\right] = \mu_{(k)}\left(\boldsymbol{X}_{(k)}^T\boldsymbol{\beta}_{(k)}\right) \qquad k = 1, 2, \ldots, K, \tag{1.14}$$

where $\mu_{(k)}$ is the inverse of the link function $g_k(.)$ and $\boldsymbol{\beta}_{(k)}$ are the regression coefficients associated with the $k$-th mean model.

The formulation of the joint distribution of the vector response $\boldsymbol{Y}$ given covariates $\boldsymbol{X}$ in the parametric VGLM is different across various frameworks. Fahrmeir and Gerhard (2001) considers conditional distributions coming from a multivariate exponential dispersion model, where the joint density with respect to a $\sigma$-finite measure is of the form

$$dF(\boldsymbol{y}; \boldsymbol{\theta}, \boldsymbol{\phi}) = \exp\left\{\frac{b(\boldsymbol{\theta}) + c(\boldsymbol{\theta})\mathcal{S}(\boldsymbol{y})}{a(\boldsymbol{\phi})} + d(\boldsymbol{y}, \boldsymbol{\phi})\right\}, \tag{1.15}$$

where $\mathcal{S}(\boldsymbol{y})$ is a $q \times 1$ dimensional sufficient statistic ($q \geq d$), $c(\boldsymbol{\theta})$ is a $q \times 1$ dimensional canonical parameter, $b(\boldsymbol{\theta})$ is the c.g.f, $a(\boldsymbol{\phi}) \geq 0, d(y, \boldsymbol{\phi}) \geq 0$, and $\phi$ is a dispersion parameter. Fahrmeir and Gerhard (2001) suppose no constraints on $\boldsymbol{\beta}$ between the components and use the maximum likelihood estimation to derive an estimate for $\boldsymbol{\beta}$ in analogy to the univariate GLM. Suppose that $\eta_i = \boldsymbol{X}_i^T\boldsymbol{\beta}$ where $\boldsymbol{X}_i$ is the $q \times K$, and $g(\boldsymbol{\mu}_i) = \eta_i = \boldsymbol{X}_i^T\boldsymbol{\beta}$ where $g : \mathbb{R}^K \to \mathbb{R}^K$. Note that the structure of the design matrix $\boldsymbol{X}_i$ changes according to the problem. Then over $n$ i.i.d observations, the ML estimate $\hat{\boldsymbol{\beta}}$ is the root to the score equation for $\boldsymbol{\beta}$ is given by

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} D_i(\boldsymbol{\beta})\Sigma_i^{-1}(\boldsymbol{\beta})\left(\boldsymbol{Y}_i - \boldsymbol{\mu}(\boldsymbol{X}_i^T\boldsymbol{\beta})\right) \ , \tag{1.16}$$

where $\Sigma_i(\boldsymbol{\beta}) = \text{Cov}(\boldsymbol{Y}_i|\boldsymbol{X}_i)$ and $D_i(\boldsymbol{\beta}) := \frac{\partial \boldsymbol{\mu}(\boldsymbol{X}_i^T\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$. Furthermore, the expected Fisher information is given by

$$I(\boldsymbol{\beta}) = \mathbb{E}\left[S(\boldsymbol{\beta})S(\boldsymbol{\beta})^T\right] = \sum_{i=1}^{n} D_i(\boldsymbol{\beta})\Sigma_i^{-1}(\boldsymbol{\beta})\boldsymbol{D}_i(\boldsymbol{\beta})^T \ , \tag{1.17}$$

and asymptotic normality of the ML estimate $\hat{\boldsymbol{\beta}}$ is given under regularity assumptions and the assumption of correct specification,

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \rightsquigarrow \mathcal{N}(\boldsymbol{0}, I^{-1}(\boldsymbol{\beta})) \ . \tag{1.18}$$

The proof of asymptotic normality is analogous to the derivations for the univariate GLM and can be found in Appendix A.2 of Fahrmeir and Gerhard (2001).

Another parametric approach to VGLMs is through the use of Gaussian Couplas explored by Song (2007). Suppose that the conditional distribution of $\boldsymbol{Y}$ given $\boldsymbol{X}$ is given by the exponential dispersion

model (1.1) and let $G_j(y_{(j)};\mu_{(j)},\phi_{(j)})$ denote that $j$-th marginal cumulative distribution function (CDF) where $\mu_{(j)}=\rho(\theta), \text{Var}(\boldsymbol{Y})=\phi\text{Var}(\boldsymbol{\mu})$. A mapping $C:(0,1)^n\to(0,1)$ is called a copula if it's a continuous distribution function and each margin is a uniform distribution. Thus, a joint CDF for the multivariate exponential dispersion model can be constructed by the Gaussian Coupla of the form

$$F(\boldsymbol{y};\boldsymbol{\mu},\boldsymbol{\phi},\Gamma)=C\left\{G_1(y_{(1)};\mu_{(1)},\phi_{(1)}),\ldots,G_m(y_{(m)};\mu_{(m)},\phi_{(m)})|\Gamma(\boldsymbol{\alpha})\right\} \tag{1.19}$$

where $C(\cdot)$ is a Gaussian coupla with CDF given by

$$C(\boldsymbol{u}|\Gamma)=\Phi_m\left\{\Phi^{-1}(u_{(1)}),\ldots,\Phi^{-1}(u_{(m)})|\Gamma(\boldsymbol{\alpha})\right\}\ . \tag{1.20}$$

Above, $\Phi_m$ is the CDF of a $m$-variate normal with mean 0, correlation matrix $\Gamma(\boldsymbol{\alpha})$ and $\Phi$ is the standard normal CDF. As a result, by finding the associated density and establishing a VGLM with this conditional distribution, the VGLM yields a large class of multivariate regression models for both discrete, continuous and mixed data. This joint density has to be constructed for each particular problem, and then the vector of parameters $\boldsymbol{\psi}=(\boldsymbol{\beta},\boldsymbol{\phi},\boldsymbol{\alpha})$ is estimated using by Maximum Likelihood Estimation. Rather than finding a closed form for the MLE which requires analytically finding the second-order derivative of the log-likelihood, it is typically numerically estimated using a Gauss-Newton type algorithm

$$\boldsymbol{\psi}^{k+1}=\boldsymbol{\psi}^k+\epsilon\left(\frac{1}{n}\sum_{i=1}^n\dot{\ell}_i(\boldsymbol{\psi}^k;\boldsymbol{Y}_i,\boldsymbol{X}_i)\dot{\ell}_i(\boldsymbol{\psi}^k;\boldsymbol{Y}_i,\boldsymbol{X}_i)^T\right)^{-1}\dot{\ell}(\boldsymbol{\psi}^k) \tag{1.21}$$

where $\epsilon$ is a step term, $\dot{\ell}(\psi)$ is the derivative of the log-likelihood function given by

$$\ell(\boldsymbol{\psi};\boldsymbol{Y},\boldsymbol{X})=\sum_{i=1}^n\ell_i(\boldsymbol{\psi};\boldsymbol{Y}_i,\boldsymbol{X}_i).$$

The use of Gaussian Couplas allows for the modelling of both positive and negative correlations as well as responses with mixed types, and if all margins are Gaussian it reduces back to a multivariate normal regression whereas other coupla methods do not. An example of how to construct a VGLM using Gaussian Couplas will be further explored in Chapter 4.

More recently, Yee (2015) establishes a broader VGLM framework where the conditional distribution of $\boldsymbol{Y}$ is given by

$$f(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\beta},\phi)=f(\boldsymbol{y},\boldsymbol{\mu},\phi), \tag{1.22}$$

for some known joint density $f(\cdot)$. Thus, the joint density doesn't necessarily need to originate from a multivariate exponential family, as multivariate generalizations of known 1-parameter exponential families are difficult to express in a parametric form. As a result of leaving the form of the joint density unspecified, there is no general form for the score equations for $\boldsymbol{\beta}$, and instead maximum likelihood estimates are obtained using Iteratively Reweighted Least Squares (IRLS). For models of the form (1.22) where $\eta_k=\boldsymbol{X}^T\boldsymbol{\beta}_{(k)}$, the log-likelihood can be expressed in the form

$$\ell(\boldsymbol{\beta})=\sum_{i=1}^n w_i\ell_i(\eta_1,\ldots,\eta_{K'}) \tag{1.23}$$

where $w_i$ are known fixed positive prior weights. Then the iterative update for $\boldsymbol{\beta}^{(a-1)}$ is given as

$$\boldsymbol{\beta}^{(a)}=\boldsymbol{\beta}^{(a-1)}+\left(\boldsymbol{X}_{\text{VLM}}^T\boldsymbol{W}^{(a-1)}\boldsymbol{X}_{\text{VLM}}\right)^{-1}\boldsymbol{X}_{\text{VLM}}^T\boldsymbol{W}^{(a-1)}\boldsymbol{z}^{(a)} \tag{1.24}$$

where

$$\boldsymbol{W}_i = -\mathbb{E}\left(\frac{\partial^2 \ell_i}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}}\right), \ \boldsymbol{u}_i = \frac{\partial \ell_i}{\partial \eta}, \ \boldsymbol{z}_i^{a-1} = \boldsymbol{X}_i \boldsymbol{\beta}^{(a-1)} + \boldsymbol{W}_i^{-1(a-1)} \boldsymbol{u}_i^{a-1}, \tag{1.25}$$

and $\boldsymbol{X}_{\text{VLM}}$ is a block matrix in terms of constraint matrices $\mathbf{H}_k$ for $k = 1, 2, \ldots, K'$ where there are $K'$ mean models. More explicitly, if for $i = 1, 2, \ldots, n$ and $k = 1, 2, \ldots, K'$, denoting $x_{ikj}$ as the $j$-th value of $x_k$ for observation i,

$$\boldsymbol{X}_{ik}^{\#} = \text{diag}(x_{ik1}, \ldots, x_{ikK'}) \tag{1.26}$$

$$\boldsymbol{X}_{\text{VLM}} = \begin{pmatrix} \boldsymbol{X}_{11}^{\#}\mathbf{H}_1 & \ldots & \boldsymbol{X}_{1K'}^{\#}\mathbf{H}_{K'} \\ \vdots & & \vdots \\ \boldsymbol{X}_{n1}^{\#}\mathbf{H}_1 & \ldots & \boldsymbol{X}_{nK'}^{\#}\mathbf{H}_{K'} \end{pmatrix}. \tag{1.27}$$

Therefore, the use of constraint matrices in the estimation procedure allows for constraints to be placed on the coefficients between the components, creating a very flexible VGLM framework. As the estimates $\hat{\boldsymbol{\beta}}$ from the model are maximum likelihood estimates, they satisfy the standard asymptotic properties from VGLMs such as consistency and asymptotic normality.

### 1.1.4 Semiparametric VGLMs

The main challenge with the parametric VGLMs presented by Fahrmeir and Gerhard (2001), Song (2007) and Yee (2015) is that they also require the correct specification of the joint distribution, the mean-variance relationship of each marginal distribution or covariance structure between the components. Therefore, it is also of interest to consider semiparametric extensions of VGLM frameworks, or vector response generalisations of existing semiparametric frameworks to overcome misspecification.

An example of a semiparametric VGLM are models that use Generalized Estimating Equations (GEEs) introduced by Liang and Zeger (1986). GEEs in general vector regression problems are able to deal with longitudinal data and data where marginal distributions are of mixed types (eg. continuous-discrete response pairs). The advantage of using GEEs is that they can create robust models that account for within-vector correlations without requiring a correct specification for the model's variances, correlations or underlying joint distribution (Huang 2017). Similar to QL methods, the vector component of the response vector $\boldsymbol{Y}$ is assumed to have a mean-variance relationship given by

$$\mathbb{E}\left[Y_{(k)}|\boldsymbol{X}_{(k)}\right] = \mu_{(k)} = \mu_{(k)}(\boldsymbol{X}_{(k)}^T \boldsymbol{\beta}_{(k)}) \tag{1.28}$$

$$\text{Var}\left[Y_{(k)}|\boldsymbol{X}_{(k)}\right] = \sigma_{(k)}^2 = \phi_{(k)} V_{(k)}(\mu_{(k)}) \tag{1.29}$$

for $k = 1, 2, \ldots, K$ where $\boldsymbol{\beta}_{(k)}$ and $\phi_{(k)}$ are the respectively the mean-model and dispersion parameters for each component. Additionally, a within-vector correlation is specified between any two components

$$\text{cor}(Y_{(k_1)}, Y_{(k_2)}|\boldsymbol{X}) = \rho_{k_1 k_2}(\gamma; \mu_{(k_1)}, \mu_{(k_2)}) \tag{1.30}$$

depending on a vector of correlation parameters $\boldsymbol{\gamma}$. Similar to the constraints seen in a VGLM framework, GEEs use constraints on the mean-model parameters to create a longitudinal framework. To perform estimation, a working covariance matrix $\boldsymbol{W}_i$ is constructed for each observation where

$$\boldsymbol{W}_i = \text{Diag}(\boldsymbol{\sigma}_i^2)^{1/2} \boldsymbol{R}_i(\boldsymbol{\gamma}) \text{Diag}(\boldsymbol{\sigma}_i^2)^{1/2} \tag{1.31}$$

where $\boldsymbol{\sigma}_i = (\sigma_{i1}, \ldots, \sigma_{iK})^T$ and $\boldsymbol{R}_i(\gamma)$ is the correlation matrix estimated by $\boldsymbol{\gamma}$. Then, an estimate for $\boldsymbol{\beta}$ similar to QL-methods is defined as the solution to the set of generalized estimating equations

$$\mathbf{0} = \sum_{i=1}^{n} D_i(\boldsymbol{\beta}) \boldsymbol{W}_i^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}_i) \tag{1.32}$$

where $D_i(\boldsymbol{\beta}) = \frac{\partial \boldsymbol{\mu}(\boldsymbol{X}_i^T\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$. Furthermore, supposing that $\boldsymbol{\beta}_{(k)} \in \mathbb{R}^{q_k}$, we can estimate the dispersion parameter $\phi_{(k)}$ via

$$\hat{\phi}_{(k)}^{-1} = \frac{1}{n - q_k} \sum_{i=1}^{n} \frac{(Y_{i(k)} - \hat{\mu}_{i(k)})^2}{V_{(k)}(\hat{\mu}_{i(k)})}, \; k = 1, \ldots, K \; . \tag{1.33}$$

Furthermore, in Theorem 2 of Liang and Zeger (1986), asymptotic normality of the GEE estimator is established, whereas $n \to \infty$, $\sqrt{n}\,(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \rightsquigarrow N(\mathbf{0}, V_{\boldsymbol{\beta}})$, where,

$$V_{\boldsymbol{\beta}} = \lim_{n\to\infty} n \left( \sum_{i=1}^{n} D_i^T \boldsymbol{W}_i^{-1} D_i \right)^{-1} \left\{ \sum_{i=1}^{n} D_i^T \boldsymbol{W}_i^{-1} \mathrm{Cov}(\boldsymbol{Y}_i) \boldsymbol{W}_i^{-1} D_i \right\} \left( \sum_{i=1}^{n} D_i^T \boldsymbol{W}_i^{-1} D_i \right)^{-1} \tag{1.34}$$

The asymptotic variance $V_{\boldsymbol{\beta}}$ can be consistently estimated by the plug-in sandwich estimator

$$\hat{V}_{\boldsymbol{\beta}} = n \left( \sum_{i=1}^{n} \hat{D}_i^T \hat{W}_i^{-1} \hat{D}_i \right)^{-1} \left\{ \sum_{i=1}^{n} \hat{D}_i^T \hat{\boldsymbol{W}}_i^{-1} \hat{\Sigma}_{\boldsymbol{Y}} \hat{\boldsymbol{W}}_i^{-1} \hat{D}_i \right\} \left( \sum_{i=1}^{n} \hat{D}_i^T \hat{\boldsymbol{W}}_i^{-1} \hat{D}_i \right)^{-1} \tag{1.35}$$

where $\hat{\Sigma}_{\boldsymbol{Y}_i} = (\boldsymbol{Y}_i - \hat{\boldsymbol{\mu}}_i)(\boldsymbol{Y}_i - \hat{\boldsymbol{\mu}}_i)^T$. However, similar to QL methods the standard errors coming from a sandwich estimator for variance can perform poorly even when a correct working variance function is specified for moderate to large sample sizes (Kauermann and Carroll 2001).

Another example of a semiparametric VGLM is the multivariate density ratio model (MDRM) proposed by Marchese and Diao (2017), extending the univariate density ratio model explored by Luo and Tsai (2012) to handling vector responses. The univariate density ratio models assumes the canonical parameter $c(\boldsymbol{\theta})$ in an exponential family has a linear-predictor for $\boldsymbol{x}^T\boldsymbol{\beta}$ and sufficient statistic $\mathcal{S}(\boldsymbol{y}) \equiv \boldsymbol{y}$, giving densities of the form

$$dF(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\beta}) = \frac{\exp\{(\boldsymbol{x}^T\boldsymbol{\beta})\, y\}}{\int_{\mathcal{Y}} \exp\{(x^T\boldsymbol{\beta})\, y\}\, \mathrm{d}F(\boldsymbol{y})} dF(\boldsymbol{y}), \tag{1.36}$$

where $\boldsymbol{y} \in \mathbb{R}$, $\boldsymbol{x} \in \mathbb{R}^q$, $\boldsymbol{\beta} \in \mathbb{R}^q$. Then similar to Rathouz and Gao (2009) and Huang (2014), the unknown density $dF(\boldsymbol{y})$ is jointly estimated along with the mean-model parameters $\boldsymbol{\beta}$, via non-parametric maximum likelihood estimation (NPMLE). In MDRM the joint distribution for a vector response $\boldsymbol{y} \in \mathbb{R}^K$ is of the form

$$dF(\boldsymbol{y}|\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{\beta}, F) = \frac{dF_0(\boldsymbol{y}) \exp\{\sum_{k=1}^{K} h_k(y_k)(\boldsymbol{x}^T\boldsymbol{\beta})_k\}}{\int_{\mathcal{Y}} \exp\{\sum_{k=1}^{K} h_k(z_k)(\boldsymbol{x}^T\boldsymbol{\beta})_k\} dF_0(\boldsymbol{z})} \tag{1.37}$$

where $\boldsymbol{x} \in \mathbb{R}^q, \boldsymbol{\beta} \in \mathbb{R}^{q\times K}$ and $dF_0$ is the baseline CDF when $\boldsymbol{x} = 0$, and $h_k$ is a function which maps $y_k$ to the sufficient statistic scale. For estimation, an empirical likelihood is considered, discretising the baseline density $f \equiv dF_0(\boldsymbol{y})$ at unique observed values $\boldsymbol{u}_m$ with an associated multiplicity term $\alpha_m$ for $m = 1, \ldots, M$. The empirical log-likelihood for $(\boldsymbol{\beta}, f)$ is given by

$$\ell_n(\boldsymbol{\beta}, f) = \sum_{m=1}^{M} \alpha_m \log\{f_m\} + \sum_{i=1}^{n} h(\boldsymbol{y}_i) \cdot \boldsymbol{x}_i^T - \sum_{i=1}^{n} \log \left\{ \sum_{m=1}^{M} f_m \exp\big(\boldsymbol{u}_m \cdot \boldsymbol{x}_i^T\boldsymbol{\beta}\big) \right\} . \tag{1.38}$$

Thus, the empirical log-likelihood is maximised to find the NPMLE's, where for a fixed $\boldsymbol{\beta}$,

$$\hat{f}_z = n_z \left\{ \sum_{i=1}^{n} \frac{\exp\big(\boldsymbol{u}_z \cdot \boldsymbol{x}_i^T\boldsymbol{\beta}\big)}{\sum_{m=1}^{M} f_m \exp\big(\boldsymbol{u}_m \cdot \boldsymbol{x}_i^T\boldsymbol{\beta}\big)} \right\}^{-1}, \quad z = 1, \ldots, M$$

The framework admits consistent estimates as well as joint asymptotic normality of the estimates, namely $\sqrt{n}\,(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*, \hat{F} - F^*)$ converges weakly to a tight zero-mean Gaussian process, where $\hat{F}$ is

an empirical CDF with jumps of size $\hat{f}_z$. As a result of these asymptotic properties, inference of the parameters $\hat{\boldsymbol{\beta}}$ can be performed using both Wald test and empirical likelihood ratio tests. These asymptotic properties are shared with similar univariate models such as Luo and Tsai (2012) and Huang (2014), but now hold in the setting of vector regression. However, the mean-model parameters $\boldsymbol{\beta}$ are modelled on the canonical scale and as a result must be re-scaled using an estimate for the asymptotic covariance matrix for $\boldsymbol{\beta}$, $\hat{\Sigma}$. For simple cases such as having multivariate normally distributed responses, the regression parameters on the mean scale can be recovered by $\hat{\boldsymbol{\beta}}\hat{\Sigma}^{-1}$, but more generally it can be challenging to find this one-to-one correspondence to the parameters on the mean scale, which can be seen in some examples presented in Marchese and Diao (2017). Thus, MDRM performing regression on the scale of the canonical parameters results in fewer parameters to optimize, but has a downside of interpretability compared to regressing on the mean function $\boldsymbol{\mu}$ as is done by Rathouz and Gao (2009) and Huang (2014). This is because regression on the mean $\boldsymbol{\mu}$ allows for interpretation of the changes in the mean-model parameters $\boldsymbol{\beta}$ independent of the reference distribution $F$.

Therefore, the frameworks above motivate a semiparametric VGLM framework which can be applied to a variety of scenarios with minimal assumptions and produce an appropriate interpretable model that remains within a full probability framework. Thus, the focus of this thesis is to extend the framework of semiparametric GLMs based on exponential tilting explored by Rathouz and Gao (2009) and Huang (2014) to a VGLM framework, which we will introduce in the following chapter.

## 1.2 Preliminaries

In order to derive the proposed model and prove its asymptotic properties, we utilise concepts from empirical processes and semiparametric theory and draw on definitions from functional analysis and exponential family literature. We will provide a brief review of the concepts and definitions used in the derivations given in Chapter 2 and 5.

### 1.2.1 Exponential Family and Exponential Tilting

As mentioned in section 1.1 for GLMs it's assumed that the distribution of the response $\boldsymbol{Y}$ conditional on the covariates $\boldsymbol{X}$ comes from an exponential dispersion family. A closely related family of distribution that we use as a part of the proposed model is the exponential family, which is formally defined below.

**Definition 1.1 (Exponential Family, Jørgensen and Labouriau 1992)** *Suppose a measurable space $(\mathcal{X}, \mathcal{G})$, a random variable $\boldsymbol{Y}$ and a reference probability measure $\nu$. A family of probability measures $\{F_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta\}$ indexed by a d-dimensional parameter $\boldsymbol{\theta}$ is called an d-parameter exponential family if there exists a $\sigma$-finite dominating measure $\upsilon$ such that the density of $F_{\boldsymbol{\theta}}$ with respect to $\upsilon$ has the form*

$$\frac{\mathrm{d}F_{\boldsymbol{\theta}}}{\mathrm{d}\upsilon}(\boldsymbol{y}) = h(\boldsymbol{y}) \exp\{b(\boldsymbol{\theta}) + c(\boldsymbol{\theta})^T \mathcal{S}(\boldsymbol{y})\}, \tag{1.39}$$

*where $\mathcal{S}(\boldsymbol{y})$ is a $q \times 1$ dimensional sufficient statistic ($q \geq d$), $c(\boldsymbol{\theta})$ is a $q \times 1$ dimensional canonical parameter, $h(\boldsymbol{y})$ is a non-negative function often called the base measure and $b(\boldsymbol{\theta})$ is the distribution's cumulant generating function (c.g.f) defined as*

$$b(\boldsymbol{\theta}) = -\log\left\{\int_{\mathcal{Y}} h(\boldsymbol{y}) \exp\left(c(\boldsymbol{\theta})^T \mathcal{S}(\boldsymbol{y})\right) \mathrm{d}\upsilon(\boldsymbol{y})\right\}. \tag{1.40}$$

*If the reference measure $\nu$ is dominated by $\upsilon$ with a density given by $h(\boldsymbol{y})$,*

$$\frac{\mathrm{d}\nu}{\mathrm{d}\upsilon}(\boldsymbol{y}) = h(\boldsymbol{y}),$$

*then $h(\boldsymbol{y})$ is absorbed into the dominating measure and $F_{\boldsymbol{\theta}}$ has density functions with respect to $\nu$ of the form*

$$\frac{\mathrm{d}F_{\boldsymbol{\theta}}}{\mathrm{d}\nu}(\boldsymbol{y}) = \exp\{b(\boldsymbol{\theta}) + c(\boldsymbol{\theta})^T \mathcal{S}(\boldsymbol{y})\}, \tag{1.41}$$

*where the c.g.f is defined as*

$$b(\boldsymbol{\theta}) = -\log\left\{\int_{\mathcal{Y}} \exp\left\{c(\boldsymbol{\theta})^T \mathcal{S}(\boldsymbol{y})\right\} \mathrm{d}\nu(\boldsymbol{y})\right\} . \tag{1.42}$$

*The parameter space $\boldsymbol{\Theta}$ is a d-dimensional convex set such that (1.41) defines a density for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, which occurs when $\mathbb{E}\left[\exp\{c(\boldsymbol{\theta})^T \mathcal{S}(\boldsymbol{y})\}\right] < \infty$.*

The reference probability measure $\nu$ is a Lebesgue–Stieltjes measure associated with a reference distribution $F$ which is a cumulative distribution function (CDF), where

$$\nu\left((\boldsymbol{y}, \boldsymbol{y} + \mathrm{d}\boldsymbol{y}]\right) = F(\boldsymbol{y} + \mathrm{d}\boldsymbol{y}) - F(\boldsymbol{y}).$$

We follow the convention of expressing the Lebesgue-Stieltjes integral of a measurable function $g$ as being with respect to $F$,

$$\int_{\mathcal{Y}} g(\boldsymbol{y}) \exp\{b(\boldsymbol{\theta}) + c(\boldsymbol{\theta})^T \mathcal{S}(\boldsymbol{y})\} \, \mathrm{d}\nu(\boldsymbol{y}) = \int_{\mathcal{Y}} g(\boldsymbol{y}) \exp\{b(\boldsymbol{\theta}) + c(\boldsymbol{\theta})^T \mathcal{S}(\boldsymbol{y})\} \, \mathrm{d}F(\boldsymbol{y}) . \tag{1.43}$$

Supposing that $F_{\boldsymbol{\theta}}$, $F$ are absolutely continuous with respect to some appropriate common measure $\lambda$, we can equivalently express the exponential family as a family of probability measures $\{F_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta\}$ with densities $dF_{\boldsymbol{\theta}}$ of the form

$$dF_{\boldsymbol{\theta}}(\boldsymbol{y}) = \exp\{b(\boldsymbol{\theta}) + c(\boldsymbol{\theta})^T \mathcal{S}(\boldsymbol{y})\} dF(\boldsymbol{y}) . \tag{1.44}$$

Furthermore as given by Morris (1982), if $b(\boldsymbol{\theta})$ on $\Theta$ exists, then $\{F_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta\}$ defined by

$$dF_{\boldsymbol{\theta}}(\boldsymbol{y}) = \exp\{b(\boldsymbol{\theta}) + \boldsymbol{\theta}^T \boldsymbol{y}\} dF(\boldsymbol{y}), \tag{1.45}$$

forms a *natural exponential family.*

There are many useful properties of the exponential family and Rathouz and Gao (2009) explored utilise these by re-expressing the density of the response variable as a natural exponential family via exponential tilting of the reference distribution $F$.

**Definition 1.2 (Tilted Exponential Family, Hiejima 1997)** *Suppose a measurable space $(\mathcal{X}, \mathcal{G})$, let $\boldsymbol{Y}$ be a random variable, $\boldsymbol{\theta} \in \Theta$ denote the tilting parameter and suppose $\mathcal{P}_{\boldsymbol{\theta}}$ is a set of probability models where for $\nu \in \mathcal{P}_{\boldsymbol{\theta}}$,*

$$\mathbb{E}\left[\exp\left\{\boldsymbol{\theta}^T \boldsymbol{y}\right\}\right] = \int_{\mathcal{Y}} \exp\left\{\boldsymbol{\theta}^T \boldsymbol{y}\right\} \, \mathrm{d}\nu(\boldsymbol{y}) < \infty. \tag{1.46}$$

*Then, a tilted exponential family $\{F_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ has densities with respect to $\nu$ of the form*

$$\frac{\mathrm{d}F_{\boldsymbol{\theta}}}{\mathrm{d}\nu}(\boldsymbol{y}) = \exp\{b(\boldsymbol{\theta}) + \boldsymbol{\theta}^T \boldsymbol{y}\}, \tag{1.47}$$

*where $b(\boldsymbol{\theta})$ is the c.g.f of $\boldsymbol{y}$,*

$$b(\boldsymbol{\theta}) = -\log\left\{\int_{\mathcal{Y}} \exp\left(\boldsymbol{\theta}^T \boldsymbol{y}\right) \mathrm{d}\nu(\boldsymbol{y})\right\} . \tag{1.48}$$

Thus, supposing absolute continuity with respect to a common measure $\lambda$, we can equivalently represent the exponentially tilted density $dF_{\boldsymbol{\theta}}$ to be of the form

$$dF_{\boldsymbol{\theta}}(\boldsymbol{y}) = \exp\left\{b(\boldsymbol{\theta}) + \boldsymbol{\theta}^T \boldsymbol{y}\right\} dF(\boldsymbol{y}) . \tag{1.49}$$

Note that if we fix this reference distribution $F$, $F_{\boldsymbol{\theta}}$ belongs to the natural exponential family (1.45).

### 1.2.2 Stochastic Convergence

Consider a random vector $\boldsymbol{X} = (X_1, X_2, \ldots, X_p) \in \mathbb{R}^p$ of real random variables. Then a sequence of random vectors $\boldsymbol{X}_n$ converges in distribution, or converges weakly to a random vector $\boldsymbol{X}$ if

$$\lim_{n\to\infty} \mathbb{P}(\boldsymbol{X}_n \leq \boldsymbol{x}) \to \mathbb{P}(\boldsymbol{X} \leq \boldsymbol{x}) \tag{1.50}$$

for all $\boldsymbol{x}$ at which the distribution function $\boldsymbol{x} \mapsto \mathbb{P}(\boldsymbol{X} \leq \boldsymbol{x})$ is continuous. We denote this as $\boldsymbol{X}_n \xrightarrow{d} \boldsymbol{X}$ or $\boldsymbol{X}_n \rightsquigarrow \boldsymbol{X}$, where throughout the thesis we adopt the latter. There are various equivalent descriptions of weak convergence which are given in the Portmanteau Theorem (Lemma 2.2 Van Der Vaart 1998), stating that the following statements are equivalent

(i) $\boldsymbol{X}_n \rightsquigarrow \boldsymbol{X}$

(ii) $\mathbb{E}\left[f(\boldsymbol{X}_n)\right] \to \mathbb{E}\left[f(\boldsymbol{X})\right]$ for all bounded continuous functions $f$

(iii) $\mathbb{E}\left[f(\boldsymbol{X}_n)\right] \to \mathbb{E}\left[f(\boldsymbol{X})\right]$ for all bounded Lipschitz functions $f$

(iv) $\liminf \mathbb{E}\left[f(\boldsymbol{X}_n)\right] \geq \mathbb{E}\left[f(\boldsymbol{X})\right]$ for all nonnegative, continuous functions $f$

(v) $\liminf \mathbb{P}\left(\boldsymbol{X}_n \in G\right) \geq \mathbb{P}(\boldsymbol{X} \in G)$ for every open set $G$

(vi) $\limsup \mathbb{P}\left(\boldsymbol{X}_n \in F\right) \leq \mathbb{P}\left(\boldsymbol{X}_n \in F\right)$ for every closed set $F$

(vii) $\mathbb{P}\left(\boldsymbol{X}_n \in B\right) \to \mathbb{P}(\boldsymbol{X} \in B)$ for all Borel sets $B$ with $\mathbb{P}(\boldsymbol{X} \in \delta B) = 0$, where $\delta B$ is the boundary of $B$.

The alternative definitions of weak convergence can often be useful for satisfying certain properties of given theorems. When discussing weak convergence, we have that $\boldsymbol{X}_n \rightsquigarrow \boldsymbol{X}$ in some metric space $\mathbb{D}$, where above we took $\mathbb{D} = \mathbb{R}^p$. But generally, there could be multiple spaces $\mathbb{D}$ for which we can consider weak convergence. Fortunately, the choice of the particular space is not important as by Lemma 7.8 in Kosorok (2008), if $\mathbb{D}_0 \subset \mathbb{D}$ have the same metric and $\boldsymbol{X}, \boldsymbol{X}_n$ takes values in $\mathbb{D}_0$, then $\boldsymbol{X}_n \rightsquigarrow \boldsymbol{X}$ if and only if $\boldsymbol{X}_n \rightsquigarrow \boldsymbol{X}$ in $\mathbb{D}$.

When we consider the weak convergence of empirical processes, the limiting term is generally assumed to be tight. For a random vector $\boldsymbol{X}$, $\boldsymbol{X}$ is tight if for all $\epsilon > 0$, there is a compact set $K$ where

$$\mathbb{P}(\boldsymbol{X} \in K) \geq 1 - \epsilon \tag{1.51}$$

Therefore a sequence of random vectors $\boldsymbol{X}_n = \{\boldsymbol{X}_\alpha : \alpha \in \mathbb{N}\}$ is called uniformly tight if for every $\epsilon > 0$, there is a compact set $K$ such that for all $\alpha$,

$$\mathbb{P}(\boldsymbol{X}_\alpha \in K) \geq 1 - \epsilon\,. \tag{1.52}$$

Uniform tightness is also called bounded in probability, and we have that every weakly converging sequence $\{\boldsymbol{X}_n\}_{n\geq 1}$ is uniformly tight.

Often stochastic $o$ and $O$ symbols are used as notation to denote terms that converge in probability to zero or that are uniformly tight. The notation $o_P(1)$ denotes a sequence of random vectors that converge to zero in probability or more generally, we have that for a sequence of random variables $\boldsymbol{R}_n$ that

$$\boldsymbol{X}_n = o_P(\boldsymbol{R}_n), \implies \boldsymbol{X}_n = \boldsymbol{Y}_n\boldsymbol{R}_n,\ \boldsymbol{Y}_n \xrightarrow{\mathbb{P}} \boldsymbol{0}\,. \tag{1.53}$$

The expression $O_P(1)$ denotes that a sequence is uniformly tight, or more generally

$$\boldsymbol{X}_n = O_P(\boldsymbol{R}_n), \implies \boldsymbol{X}_n = \boldsymbol{Y}_n\boldsymbol{R}_n, \boldsymbol{Y}_n = O_P(1)\,. \tag{1.54}$$

### 1.2.3 Empirical Processes

If we suppose an independent and identically distributed (i.i.d) random sample $X_1, \ldots, X_n$ coming from a probability distribution $F$ on $\mathbb{R}$, then the empirical distribution function is denoted by

$$\mathbb{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1} \{X_i \leq x\} \ , \tag{1.55}$$

where $x \in \mathbb{R}$. By the Law of Large Numbers, we have that

$$\mathbb{F}_n(x) \xrightarrow{a.s} F(x) \text{ for each } x, \tag{1.56}$$

where a.s denotes almost sure convergence. Glivenko (1933) and Cantelli (1933) extended the Law of Large Numbers to uniform convergence (Glivenko-Cantelli Theorem) by showing that

$$||\mathbb{F}_n - F||_\infty = \sup_{x \in \mathbb{R}} |\mathbb{F}_n(x) - F(x)| \xrightarrow{a.s} 0 \ ,$$

where $||\mathbb{F}_n - F||_\infty$ is known as the Kolmogorov-Smirnov Statistic. Furthermore, by the Central Limit Theorem we asymptotic normality of the empirical distribution function

$$\sqrt{n} \left(\mathbb{F}_n(x) - F(x)\right) \rightsquigarrow \mathcal{N}\left(0, F(x)(1 - F(x))\right) \tag{1.57}$$

This can be extended to multiple $x$'s by the Multivariate Central Limit Theorem where for every $x$ in a finite set $T_k = \{x_1, x_2, \ldots, x_k\} \in \mathbb{R}$,

$$\sqrt{n} \begin{pmatrix} \mathbb{F}_n(x_1) - F(x_1) \\ \mathbb{F}_n(x_2) - F(x_2) \\ \vdots \\ \mathbb{F}_n(x_k) - F(x_k) \end{pmatrix} \rightsquigarrow \begin{pmatrix} G_F(x_1) \\ G_F(x_2) \\ \vdots \\ G_F(x_k) \end{pmatrix} \ , \tag{1.58}$$

where $\boldsymbol{G}_F = (G_F(x_1), \ldots G_F(x_k))^T$ is a multivariate normal distribution with mean zero and covariance

$$\text{Cov}(G(x_i), G(x_j)) = F(\min(x_i, x_j)) - F(x_i)F(x_j), \quad i, j = 1, \ldots, k$$

This limiting process $\boldsymbol{G}_F$ is also known as a $F$-Brownian Bridge.

Now, if we consider $x \mapsto \mathbb{F}_n(x)$ as a random function instead of a real-valued estimator, the results above can be extended more generally to a class of functions and similar notions can be considered for more general classes of functions. Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ be a random sample from a probability distribution $F$ with underlying measure $P$ on a measurable space $(\mathcal{X}, \mathcal{G})$. Denote $\mathbb{P}_n$ to be the empirical measure,

$$\mathbb{P}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i} \ ,$$

where $\delta_x$ is the measure that assigns mass 1 at $x$ and 0 elsewhere. Therefore, given a collection $\mathcal{H}$ of measurable functions $h : \mathcal{X} \to \mathbb{R}$, we denote the expectation of $h$ under the empirical measure by the operator

$$\mathbb{P}_n h = \frac{1}{n} \sum_{i=1}^{n} h(X_i) \ . \tag{1.59}$$

The empirical process $\{\mathbb{P}_n h, h \in \mathcal{H}\}$ can be viewed here as a stochastic process indexed by $h \in \mathcal{H}$. Furthermore, we can denote the $Ph$ as the expectation of $h$ under $P$ by

$$Ph = \int h(x) \, \mathrm{d}P(x) = \int h(x) \, \mathrm{d}F(x) \ . \tag{1.60}$$

Then, we can consider the notion of uniform consistency of an empirical measure over some given class of functions $\mathcal{H}$ with a corresponding measure $P$. For a measure $P$, a class of functions $\mathcal{H}$ is called *P-Glivenko-Cantelli* if

$$||\mathbb{P}_n - P||_{\mathcal{H}} = \sup_{h \in \mathcal{H}} |\mathbb{P}_n h - Ph| \xrightarrow{a.s} 0$$

Note that this corresponds to the Glivenko Cantelli theorem if we take $\mathcal{H}$ to be the class $\mathcal{H}_L = \{\mathbb{1}(x \leq t) \,|t \in \mathbb{R}\}$, of left-hand indicator functions. Donsker (1952) strengthened the multivariate central limit theorem to a uniform central limit theorem as follows. First, define the following empirical process evaluated at $h \in \mathcal{H}$,

$$\mathbb{G}_n h = \sqrt{n}\left(\mathbb{P}_n h - Ph\right). \tag{1.61}$$

Next, let us denote $\ell^\infty(T)$ to be the set of all uniformly bounded, real functions on an arbitrary set $T$, where for $z : T \to \mathbb{R}$,

$$||z||_T := \sup_{t \in T} |z(t)| < \infty. \tag{1.62}$$

Therefore, for a stochastic process $\{X(t) : t \in T\}$ with bounded sample paths, the stochastic process yields a map $X : \Omega \to \ell^\infty(T)$. Next, denote $\mathbb{G}_P$ as a mean-zero Gaussian Process indexed by $\mathcal{H}$ with covariance function

$$Ph_1h_2 - Ph_1Ph_2 = \mathbb{E}\left[h_1h_2\right] - \mathbb{E}\left[h_1\right]\mathbb{E}\left[h_2\right] \ , \tag{1.63}$$

$h_1, h_2 \in \mathcal{H}$. Furthermore, as emphasised in Van Der Vaart and Wellner (1996), it is assumed when dealing with a uniform central limit theorem that $\sup_{h \in \mathcal{H}} |h(\boldsymbol{x}) - Ph| < \infty$ for all $\boldsymbol{x} \in \mathcal{X}$ . Thus, re-expressing the multivariate central limit theorem, given a finite set of measurable functions $h_1, \ldots h_k \in \mathcal{H}_L$ with $Ph_i^2 < \infty$ for $i = 1, 2, \ldots, k$,

$$\begin{pmatrix} \mathbb{G}_n h_1 \\ \mathbb{G}_n h_2 \\ \vdots \\ \mathbb{G}_n h_k \end{pmatrix} \rightsquigarrow \begin{pmatrix} \mathbb{G}_P h_1 \\ \mathbb{G}_P h_2 \\ \vdots \\ \mathbb{G}_P h_k \end{pmatrix} \tag{1.64}$$

in $\ell^\infty(\mathcal{H}_L)$. Importantly, $\mathbb{G}_P$ is indexed by $h \in \mathcal{H}_L$ and is a tight Borel measurable map from some probability space to $\ell^\infty(\mathcal{H}_L)$. Therefore, we can extend the result above can hold for a more general class of functions $\mathcal{H}$, which is dependent on the underlying measure $P$ as follows.

**Definition 1.3 (*P*-Donsker Class)** *A given measure $P$, a class $\mathcal{H}$ of measurable functions $h : \mathcal{X} \to \mathbb{R}$ is a P-Donsker class if*

$$\left\{\sqrt{n}\left(\mathbb{P}_n - P\right)h\right\}_{h \in \mathcal{H}} \rightsquigarrow \mathbb{G}_P \text{ in } \ell^\infty(\mathcal{H}) \ , \tag{1.65}$$

*where the tight limiting process $\mathbb{G}_P$ is a Gaussian process indexed by $h \in \mathcal{H}$ with mean zero and covariance*

$$Cov(h_1, h_2) = Ph_1h_2 - Ph_1Ph_2, \quad h_1, h_2 \in \mathcal{H}. \tag{1.66}$$

In literature, a $P$-Donsker class is often just called as *Donsker class* when it's clear what the underlying measure is. If for a particular class, $\mathcal{H}$ the Donsker result holds for any underlying measure $P$, then the class is called a *universal Donsker class*. An example of a universal Donsker class is the class of all indicator functions of lower rectangles $\{\mathbb{1}(-\infty, \boldsymbol{r} : r) \in \mathbb{R}^d\}$.

The results above are for classes of real-valued functions, however, the thesis will predominately focus on showing vector-valued functions are Donsker classes. Van Der Vaart (1998) defines a class $\mathcal{H}$ of vector-valued functions $\boldsymbol{h} : \boldsymbol{x} \mapsto \mathbb{R}^k$ to be Glivenko-Cantelli or Donsker if each of the classes of coordinates $h_i : \boldsymbol{x} \mapsto \mathbb{R}$ with $\boldsymbol{h} = (h_i, \dots, h_k)$ ranging over $\mathcal{H}$ $(i = 1, 2, \dots k)$ is Glivenko-Cantelli or Donsker respectively. We will primarily consider two main Donsker classes, the first is $\mathcal{H}_L = \{\mathbb{1}(\boldsymbol{y} \leq \boldsymbol{r}) | \boldsymbol{r} \in \mathbb{R}^K\}$, the class of indicator functions of lower left rectangles which is a Donsker class as it is a Vapnik-Cervonenkis (VC) class.

**Definition 1.4 (Vapnik-Cervonenkis class, Van Der Vaart and Wellner 1996)** *Let $C$ be a collection of subsets of a set $\mathcal{X}$, and $\{x_1, x_2, \dots, x_n\}$ is an arbitrary set of $n$ points. Suppose $C$ picks out a certain subset which is of the form $c \cap \{x_1, x_2, \dots, x_n\}$ for a $c \in C$. The collection $C$ is said to shatter $\{x_1, x_2, \dots, x_n\}$ if each of its $2^n$ subsets can be picked out in this manner. The Vapnik-Cervonenkis (VC)-index $V(C)$ is the smallest $n$ for which no set of size $n$ is shattered by $C$. A collection of measurable sets $C$ is called a VC-class if its index is finite.*

*For functions, a subgraph of $f : \mathcal{X} \mapsto \mathbb{R}$ is the subset given by*

$$\{(x, t) : t < f(x)\}$$

*A collection $\mathcal{F}$ of measurable functions on a sample space is called a $VC-$subgraph class or a $VC$-class if the collection of all subgraphs of the function in $\mathcal{F}$ forms a VC-class of sets.*

The second is the class of vector-valued functions which are Lipschitz continuous in parameter, which is Donsker as seen in Theorem 5.21, and Example 19.7 of Van Der Vaart (1998), and in Example 3.3.7 in Van Der Vaart and Wellner (1996).

**Definition 1.5 (Lipschitz Continuous in Parameter)** *Suppose normed spaces $\mathcal{X}, \mathbb{D}$ and let $\mathcal{F} = \{f_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ be a collection of measurable functions $f_{\boldsymbol{\theta}} : \mathcal{X} \to \mathbb{D}$, indexed by a bounded subset $\Theta \subset \mathbb{R}^d$. The class of functions $\mathcal{F}$ is Lipschitz continuous in $\boldsymbol{\theta}$ if there exists some measurable function $m$ such that for all $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta$,*

$$||\boldsymbol{f}_{\boldsymbol{\theta}_1}(\boldsymbol{x}), \boldsymbol{f}_{\boldsymbol{\theta}_2}(\boldsymbol{x}))||_{\mathbb{D}} \leq m(\boldsymbol{x})||\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2||_{\mathcal{X}} \tag{1.67}$$

Instead of adhering to the definition of a Donsker class for each class of functions, it's often easier to use the fact that the Donsker property is preserved under certain transformations. For example, the product of uniformly bounded Donsker classes is also a Donsker class, and the product between a Donsker class and uniformly bounded measurable function is also a Donsker class.

### 1.2.4 Semiparametric Theory

A statistical model is a collection of probability measures $\{P \in \mathcal{P}\}$ on a sample space $\mathcal{Z}$. A semiparametric model is a collection of probability measures of the form $\mathcal{P} = \{P_{\boldsymbol{\beta},F} : \boldsymbol{\beta} \in \mathbb{R}^q, F \in \mathcal{F}\}$, where $\boldsymbol{\beta}$ is a finite-dimensional parameter, $F$ is an infinite-dimensional parameter and $\boldsymbol{\beta}, F$ are variationally independent. We also assume the true model $P^*$ generating the data is given by $P_{\boldsymbol{\beta}^*,F^*}$ for some $\boldsymbol{\beta}^* \in \mathbb{R}^q, F^* \in \mathcal{F}$. In this thesis, the infinite-dimensional parameter space $\mathcal{F}$ is the set of all probability distributions on $\mathcal{Z}$ and this allows for fewer restrictions on the probabilistic constraints on data compared to finite-dimensional parametric models.

In semiparametric literature, parametric submodels are used to develop the theory behind semiparametric models. Parametric submodels denoted by $\mathcal{P}_{\beta,F_t} = \{P_{\beta,F_t} | \beta \in \mathbb{R}^q, t \in T\}$ are a class of models indexed by finite dimensional parameters $\beta, F_t$ such that $\mathcal{P}_{\beta,F_t} \in \mathcal{P}$ and $P^* \in \mathcal{P}_{\beta,F_t}$. The first condition states that every distribution in $\mathcal{P}_{\beta,F_t}$ is a member of the semiparametric model $\mathcal{P}$. The

second condition says that the parametric submodel must contain the true distribution $P^* = P_{\beta^*,F_t^*}$. Here the parameter space $T$ for $t$ is taken to be some appropriate subspace of $\mathbb{R}^r$ for $1 \leq r < \infty$.

Key concepts in semiparametric theory utilise Hilbert spaces whose elements are $q \times 1$ random vectors with mean zero and finite variance, with the covariance inner product

$$\langle h_1, h_2 \rangle = \mathbb{E}\left[h_1^T h_2\right] \tag{1.68}$$

for $h_1, h_2$ being elements of the linear vector space $\mathscr{H}$. A key result for Hilbert spaces which is used in semiparametric theory is the projection theorem given below.

**Theorem 1.2 (Projection Theorem)** *Let $(h, \langle \cdot, \cdot \rangle)$ be a Hilbert space and let $\mathcal{U} \subset \mathscr{H}$ be a closed linear subspace. For any $h \in \mathscr{H}$ there exists a unique $u_0 \in U$ that is closest to $h$,*

$$u_0 = \arg\min_{u \in U} ||h - u|| \, . \tag{1.69}$$

*Furthermore, $u_0$ is characterized by the fact that $h - u_0$ is orthogonal to $U$*

$$\langle h - u_0, u \rangle = 0, \; u \in \mathcal{U} \, . \tag{1.70}$$

Note that a linear subspace is a space $\mathcal{U} \subset \mathscr{H}$ where $u_1, u_2 \in \mathcal{U}$ implies that $au_1 + bu_2 \in \mathcal{U}$ for all scalar constants $a, b$. Importantly, $u_0$ is referred to as the projection of $h$ onto the space $U$ denoted by $\Pi(h \,|\, \mathcal{U})$. We can find an expression for this unique projection of $h$ for our Hilbert space $\mathscr{H}$ by following Chapter 2, Example 2 in Tsiatis (2006).

Let $S_{\boldsymbol{\theta}}(\boldsymbol{X}, \boldsymbol{Y})$ be a $r$ dimensional score vector with mean 0 and $\langle S_{\boldsymbol{\theta}}, S_{\boldsymbol{\theta}} \rangle < \infty$. Consider the linear subspace $\mathcal{U}$ spanned by $S_{\boldsymbol{\theta}}$ given by

$$\mathcal{U} = \{B^{q \times r} S_{\boldsymbol{\theta}}(\boldsymbol{X}, \boldsymbol{Y}), \text{where } B \text{ is any arbitrary } q \times r \text{ matrix of real numbers}\},$$

where $\mathcal{U}$ is a finite-dimensional linear subspace contained in the infinite-dimensional Hilbert space $\mathscr{H}$. This linear subspace is often referred to as a tangent space. By following the example, the unique projection of an arbitrary $h(\boldsymbol{X}, \boldsymbol{Y}) \in \mathscr{H}$ onto $\mathcal{U}$ by the projection theorem is

$$\Pi(h \,|\, \mathcal{U}) = \mathbb{E}[h(\boldsymbol{X}, \boldsymbol{Y}) S_{\boldsymbol{\theta}}(\boldsymbol{X}, \boldsymbol{Y})^T] \mathbb{E}[S_{\boldsymbol{\theta}}(\boldsymbol{X}, \boldsymbol{Y}) S_{\boldsymbol{\theta}}(\boldsymbol{X}, \boldsymbol{Y})^T]^{-1} S_{\boldsymbol{\theta}}(\boldsymbol{X}, \boldsymbol{Y}) \tag{1.71}$$

In the case where $\boldsymbol{\theta}$ can be partitioned as $(\boldsymbol{\beta}^T, F)^T$, the linear subspace spanned by the nuisance score vector is known as a *nuisance tangent space.* Now we can define the notion of a parametric submodel nuisance tangent space and a semiparametric model nuisance tangent space. Note that associated with each model $P_{\boldsymbol{\beta},F_t}$ in a parametric submodel is a density $p_{\boldsymbol{\beta},F_t}(\boldsymbol{x}, \boldsymbol{y})$, a log-likelihood $l(\boldsymbol{\beta}, F_t)(\boldsymbol{x}, \boldsymbol{y}) = \log p_{\boldsymbol{\beta},F_t}(\boldsymbol{x}, \boldsymbol{y})$, and a score function for $t$

$$S_t(\boldsymbol{\beta}, F_t)(\boldsymbol{x}, \boldsymbol{y}) = \frac{\partial l(\boldsymbol{\beta}, F_t)(\boldsymbol{x}, \boldsymbol{y})}{\partial t} \in \mathbb{R}^{r \times 1} \, . \tag{1.72}$$

**Definition 1.6 (Parametric & Semiparametric Nuisance tangent space, Tsiatis 2006)** *The nuisance tangent space for a semiparametric model denoted by $\Lambda$ is defined as the mean-square closure of parametric submodel nuisance tangent spaces, where a parametric submodel is the set of elements*

$$\Lambda_t = \{B^{q \times r} \boldsymbol{S}_t(\boldsymbol{\beta^*}, F^*)(\boldsymbol{x}, \boldsymbol{y}), \textit{where } B \textit{ is any arbitrary } q \times r \textit{ matrix of real numbers}\} \, , \tag{1.73}$$

*$S_t(\boldsymbol{\beta}^*, F^*)(\boldsymbol{x}, \boldsymbol{y})$ is the score vector for the nuisance parameter $t$ for some parametric submodel, evaluated at $\boldsymbol{\beta} = \boldsymbol{\beta}^*$, $F_{t^*} = F^*$. Specifically, the mean-square closure of the spaces above is defined as the space $\Lambda \subset \mathscr{H}$, where*

$$\Lambda = \{h^{q\times 1}(\boldsymbol{X}, \boldsymbol{Y}) \in \mathscr{H}\} \tag{1.74}$$

*such that $||h(\boldsymbol{X}, \boldsymbol{Y})|| < \infty$ and there is some subsequence $B_j S_{tj}$ such that*

$$||h(\boldsymbol{X}, \boldsymbol{Y}) - B_j \boldsymbol{S}_{tj}(\boldsymbol{X}, \boldsymbol{Y})||^2 \xrightarrow{j\to\infty} 0$$

*for a sequence of parametric submodels indexed by $j$.*

The nuisance tangent space is important as we use it to define is efficient score function for $\boldsymbol{\beta}$.

**Definition 1.7 (Efficient score function for $\boldsymbol{\beta}$, Tsiatis 2006)** *The semiparametric efficient score for $\boldsymbol{\beta}$ is defined as*

$$\tilde{S}(\boldsymbol{\beta}^*, F^*) = S_{\boldsymbol{\beta}}(\boldsymbol{\beta}^*, F^*) - \Pi(S_{\boldsymbol{\beta}}(\boldsymbol{\beta}^*, F^*)|\Lambda) \tag{1.75}$$

This is the residual of the score vector with respect to $\boldsymbol{\beta}$ after projecting it onto the nuisance tangent space. Note that $\Lambda$ is a closed linear subspace and therefore the projection $\Pi$ does exist and is unique. Furthermore, by Theorem 4.1 in Tsiatis (2006), we have that the inverse of the covariance matrix of the semiparametric efficient score

$$\left(\mathbb{E}\left[\tilde{S}(\boldsymbol{\beta}^*, F^*)\tilde{S}(\boldsymbol{\beta}^*, F^*)^T\right]\right)^{-1} \tag{1.76}$$

attains the semiparametric efficiency bound for estimating $\boldsymbol{\beta}$.

If the maximum likelihood estimator satisfies this efficient score equation, then we have that $\hat{\boldsymbol{\beta}}$ is asymptotically efficient, provided some conditions for the maximum likelihood estimator $\hat{F}$. However, the efficient score is a projection so nothing guarantees that this projection is the derivative of the log-likelihood under some parametric submodel and that the maximum likelihood estimator satisfies the efficient score equation. As suggested by Van Der Vaart (1998), this can be remedied using an approximately least-favourable submodel. First, let us introduce the notion of a least favourable submodel.

Suppose we are aiming to estimate some parameter $\psi(P^*)$ given the model $\mathcal{P}$, which is easier if we do so along a submodel $\mathcal{P}_0 = \{P_\theta : \theta \in \Theta\}$. Then for every smooth parametric submodel, we can calculate the Fisher information for estimating $\psi(P_\theta)$ and note that the Fisher information for estimating $\psi(P^*)$ is not larger than the infimum of the information over all submodels. A submodel that attains this infimum is called a *least favourable submodel*, where we define the information for the whole model as this infimum. Van Der Vaart (1998) notes that it suffices in most cases to consider one-dimensional submodels which pass through the true distribution $P^*$ and is differentiable at $P^*$.

Thus, going back to the efficient score, if there exists a least favourable path $t \mapsto F_t(\hat{\boldsymbol{\beta}}, \hat{F})$ such that $F_{\hat{\boldsymbol{\beta}}}(\hat{\boldsymbol{\beta}}, \hat{F}) = \hat{F}$ and for every $\boldsymbol{x}$,

$$\tilde{\ell}_{\hat{\boldsymbol{\beta}},\hat{F}}(\boldsymbol{X}, \boldsymbol{Y}) = \frac{\partial}{\partial t}\bigg|_{t=\hat{\boldsymbol{\beta}}} \log l_{t,F_t(\hat{\boldsymbol{\beta}},\hat{F})}(\boldsymbol{X}, \boldsymbol{Y}) \,, \tag{1.77}$$

then the maximum likelihood estimator satisfies the efficient score function. However, if a least favourable path doesn't exist, it's not immediately clear whether the MLE satisfies the efficient score function above. To overcome this uncertainty surrounding the existence of an exact least favourable

path, we replace the efficient score equation with an approximation as suggested in Van Der Vaart (1998), which is close to the efficient score function at least at the true parameter values. Thus, we define *approximately-least favourable submodels* as maps $t \mapsto F_t(\boldsymbol{\beta}, F)$ from neighbourhoods of $0 \in \mathbb{R}^q$ to $\mathcal{F}$ with $F_{\boldsymbol{\beta}}(\boldsymbol{\beta}, F) = F$ for every $(\boldsymbol{\beta}, F)$ such that

$$\ell_{\boldsymbol{\beta},F}(t)(\boldsymbol{X}, \boldsymbol{Y}) = \log l(t, F_t(\boldsymbol{\beta}, F))(\boldsymbol{X}, \boldsymbol{Y}), \tag{1.78}$$

$$\dot{\ell}_{\boldsymbol{\beta},F}(t)(\boldsymbol{X}, \boldsymbol{Y}) = \frac{\partial}{\partial t} \log l(t, F_t(\boldsymbol{\beta}, F))(\boldsymbol{X}, \boldsymbol{Y}) \tag{1.79}$$

exists and $\dot{\ell}_{\boldsymbol{\beta},F}$ is equal to the efficient score function at $(\boldsymbol{\beta}, F) = (\boldsymbol{\beta^*}, F^*)$. Thus the path $t \mapsto F_t(\boldsymbol{\beta}, F)$ must pass through $(\boldsymbol{\beta}, F)$ at $t = \boldsymbol{\beta}$ (i.e. $F_{\boldsymbol{\beta}}(\boldsymbol{\beta}, F) = F$) and at the true parameter $(\boldsymbol{\beta^*}, F^*)$ the submodel is truly least favourable for estimating $\boldsymbol{\beta}$ in that

$$\dot{\ell}_{\boldsymbol{\beta^*},F^*}(\boldsymbol{\beta^*})(\boldsymbol{X}, \boldsymbol{Y}) = \tilde{S}_{\boldsymbol{\beta^*},F^*}(\boldsymbol{X}, \boldsymbol{Y}) \, . \tag{1.80}$$

Furthermore, if $(\hat{\boldsymbol{\beta}}, \hat{F})$ maximizes the likelihood, then $(\hat{\boldsymbol{\beta}}, \hat{F})$ satisfies $\mathbb{P}_n \dot{\ell}_{\hat{\boldsymbol{\beta}},\hat{F}} = 0$. As a result, we can replace $\tilde{\ell}_{\boldsymbol{\beta},F}(\boldsymbol{X}, \boldsymbol{Y})$ with $\dot{\ell}_{\boldsymbol{\beta},F}(t)(\boldsymbol{X}, \boldsymbol{Y})$, which is done in Theorem 5.3 to bring in the approximately least favourable submodels.

### 1.2.5 Functional Analysis

For the derivations in Chapter 5 we make mention of linear operators, topological spaces and compact operators which we will define below.

**Definition 1.8 (Linear Mapping)** *A mapping $\mathcal{A}$ from a vector space $\mathcal{E}$ into a vector space $\mathcal{F}$ is said to be a linear mapping if*

$$\mathcal{A}(\alpha \boldsymbol{x}_1 + \beta \boldsymbol{x}_2) = \alpha \mathcal{A}(\boldsymbol{x}_1) + \beta \mathcal{A}(\boldsymbol{x}_2) \tag{1.81}$$

*for all $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{E}$.*

Linear mappings of $\mathcal{X}$ into $\mathcal{X}$ are called linear operators on $\mathcal{X}$. However in literature, linear transformations between $\mathcal{X} \to \mathcal{Y}$ are also called linear operators.

**Definition 1.9 (Bounded Operator)** *A linear mapping $\mathcal{A} : \mathcal{E} \to \mathcal{F}$ where $\mathcal{E}, \mathcal{F}$ are normed vector spaces is bounded if and only if there exists some $M > 0$ such that for all $\boldsymbol{x} \in \mathcal{E}$,*

$$||\mathcal{A}\boldsymbol{x}||_{\mathcal{F}} \leq M ||\boldsymbol{x}||_{\mathcal{E}} \tag{1.82}$$

Furthermore, an operator between two normed spaces is a bounded linear operator if and only if it is a continuous linear operator. Now, let us review the notion of a weak topology.

**Definition 1.10 (Topological Space)** *A topology of a nonempty set $X$ is a collection $\tau$ of subsets of $X$. Each element of $\tau$ is an open set where*

- *$X \in \tau$, $\emptyset \in \tau$,*
- *if $U_\alpha \in \tau$ for each $\alpha \in I_\alpha$ then $\cup_{\alpha \in I_\alpha} U_i \in \tau$,*
- *if $U_i \in \tau$ for $i = 1, 2, \ldots, n$ then $\cap_{i=1}^n U_i \in \tau$*

*The pair $(X, \tau)$ is a topological space, but often $X$ is referred to as a topological space.*

**Definition 1.11 (Dual Space)** *The dual space of a topological space $X$ is the vector space $X^*$ whose elements are linear mappings on $X$.*

**Definition 1.12 (Weak Topology)** *The weak topology on a topological space $X$ with a dual $X^*$ which separates points ($\forall x, y \in X, x \neq y$, there exists a bounded linear function $f \in X^*$such that $f(x) \neq f(y)$) is defined as the coarsest topology (one with the fewest open sets) in which all maps $f \in X^*$ are continuous on $X$.*

Now, let us review the notion of compactness.

**Definition 1.13 (Open Cover & Compact Set, Rudin 1973)** *Let $(X, d)$ be a metric space. An open cover of a set $E$ in $X$, is a collection $\{U_\alpha\}_{\alpha \in I_\alpha}$ of open subsets of $X$ such that $E \subseteq \cup_{\alpha \in I_\alpha} U_\alpha$*

*A subset $K$ of a metric space $X$ is compact if every open cover of $K$ contains a finite subcover. Or more explicitly, it $K \subseteq \cup_{\alpha \in I_\alpha} U_\alpha$ for a collection of open subsets $\{U_\alpha\}_{\alpha \in I_\alpha}$, then there exists finitely many $\alpha_i \in I_\alpha$ such that $K \subseteq \cup_{i=1}^{n} U_{\alpha_k}$ for some sub-collection.*

Therefore, a topological space $X$ is called compact if every open cover of $X$ has a finite subcover. As a result, we can now define the notion of a compact operator.

**Definition 1.14 (Compact Operator, Zimmer 1990)** *If $\mathcal{E}$ and $\mathcal{F}$ are Banach spaces, then a bounded linear operator $\mathcal{A} : \mathcal{E} \to \mathcal{F}$ is called a compact operator if $\overline{\mathcal{A}(B)}$ is compact in $\mathcal{F}$ for all bounded sets $B \subset \mathcal{E}$. Here $\overline{\mathcal{A}(B)}$ denotes the closure of the set $\mathcal{A}(B)$.*

Note that by definition compact operators are bounded and hence continuous. We also have another way of defining a compact operator as given by Theorem 3.1.5 in Zimmer (1990).

**Definition 1.15 (Compact Operator, Zimmer 1990)** *If $\mathcal{X}$ is a compact space with finite measure $\mu$, if $K \in C(\mathcal{X} \times \mathcal{X})$, then the integral operator*

$$(Tf)(x) = \int_{\mathcal{X}} K(x, y) f(y) d\mu(y) \tag{1.83}$$

*defines a compact operator $T : L^2(\mathcal{X}) \to L^2(\mathcal{X})$.*

This definition will be the one which we mainly adhere to in the derivations of the technical details in Chapter 5.

# Chapter 2

# Model and Parameter Estimation

## 2.1 Introduction

This chapter introduces a Vector Semiparametric Generalized Linear model (VSPGLM), which is a generalization of a semiparametric GLM (SPGLM) published by Huang (2014), to handle problems with vector responses. The VSPGLM is a member of a family of SPGLMs introduced by Rathouz and Gao (2009) which utilises an exponential tilt representation of the response's distribution conditional on the covariates. Taking this approach, the reference distribution of $F$ in a vector GLM can be left unspecified and treated as an infinite-dimensional parameter that is jointly estimated along with the usual mean-model parameters $\boldsymbol{\beta}$. As a result, the parametric specification of the joint distribution is not required, and there are no working covariance functions or dispersion parameters in the model. This overcomes problems of model misspecification in parametric models which lead to biased inference on the mean-model parameters, and handles problems with vector correlations and mixed data types where the reference distribution is not a standard multivariate distribution.

The chapter will introduce the model and provide derivations of the score functions, drawing comparisons to the SPGLM. The chapter then extends the results presented in Huang (2014), exploring parameter estimation using maximum empirical likelihood estimation (MELE), the asymptotic properties of the model and how to conduct inference on the $\boldsymbol{\beta}$ parameters.

## 2.2 A Semiparametric Extension of VGLMs (VSPGLM)

Let $(\boldsymbol{X}, \boldsymbol{Y}) \in \mathcal{X} \times \mathcal{Y} \subset \mathbb{R}^q \times \mathbb{R}^K$, where the covariates $\boldsymbol{X}$ are sampled according to some design measure $G_{\boldsymbol{X}}$ on $\mathcal{X}$, being either random samples from some population or fixed by design, not dependent on $\boldsymbol{Y}$. Each vector-valued response $\boldsymbol{Y}$ has $K$ components and is sampled from some multivariate distribution $F(\boldsymbol{y}|\boldsymbol{x})$, which is conditional on $\boldsymbol{X} = \boldsymbol{x}$. The data then consists of independent copies $(\boldsymbol{X}_1, \boldsymbol{Y}_1), (\boldsymbol{X}_2, \boldsymbol{Y}_2) \ldots (\boldsymbol{X}_n, \boldsymbol{Y}_n)$ of $(\boldsymbol{X}, \boldsymbol{Y})$.

Extending McCullagh and Nelder (1989), a Vector GLM is specified by two key components which we will provide for the proposed model. The first is a conditional mean model for each component of the form

$$\mathbb{E}\left[Y_{(k)}|\boldsymbol{X}\right] = \mu_{(k)}\left(\boldsymbol{X}_{(k)}^T \boldsymbol{\beta}_{(k')}\right) \qquad k' \in \{1, \ldots, K'\},\ k = 1, 2, \ldots, K\ . \tag{2.1}$$

Here $\mu_{(k)}$ is the user-specified inverse link function or the mean function, $\boldsymbol{\beta}_{(k')}$ is a $q_k$ vector of unknown coefficients associated with the $k$-th component with the $k'$-th mean model. Throughout the model, we

will be using the covariate and vector notation defined in Appendix A.1. To summarise, it's important to note the distinction between the number of components denoted by $K$ and the number of mean models denoted by $K' \leq K$. In the proposed framework components of the response can be constrained to having the same link function and coefficients, or in other words share the same mean model.

The notation is easier to understand through an example, so consider jointly modelling two longitudinal studies on Raven and Arithmetic scores of Kenyan school children across 4 visits as is explored in Section 4.7. This generates a response vector $\boldsymbol{Y}$ with 8 components corresponding to the measurements of Raven score and Arithmetic score across the 4 time periods. Given we are considering a longitudinal study, we want to constrain all of the components corresponding to Raven scores to share the same coefficients $\boldsymbol{\beta}_{(1)}$, which are different to the coefficients $\boldsymbol{\beta}_{(2)}$ shared between all of the components with Arithmetic scores as seen in (2.2).

$$\begin{aligned}
\mathbb{E}\left[\text{Raven}_{i(k)}|\boldsymbol{X}_i\right] =& \beta_{(1)0} + \beta_{(1)1}\,\text{age}_i + \beta_{(1)2}\,\text{ses}_i + \beta_{(1)3}\,\text{braven}_i + \beta_{(1)4}\,\mathbb{1}(\text{male}_i) \\
&+ \left(\beta_{(1)5} + \beta_{(1)6}\,\text{milk}_i + \beta_{(1)7}\,\text{meat}_i + \beta_{(1)8}\,\text{energy}_i\right) \times \text{rel_time}_{ik}, \\
\mathbb{E}\left[\text{Arithmetic}_{i(k+4)}|\boldsymbol{X}_i\right] =& \beta_{(2)0} + \beta_{(2)1}\,\text{age}_i + \beta_{(2)2}\,\text{ses}_i + \beta_{(2)3}\,\text{barithmetic}_i + \beta_{(2)4}\,\mathbb{1}(\text{male}_i) \\
&+ \left(\beta_{(2)5} + \beta_{(2)6}\,\text{milk}_i + \beta_{(2)7}\,\text{meat}_i + \beta_{(2)8}\,\text{energy}_i\right) \times \text{rel_time}_{ik}\,, \qquad (2.2)
\end{aligned}$$

for $i = 1, \ldots, n$ and $k = 1, 2, 3, 4$. Importantly above, we note that there are $K' = 2$ mean models and $K = 8$ components, where the set of mean model parameters is denoted by $\boldsymbol{\beta} = \{\boldsymbol{\beta}_{(k')} \in \mathbb{R}^{q_k} | k' = 1, 2, \ldots, K'\}$ and the total number of mean model parameters is $\sum_{k=1}^{K} q_k = q$. The aim of the notation is to allow for flexibility to constrain coefficients between components such seen in the VGLM framework proposed by Yee (2015) which uses explicit user-specified constraint matrices. Each component $k$ will have an associated vector of covariates $\boldsymbol{X}_{(k)}$ and note that components with the same mean-model can have different covariates as seen in (2.2) with the rel_time covariate. More details on the notation can be found in Appendix A.1, as well as an example of the in Appendix A.3.

Secondly, the joint distribution $F(\boldsymbol{y}|\boldsymbol{x})$ is assumed to come from some multivariate exponential family. Utilising the exponential tilt formulation of the multivariate exponential family $F_{\boldsymbol{\theta}}(\boldsymbol{y})$ given in Rathouz and Gao (2009), the density with respect to an appropriate common measure $\lambda$ can be expressed as

$$dF_{\boldsymbol{\theta}}(\boldsymbol{y}) = \exp\left(b + \boldsymbol{\theta}^T\boldsymbol{y}\right) dF(\boldsymbol{y}) \qquad (2.3)$$

for some baseline distribution $F(\boldsymbol{y})$. Here $b = b(\boldsymbol{X}, \boldsymbol{\beta}, F)$ is the cumulant generating function (c.g.f) of the distribution which can be seen as a normalizing function where

$$b = b(\boldsymbol{X}, \boldsymbol{\beta}, F) = -\log\left\{\int_{\mathcal{Y}} \exp\left(\boldsymbol{\theta}^T\boldsymbol{y}\right) \mathrm{d}F(\boldsymbol{y})\right\}\,. \qquad (2.4)$$

This indeed satisfies the normalization constraint as

$$\int_{\mathcal{Y}} \mathrm{d}F_{\boldsymbol{\theta}}(\boldsymbol{y}) = \int_{\mathcal{Y}} \exp\left(b + \boldsymbol{\theta}^T\boldsymbol{y}\right) \mathrm{d}F(\boldsymbol{y}) = \int_{\mathcal{Y}} \frac{\exp\left(\boldsymbol{\theta}^T\boldsymbol{y}\right)}{\left(\int_{\mathcal{Y}} \exp(\boldsymbol{\theta}^T\boldsymbol{y}_1)\, \mathrm{d}F(\boldsymbol{y}_1)\right)} \mathrm{d}F(\boldsymbol{y}) = 1\,. \qquad (2.5)$$

For notation, as $\frac{\mathrm{d}F(\boldsymbol{y})}{\mathrm{d}\lambda(\boldsymbol{y})} = dF(\boldsymbol{y})$, for any measurable function $g(\boldsymbol{y})$ the following are equivalent by the definition of a density,

$$\int_{\mathcal{Y}} g(\boldsymbol{y})\, \mathrm{d}F(\boldsymbol{y}) = \int_{\mathcal{Y}} g(\boldsymbol{y}) dF(\boldsymbol{y})\, \mathrm{d}\lambda(\boldsymbol{y})\,. \qquad (2.6)$$

Similar to a parametric GLM, the reference distribution $F$ is required to have a Laplace transformation in some neighbourhood of the origin so that (2.4) exists and is well defined.

The canonical or tilt parameter $\boldsymbol{\theta} = \left(\boldsymbol{\theta}_{(1)}(\boldsymbol{X};\boldsymbol{\beta},F),\ldots,\boldsymbol{\theta}_{(K)}(\boldsymbol{X};\boldsymbol{\beta},F)\right)^T$ is implicitly defined as the solution to the mean constraint for each component

$$\mu_{(k)}\left(\boldsymbol{X}_{(k)}^T\boldsymbol{\beta}_{(k')}\right) = \int_{\mathcal{Y}} y_{(k)} \exp\left(b+\boldsymbol{\theta}^T\boldsymbol{y}\right) \mathrm{d}F(\boldsymbol{y}) \quad k' \in \{1,\ldots,K'\},\ k = 1,2,\ldots,K \tag{2.7}$$

or using our vector integral notation,

$$\boldsymbol{\mu}\left(\boldsymbol{X}^T\boldsymbol{\beta}\right) = \int_{\mathcal{Y}} \boldsymbol{y} \exp\left(b+\boldsymbol{\theta}^T\boldsymbol{y}\right) \mathrm{d}F(\boldsymbol{y})\ . \tag{2.8}$$

As $b$ and $\boldsymbol{\theta}$ are functions of $\boldsymbol{X}$, the notation of conditioning on $\boldsymbol{X}$ is usually dropped in favour of indexing by $\boldsymbol{\theta}$. Thus, each density $dF_{\boldsymbol{\theta}}(\boldsymbol{y})$ is an exponential tilt of some reference density $dF(\boldsymbol{y})$ where the amount of tilt given to each component of the response is dependent on the response mean $\mu_{(k)}$ for $k = 1,2,\ldots,K$. As any multivariate exponential family can be re-expressed in an exponential tilt representation, (2.1)-(2.7), it encompasses all classical multivariate GLMs with the appropriate choice of $F$, along with all univariate GLMs when $K = 1$.

The key advantage of the exponential tilt formulation is that it naturally leads itself to a semiparametric extension where the reference distribution $F$ can be left unspecified, removing the potential for model misspecification. Then, the reference distribution $F$ is treated as an infinite-dimensional parameter which is estimated along with the mean-model parameters $\boldsymbol{\beta}$. Thus, the semiparametric log-likelihood of the data $(\boldsymbol{X}_1,\boldsymbol{Y}_1),(\boldsymbol{X}_2,\boldsymbol{Y}_2),\ldots,(\boldsymbol{X}_n,\boldsymbol{Y}_n)$ becomes

$$\ell(\boldsymbol{\beta},F|\boldsymbol{X},\boldsymbol{Y}) = \sum_{i=1}^{n}\left\{\log\left\{dF(\boldsymbol{Y}_i)\right\} + b(\boldsymbol{X}_i;\boldsymbol{\beta},F) + \boldsymbol{\theta}(\boldsymbol{X}_i;\boldsymbol{\beta},F)^T\boldsymbol{Y}_i\right\}\ , \tag{2.9}$$

where $b$ and $\boldsymbol{\theta}$ are defined in (2.4) and (2.7).

Note that the log-likelihood (2.9) is invariant to any exponential tilting of $F$, meaning that the reference distribution $F$ is not identifiable. To make $F$ identifiable, we constrain $F$ to have some mean $\boldsymbol{\mu}$ in the interior of $\mathcal{Y}$, where the particular choice of $\boldsymbol{\mu}$ is irrelevant due to the invariance of the log-likelihood. Thus, the parameter space for $F$ denoted $\mathcal{F}_{\boldsymbol{\mu}}$ is the class of all distributions with mean $\boldsymbol{\mu}$ and a Laplace transformation in a neighbourhood of the origin. To emphasise the dependence on $\boldsymbol{\mu}$, when needed for clarity we equivalently express $F$ as $F_{\boldsymbol{\mu}}$.

The proposed VSPGLM has the following orthogonality property shown by Huang and Rathouz (2017) for a single-valued response, where the extension to the vector-valued response case is formally verified in Section 5.2.1. Orthogonality in this context refers to the score functions for the finite-dimensional parameter $\boldsymbol{\beta}$ being orthogonal to the nuisance tangent space for the infinite-dimensional parameter $F$. The implications of Lemma 2.1 is that any estimation and inference of $\boldsymbol{\beta}$ is asymptotically efficient and independent of any estimation of $F$.

**Lemma 2.1 (Orthogonality)** *The mean model parameters $\boldsymbol{\beta}$ and the reference distribution $F$ in any Vector Generalized Linear Model are orthogonal.*

The key observation here is that the nuisance tangent space for $F$ is a subspace of the nuisance tangent space for a more general class of semiparametric restricted moment models. In fact, as highlighted by Huang and Rathouz (2017), the nuisance tangent space for any model for which $F$ is characterized by a finite number of parameters (parametric model) is necessarily a subspace of the semiparametric nuisance tangent space. This results in the following corollary which immediately extends to the case of vector responses.

**Corollary 2.1 (Orthogonality of parametric models)** *If the reference distribution is parameterised by a finite vector of nuisance parameters $\boldsymbol{\phi}$, then the mean model parameter $\boldsymbol{\beta}$ is orthogonal to $\phi$.*

Furthermore, as the model utilises exponential tilt families, we can infer a connection to QL-methods through the following result by Hiejima (1997). This is, for any mean-variance relationship, there exists an exponential family whose roots of the score equation are asymptotically close to the roots of the corresponding QL-score equation. Therefore, similar to Huang (2014) the proposed VSPGLM can approximate asymptotically well any correctly specified QL-method while avoiding any model specification, as the model allows the data to dictate which exponential family is selected. Therefore, the model and its inferences are asymptotically valid as long as the variance of the true distribution is a function of the mean and not a function of the covariates.

Finally, in contrast to the MDRM proposed by Marchese and Diao (2017) which performs regression on the scale of the canonical parameters $\boldsymbol{\theta} = (\boldsymbol{x}^T\boldsymbol{\beta})$, VSPLM models the mean function $\mu_{(k)}(\boldsymbol{X}_{(k)}^T\boldsymbol{\beta}_{(k')})$ directly. As a result, the mean-model parameters $\boldsymbol{\beta}$ remain interpretable like in a standard GLM, where $\boldsymbol{\beta}$ are the mean contrasts or treatment effects present in the data, independent of the underlying reference distribution $F(\boldsymbol{y})$.

## 2.3 Derivations of the Score Function

In this section, we will derive the score equations for both $\boldsymbol{\beta}$ and $F$, which are the partial derivatives of the log-likelihood with respect to the appropriate parameters.

### 2.3.1 Score Function for $\beta$

The derivations for the score function of $\boldsymbol{\beta}$ for VSPGLM will extend the working in Rathouz and Gao (2009), but for a vector-valued response. For this derivation of the score function for $\boldsymbol{\beta}$, $F$ is held fixed. The score function for $\boldsymbol{\beta}$ is defined as

$$\boldsymbol{S}_{\boldsymbol{\beta},F}(\boldsymbol{X},\boldsymbol{Y}) = \sum_{i=1}^{n} \frac{\partial}{\partial\boldsymbol{\beta}}\ell(\boldsymbol{\beta},F|\boldsymbol{X}_i,\boldsymbol{Y}_i) = \sum_{i=1}^{n} S_{\boldsymbol{\beta},F}(\boldsymbol{X}_i,\boldsymbol{Y}_i)\ . \tag{2.10}$$

Let us consider a single generic data pair $(\boldsymbol{X},\boldsymbol{Y})$ where the log-likelihood is given by

$$\ell(\boldsymbol{\beta},F|\boldsymbol{X},\boldsymbol{Y}) = \log\{dF(\boldsymbol{Y})\} + b(\boldsymbol{X};\boldsymbol{\beta},F) + \boldsymbol{\theta}(\boldsymbol{X};\boldsymbol{\beta},F)^T\boldsymbol{Y}, \tag{2.11}$$

and the score function for $\boldsymbol{\beta}$ is given by

$$S_{\boldsymbol{\beta},F}(\boldsymbol{X},\boldsymbol{Y}) = \frac{\partial}{\partial\boldsymbol{\beta}}\ell(\boldsymbol{\beta},F|\boldsymbol{X},\boldsymbol{Y})\ , \tag{2.12}$$

where $S_{\boldsymbol{\beta},F}(\boldsymbol{X},\boldsymbol{Y})$ is a $q\times 1$ vector. By an application of a multivariate chain rule, we have

$$\frac{\partial l}{\partial\boldsymbol{\beta}} = \frac{\partial\boldsymbol{\mu}}{\partial\boldsymbol{\beta}}\frac{\partial\boldsymbol{\theta}}{\partial\boldsymbol{\mu}}\frac{\partial l}{\partial\boldsymbol{\theta}} \tag{2.13}$$

To find these partial derivatives, let us consider the following identities. Through an application of the Leibniz integral rule due to the assumptions in Appendix A.2, we can bring the partial derivative inside of the integral,

$$\begin{aligned} b &= -\log\left[\int_{\mathcal{Y}} \exp\left(\boldsymbol{\theta}^T\boldsymbol{y}\right)\mathrm{d}F(\boldsymbol{y})\right] \\ -\frac{\partial b}{\partial\boldsymbol{\theta}} &= \frac{\int_{\mathcal{Y}}\frac{\partial}{\partial\boldsymbol{\theta}}\exp\left(\boldsymbol{\theta}^T\boldsymbol{y}\right)\mathrm{d}F(\boldsymbol{y})}{\int_{\mathcal{Y}}\exp(\boldsymbol{\theta}^T\boldsymbol{y})\,\mathrm{d}F(\boldsymbol{y})} = \frac{\int_{\mathcal{Y}}\boldsymbol{y}\exp\left(\boldsymbol{\theta}^T\boldsymbol{y}\right)\mathrm{d}F(\boldsymbol{y})}{\int_{\mathcal{Y}}\exp(\boldsymbol{\theta}^T\boldsymbol{y})\,\mathrm{d}F(\boldsymbol{y})}\ . \end{aligned}$$

By re-introducing the normalizing function $b$,

$$-\frac{\partial b}{\partial \boldsymbol{\theta}} = \frac{\int_{\mathcal{Y}} \boldsymbol{y} \exp\left(b + \boldsymbol{\theta}^T \boldsymbol{y}\right) \mathrm{d}F(\boldsymbol{y})}{\int_{\mathcal{Y}} \exp(b + \boldsymbol{\theta}^T \boldsymbol{y}) \, \mathrm{d}F(\boldsymbol{y})} = \int_{\mathcal{Y}} \boldsymbol{y} \exp\left(b + \boldsymbol{\theta}^T \boldsymbol{y}\right) \mathrm{d}F(\boldsymbol{y}) \, .$$

Thus, by the mean constraint (2.7),

$$-\frac{\partial b}{\partial \boldsymbol{\theta}} = \boldsymbol{\mu} \tag{2.14}$$

which is the same identity found for the multivariate exponential family. Therefore,

$$\frac{\partial l}{\partial \boldsymbol{\theta}} = \boldsymbol{Y} + \frac{\partial b}{\partial \boldsymbol{\theta}} = \boldsymbol{Y} - \boldsymbol{\mu} \, . \tag{2.15}$$

Furthermore, differentiating (2.14) with respect to $\boldsymbol{\theta}$ again gives

$$\begin{aligned}
\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\theta}} &= \frac{\partial}{\partial \boldsymbol{\theta}} \left( -\frac{\partial b}{\partial \boldsymbol{\theta}} \right) \\
&= \int_{\mathcal{Y}} \frac{\partial}{\partial \boldsymbol{\theta}} \left\{ \boldsymbol{y} \exp\left(b + \boldsymbol{\theta}^T \boldsymbol{y}\right) \right\} \mathrm{d}F(\boldsymbol{y}) \\
&= \int_{\mathcal{Y}} \left( \frac{\partial b}{\partial \boldsymbol{\theta}} + \boldsymbol{y} \right) \exp\left(b + \boldsymbol{\theta}^T \boldsymbol{y}\right) \boldsymbol{y}^T \, \mathrm{d}F(\boldsymbol{y}) \\
\implies \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\theta}} &= \int_{\mathcal{Y}} (\boldsymbol{y} - \boldsymbol{\mu}) \, \boldsymbol{y}^T \exp\left(b + \boldsymbol{\theta}^T \boldsymbol{y}\right) \mathrm{d}F(\boldsymbol{y}) \, .
\end{aligned}$$

To simplify the expression above, let $\Sigma_{\boldsymbol{Y}}(\boldsymbol{X}; \boldsymbol{\beta}, F)$ denote the conditional covariance matrix, of $\boldsymbol{Y}$ given $\boldsymbol{X}$, under $(\boldsymbol{\beta}, F)$. By definition,

$$\begin{aligned}
\Sigma_{\boldsymbol{Y}}(\boldsymbol{X}; \boldsymbol{\beta}, F) &= \mathbb{E}_{\boldsymbol{\beta}, F} \left[ (\boldsymbol{Y} - \boldsymbol{\mu})(\boldsymbol{Y} - \boldsymbol{\mu})^T | \boldsymbol{X} \right] \\
&= \mathbb{E}_{\boldsymbol{\beta}, F} \left[ (\boldsymbol{Y} - \boldsymbol{\mu}) \boldsymbol{Y}^T | \boldsymbol{X} \right] - \mathbb{E}_{\boldsymbol{\beta}, F} \left[ (\boldsymbol{Y} - \boldsymbol{\mu}) \boldsymbol{\mu}^T | \boldsymbol{X} \right] \\
&= \mathbb{E}_{\boldsymbol{\beta}, F} \left[ (\boldsymbol{Y} - \boldsymbol{\mu}) \boldsymbol{Y}^T | \boldsymbol{X} \right] - \mathbb{E}_{\boldsymbol{\beta}, F} \left[ (\boldsymbol{Y} - \boldsymbol{\mu}) | \boldsymbol{X} \right] \boldsymbol{\mu}^T \\
&= \mathbb{E}_{\boldsymbol{\beta}, F} \left[ (\boldsymbol{Y} - \boldsymbol{\mu}) \boldsymbol{Y}^T | \boldsymbol{X} \right] - \boldsymbol{0} \cdot \boldsymbol{\mu}^T \\
&= \mathbb{E}_{\boldsymbol{\beta}, F} \left[ (\boldsymbol{Y} - \boldsymbol{\mu}) \boldsymbol{Y}^T | \boldsymbol{X} \right] \, .
\end{aligned}$$

This can be seen as the model-based estimate for the conditional covariance matrix. As a result, using our vector notation we have that

$$\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\theta}} = \Sigma_{\boldsymbol{Y}}(\boldsymbol{X}; \boldsymbol{\beta}, F) \tag{2.16}$$

$$\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\mu}} = \left[ \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\theta}} \right]^{-1} = \Sigma_{\boldsymbol{Y}}^{-1}(\boldsymbol{X}; \boldsymbol{\beta}, F) \, . \tag{2.17}$$

Substituting (2.15) and (2.17) into (2.13),

$$S_{\boldsymbol{\beta}, F}(\boldsymbol{X}, \boldsymbol{Y}) = \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}} \, \Sigma_{\boldsymbol{Y}}^{-1}(\boldsymbol{X}; \boldsymbol{\beta}, F) \left( \boldsymbol{Y} - \boldsymbol{\mu} \left( \boldsymbol{X}^T \boldsymbol{\beta} \right) \right) \, , \tag{2.18}$$

where

$$\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}} = \begin{bmatrix} \frac{\partial \mu_1}{\partial \beta_1} & \frac{\partial \mu_2}{\partial \beta_1} & \cdots & \frac{\partial \mu_K}{\partial \beta_1} \\ \frac{\partial \mu_1}{\partial \beta_2} & \frac{\partial \mu_2}{\partial \beta_2} & \cdots & \frac{\partial \mu_K}{\partial \beta_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \mu_1}{\partial \beta_Q} & \frac{\partial \mu_2}{\partial \beta_Q} & \cdots & \frac{\partial \mu_K}{\partial \beta_Q} \end{bmatrix} = \begin{bmatrix} \frac{\partial \mu_{(1)}}{\partial \boldsymbol{\beta}_{(1')}} & \frac{\partial \mu_{(2)}}{\partial \boldsymbol{\beta}_{(1')}} & \cdots & \frac{\partial \mu_{(K)}}{\partial \boldsymbol{\beta}_{(1')}} \\ \frac{\partial \mu_{(1)}}{\partial \boldsymbol{\beta}_{(2')}} & \frac{\partial \mu_{(2)}}{\partial \boldsymbol{\beta}_{(2')}} & \cdots & \frac{\partial \mu_{(K)}}{\partial \boldsymbol{\beta}_{(2')}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \mu_{(1)}}{\partial \boldsymbol{\beta}_{(K')}} & \frac{\partial \mu_{(2)}}{\partial \boldsymbol{\beta}_{(K')}} & \cdots & \frac{\partial \mu_{(K)}}{\partial \boldsymbol{\beta}_{(K')}} \end{bmatrix} = \begin{bmatrix} \frac{\partial \boldsymbol{\mu}_{(1')}}{\partial \boldsymbol{\beta}_{(1')}} & \frac{\partial \boldsymbol{\mu}_{(2')}}{\partial \boldsymbol{\beta}_{(1')}} & \cdots & \frac{\partial \boldsymbol{\mu}_{(K')}}{\partial \boldsymbol{\beta}_{(1')}} \\ \frac{\partial \boldsymbol{\mu}_{(1')}}{\partial \boldsymbol{\beta}_{(2')}} & \frac{\partial \boldsymbol{\mu}_{(2')}}{\partial \boldsymbol{\beta}_{(2')}} & \cdots & \frac{\partial \boldsymbol{\mu}_{(K')}}{\partial \boldsymbol{\beta}_{(2')}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \boldsymbol{\mu}_{(1')}}{\partial \boldsymbol{\beta}_{(K')}} & \frac{\partial \boldsymbol{\mu}_{(2')}}{\partial \boldsymbol{\beta}_{(K')}} & \cdots & \frac{\partial \boldsymbol{\mu}_{(K')}}{\partial \boldsymbol{\beta}_{(K')}} \end{bmatrix}$$

is a $q \times K$ matrix. Generally, let us define

$$D(\boldsymbol{X};\boldsymbol{\beta}) := \frac{\partial \boldsymbol{\mu}(\boldsymbol{X}^T\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \tag{2.19}$$

where the evaluation of the derivative is at the parameter $\boldsymbol{\beta}$. Therefore, we find the score expression for $\boldsymbol{\beta}$ to be

$$S_{\boldsymbol{\beta},F}(\boldsymbol{X},\boldsymbol{Y}) = D(\boldsymbol{X};\boldsymbol{\beta})\,\Sigma_{\boldsymbol{Y}}^{-1}(\boldsymbol{X};\boldsymbol{\beta},F)\left(\boldsymbol{Y} - \boldsymbol{\mu}\left(\boldsymbol{X}^T\boldsymbol{\beta}\right)\right)\ . \tag{2.20}$$

To extend on this, we can simplify and find a more explicit expression for (2.19). Firstly, let us note that

$$\begin{aligned}\frac{\partial \mu_{(k)}(\boldsymbol{X}_{(k)}^T\boldsymbol{\beta}_{(k')})}{\partial \boldsymbol{\beta}_{(j')}} &= \frac{\partial(\boldsymbol{X}_{(k)}^T\boldsymbol{\beta}_{(k')})}{\partial \boldsymbol{\beta}_{(j')}}\frac{\partial \mu_{(k)}(\boldsymbol{X}_{(k)}^T\boldsymbol{\beta}_{(k')})}{\partial\left(\boldsymbol{X}_{(k)}^T\boldsymbol{\beta}_{(k')}\right)} \\ &= \frac{\partial(\boldsymbol{X}_{(k)}^T\boldsymbol{\beta}_{(k')})}{\partial \boldsymbol{\beta}_{(j')}}\mu'_{(k)}(\boldsymbol{X}_{(k)}^T\boldsymbol{\beta}_{(k')})\end{aligned}$$

for $j' \in \{1,\dots,K'\}, k' \in \{1,\dots,K'\}$ and $k = 1,\dots,K$. Therefore, noting that $j' = k'$ denotes that the coefficients are shared, we have that for $k = 1,\dots,K$,

$$\frac{\partial \mu_{(k)}(\boldsymbol{X}_{(k)}^T\boldsymbol{\beta}_{(k')})}{\partial \boldsymbol{\beta}_{(j')}} = \begin{cases}\boldsymbol{X}_{(k)}\mu'_{(k)}(\boldsymbol{X}_{(k)}^T\boldsymbol{\beta}_{(k')}) & j' = k' \\ 0 & j' \neq k'\end{cases}\ .$$

Now consider the vector $\boldsymbol{\mu}$ of components that have the same mean model, letting these components belong to an index set $I_{k'}$ of size $M_{k'}$ for the $k'$-th mean model. Then, if $j' = k'$ we have that

$$\begin{aligned}\frac{\partial \boldsymbol{\mu}_{(k')}}{\partial \boldsymbol{\beta}_{(j')}} &= \left[\frac{\partial \mu_{(k',1)}}{\partial \boldsymbol{\beta}_{(j')}} \quad \frac{\partial \mu_{(k',2)}}{\partial \boldsymbol{\beta}_{(j')}} \quad \dots \quad \frac{\partial \mu_{(k',M_{k'})}}{\partial \boldsymbol{\beta}_{(j')}}\right] \\ &= \left[\boldsymbol{X}_{(k',1)}\mu'_{(k',1)} \quad \boldsymbol{X}_{(k',2)}\mu'_{(k',2)} \quad \dots \quad \boldsymbol{X}_{(k',M_{k'})}\mu'_{(k',M_{k'})}\right] \\ &= \boldsymbol{X}_{(k')}\mathrm{Diag}(\boldsymbol{\mu}'_{(k')})\end{aligned}$$

where

$$\boldsymbol{X}_{(k')} = \left[\boldsymbol{X}_{(k',1)}, \boldsymbol{X}_{(k',2)}, \dots, \boldsymbol{X}_{(k',M_{k'})}\right] \in \mathbb{R}^{q_k \times M_{k'}}\ .$$

To re-iterate the notation, $\boldsymbol{X}_{(k',m)}$ for $k' \in \{1,\dots,K'\}$ and $m \in I_{k'}$ refers to set of covariates associated with the $m$-th component in the $k'$-th mean model, which correspond to some covariate vector $\boldsymbol{X}_{(k)}$ for $k = 1,\dots,K$. This notation of evaluating the derivative by collecting the mean models is convenient because it allows for the final expression to be given in terms of a block diagonal matrix. In the case

where $j' \neq k'$, the expression is a vector of zeros. Therefore, we can re-express $D(\boldsymbol{X};\boldsymbol{\beta})$ as follows

$$
\begin{aligned}
D(\boldsymbol{X};\boldsymbol{\beta}) &= \begin{bmatrix} \frac{\partial \mu_{(1')}}{\partial \boldsymbol{\beta}_{(1')}} & \frac{\partial \mu_{(2')}}{\partial \boldsymbol{\beta}_{(1')}} & \cdots & \frac{\partial \mu_{(K')}}{\partial \boldsymbol{\beta}_{(1')}} \\ \frac{\partial \mu_{(1')}}{\partial \boldsymbol{\beta}_{(2')}} & \frac{\partial \mu_{(2')}}{\partial \boldsymbol{\beta}_{(2')}} & \cdots & \frac{\partial \mu_{(K')}}{\partial \boldsymbol{\beta}_{(2')}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \mu_{(1')}}{\partial \boldsymbol{\beta}_{(K')}} & \frac{\partial \mu_{(2')}}{\partial \boldsymbol{\beta}_{(K')}} & \cdots & \frac{\partial \mu_{(K')}}{\partial \boldsymbol{\beta}_{(K')}} \end{bmatrix} \\
&= \begin{bmatrix} \boldsymbol{X}_{(1')}\mathrm{Diag}(\boldsymbol{\mu}'_{(1')}) & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{X}_{(2')}\mathrm{Diag}(\boldsymbol{\mu}'_{(2')}) & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{X}_{(K')}\mathrm{Diag}(\boldsymbol{\mu}'_{(K')}) \end{bmatrix} \\
&= \begin{bmatrix} \boldsymbol{X}_{(1')} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{X}_{(2')} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{X}_{(K')} \end{bmatrix} \begin{bmatrix} \mu'_{(1)} & 0 & \cdots & 0 \\ 0 & \mu'_{(2)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mu'_{(K)} \end{bmatrix} .
\end{aligned}
$$

Here the first matrix is a $q \times K$ block diagonal matrix, as $\sum_{k=1}^{K'} q_k = q, \sum_{k'=1}^{K'} M_{k'} = K$.

Therefore, we obtain the general form

$$
D(\boldsymbol{X};\boldsymbol{\beta}) = \mathrm{Diag}(\boldsymbol{X}')\ \mathrm{Diag}(\boldsymbol{\mu}') \tag{2.21}
$$

where $\mathrm{Diag}(\boldsymbol{X}')$ is a block diagonal of $\boldsymbol{X}' = (\boldsymbol{X}_{(1')}, \ldots, \boldsymbol{X}_{(k')})^T \in \mathbb{R}^Q$, $Q \geq q$. As a result, the score function for $\boldsymbol{\beta}$ can be expressed as

$$
S_{\boldsymbol{\beta},F}(\boldsymbol{X},\boldsymbol{Y}) = D(\boldsymbol{X};\boldsymbol{\beta})\ \Sigma_{\boldsymbol{Y}}^{-1}(\boldsymbol{X};\boldsymbol{\beta},F)\left(\boldsymbol{Y} - \boldsymbol{\mu}\left(\boldsymbol{X}^T\boldsymbol{\beta}\right)\right) \tag{2.22}
$$

$$
S_{\boldsymbol{\beta},F}(\boldsymbol{X},\boldsymbol{Y}) = \mathrm{Diag}(\boldsymbol{X}')\ \mathrm{Diag}(\boldsymbol{\mu}')\ \Sigma_{\boldsymbol{Y}}^{-1}(\boldsymbol{X};\boldsymbol{\beta},F)\left(\boldsymbol{Y} - \boldsymbol{\mu}\left(\boldsymbol{X}^T\boldsymbol{\beta}\right)\right) . \tag{2.23}
$$

The expectation of $S_{\boldsymbol{\beta},F}$ for any $(\boldsymbol{\beta}, F)$ conditional on $\boldsymbol{X}$ is given by

$$
\begin{aligned}
\mathbb{E}_{\boldsymbol{\beta},F}\left[S_{\boldsymbol{\beta},F}(\boldsymbol{X},\boldsymbol{Y})|\boldsymbol{X}\right] &= \mathbb{E}_{\boldsymbol{\beta},F}\left[D(\boldsymbol{X};\boldsymbol{\beta})\Sigma_{\boldsymbol{Y}}^{-1}(\boldsymbol{X};\boldsymbol{\beta},F)\left(\boldsymbol{Y}-\boldsymbol{\mu}\right)|\boldsymbol{X}\right] \\
&= D(\boldsymbol{X};\boldsymbol{\beta})\Sigma_{\boldsymbol{Y}}^{-1}(\boldsymbol{X};\boldsymbol{\beta},F)\mathbb{E}_{\boldsymbol{\beta},F}\left[\left(\boldsymbol{Y}-\boldsymbol{\mu}\right)|\boldsymbol{X}\right] \\
&= 0
\end{aligned}
$$

Thus for any $(\boldsymbol{\beta}, F)$, the score equation for $\boldsymbol{\beta}$ is unbiased,

$$
\mathbb{E}_{\boldsymbol{\beta},F}\left[S_{\boldsymbol{\beta},F}(\boldsymbol{X},\boldsymbol{Y})\right] = \mathbb{E}^{\mathcal{X}}\left[\mathbb{E}_{\boldsymbol{\beta},F}\left(S_{\boldsymbol{\beta},F}(\boldsymbol{X},\boldsymbol{Y})|\boldsymbol{X}\right)\right] = 0 . \tag{2.24}
$$

The expressions handle a variety of cases such as; all components sharing coefficients, some components sharing coefficients and others not, or all components not sharing any coefficients or mean models. To further see this and help understand the notation and expressions, we have included some generic examples in Appendix A.3 for the case of $K = 3$ components. Importantly, in the case that the response components are independent, we recover the expressions given in Huang (2014).

### 2.3.2 Score Function for $F$

To define a score function for the infinite-dimensional parameter $F$, we need to use parametric submodels and take directional derivatives on the score surface in a particular direction. Suppose a submodel $t \mapsto F_t$ which passes through $F$ at $t = 0$ is a perturbation of $F$ in some fixed direction $h \in \mathcal{H}$ with

magnitude $t$. A score function for $F$ in the direction of $h$ along this parametric submodel denoted by $F_t$, evaluated at $(\boldsymbol{\beta}, F)$ is defined by

$$A_{\boldsymbol{\beta},F} h(\boldsymbol{X}, \boldsymbol{Y}) := \frac{\partial}{\partial t} \ell(\boldsymbol{\beta}, F_t)(\boldsymbol{X}, \boldsymbol{Y}) \Big|_{t=0} , \tag{2.25}$$

where $A_{\boldsymbol{\beta},F} : \mathcal{H} \to \ell^\infty(\mathcal{H})$ is the score operator. As seen in Murphy and Van Der Vaart (2001), for a reference distribution $F$ and some bounded function $h$ and $t$ sufficiently close to 0, we can consider submodels $F_t$ with probability densities of the form

$$dF_t(\boldsymbol{y}) = \left(1 + t\left(h(\boldsymbol{y}) - \int h(\boldsymbol{y}) \, \mathrm{d}F(\boldsymbol{y})\right)\right) dF(\boldsymbol{y}) . \tag{2.26}$$

Huang (2011) notes that any distribution function $F$ can be reparametrised by writing its density as $dF = dP / \int_{\mathcal{Z}} \mathrm{d}P$ for some measure $P$, removing the need for normalisation. As a result, submodels $P_t$ with densities of the form $dP_t = (1 + th)dP$ also suffice for $t$ sufficiently close to 0 and bounded functions $h$.

Similar to Huang (2014), let's consider inserting into the log-likelihood the parametric submodel $F_t$ with densities of the form

$$dF_t(\boldsymbol{y}) = (1 + th(\boldsymbol{y}))dF(\boldsymbol{y}) . \tag{2.27}$$

Here we take $\mathcal{H}$ to be the class of all left indicators on the hyperrectangle $\mathcal{Y}$ denoted as $\mathcal{H}_L := \{\mathbb{1}(\boldsymbol{y} \leq \boldsymbol{r}) : \boldsymbol{r} \in \mathcal{Y}\}$. No normalization or tilting is required in the submodel due to the invariance of the log-likelihood, and indeed $F_t$ is a valid measure for $t$ sufficiently close to 0. The log-likelihood function for a generic data pair $(\boldsymbol{X}, \boldsymbol{Y})$ corresponding to this submodel is given by

$$\ell(\boldsymbol{\beta}, F_t | \boldsymbol{X}, \boldsymbol{Y}) = \log(1 + th(\boldsymbol{Y})) + \log dF(\boldsymbol{Y}) + b(\boldsymbol{X}; \boldsymbol{\beta}, F_t) + \boldsymbol{\theta}^T(\boldsymbol{X}; \boldsymbol{\beta}, F_t)\boldsymbol{Y} .$$

Thus, we can express the score function for $F$ in the direction of $h$ as

$$\frac{\partial}{\partial t} \ell(\boldsymbol{\beta}, F_t | \boldsymbol{X}, \boldsymbol{Y}) \Big|_{t=0} = h(\boldsymbol{Y}) + \frac{\partial b_t}{\partial t} \Big|_{t=0} + \frac{\partial \boldsymbol{\theta}_t}{\partial t} \Big|_{t=0} \boldsymbol{Y} , \tag{2.28}$$

where for notational convenience, $b_t = b(\boldsymbol{X}; \boldsymbol{\beta}, F_t), \boldsymbol{\theta}_t = \boldsymbol{\theta}(\boldsymbol{X}; \boldsymbol{\beta}, F_t)$. Note the partial derivative of $\boldsymbol{\theta}$ with respect to $t$ is a row vector using our denominator convention and that normalization and mean constraints are still enforced,

$$b_t = b(\boldsymbol{X}, \boldsymbol{\beta}, F_t) = -\log\left\{\int_{\mathcal{Y}} \exp\left(\boldsymbol{\theta}(\boldsymbol{X}; \boldsymbol{\beta}, F_t)^T \boldsymbol{y}\right) (1 + th(\boldsymbol{y})) \, \mathrm{d}F(\boldsymbol{y})\right\}$$

and $\boldsymbol{\theta}(\boldsymbol{X}; \boldsymbol{\beta}, F_t)$ satisfies

$$\boldsymbol{\mu} = \int_{\mathcal{Y}} \boldsymbol{y} \exp\left(b_t + \boldsymbol{\theta}_t^T \boldsymbol{y}\right) (1 + th(\boldsymbol{y})) \, \mathrm{d}F(\boldsymbol{y}) \tag{2.29}$$

using the notation defined in (2.6). To find the partial derivative of $\boldsymbol{\theta}$ with respect to $t$ we can first perform implicit differentiation of the mean constraint re-expressed as

$$\mathbf{0} = \int_{\mathcal{Y}} (\boldsymbol{y} - \boldsymbol{\mu}) \exp\left(b_t + \boldsymbol{\theta}_t^T \boldsymbol{y}\right) (1 + th(\boldsymbol{y})) \, \mathrm{d}F(\boldsymbol{y}) .$$

Note that we can bring the partial derivative inside of the integral by the Leibniz integral rule, due to the Assumptions in A.2.

$$\mathbf{0} = \frac{\partial}{\partial t} \int_{\mathcal{Y}} (\boldsymbol{y} - \boldsymbol{\mu}) \exp\big(b_t + \boldsymbol{\theta}_t^T \boldsymbol{y}\big) \, (1 + th(\boldsymbol{y})) \, \mathrm{d}F(\boldsymbol{y})$$

$$\mathbf{0} = \int_{\mathcal{Y}} (\boldsymbol{y} - \boldsymbol{\mu}) \frac{\partial}{\partial t} \exp\big(b_t + \boldsymbol{\theta}_t^T \boldsymbol{y}\big)(1 + th(\boldsymbol{y})) \, \mathrm{d}F(\boldsymbol{y}) \tag{2.30}$$

$$+ \int_{\mathcal{Y}} (\boldsymbol{y} - \boldsymbol{\mu}) \exp\big(b_t + \boldsymbol{\theta}_t^T \boldsymbol{y}\big) \frac{\partial}{\partial t} \, (1 + th(\boldsymbol{y})) \, \mathrm{d}F(\boldsymbol{y}) \, . \tag{2.31}$$

Considering the term (2.30),

$$\begin{aligned}
&\int_{\mathcal{Y}} (\boldsymbol{y} - \boldsymbol{\mu}) \frac{\partial}{\partial t} \exp\big(b_t + \boldsymbol{\theta}_t^T \boldsymbol{y}\big)(1 + th(\boldsymbol{y})) \, \mathrm{d}F(\boldsymbol{y}) \\
&= \int_{\mathcal{Y}} (\boldsymbol{y} - \boldsymbol{\mu}) \frac{\partial \boldsymbol{\theta}_t}{\partial t} \frac{\partial \exp\big(b_t + \boldsymbol{\theta}_t^T \boldsymbol{y}\big)}{\partial \boldsymbol{\theta}_t} \, \mathrm{d}F(\boldsymbol{y}) \\
&= \int_{\mathcal{Y}} (\boldsymbol{y} - \boldsymbol{\mu}) \frac{\partial \boldsymbol{\theta}_t}{\partial t} \boldsymbol{y} \exp\big(b_t + \boldsymbol{\theta}_t^T \boldsymbol{y}\big)(1 + th(\boldsymbol{y})) \, \mathrm{d}F(\boldsymbol{y}) \\
&= \int_{\mathcal{Y}} (\boldsymbol{y} - \boldsymbol{\mu}) \, \boldsymbol{y}^T \left[\frac{\partial \boldsymbol{\theta}_t}{\partial t}\right]^T \exp\big(b_t + \boldsymbol{\theta}_t^T \boldsymbol{y}\big)(1 + th(\boldsymbol{Y})) \, \mathrm{d}F(\boldsymbol{y}) \\
&= \left(\int_{\mathcal{Y}} (\boldsymbol{y} - \boldsymbol{\mu}) \, \boldsymbol{y}^T \exp\big(b_t + \boldsymbol{\theta}_t^T \boldsymbol{y}\big)(1 + th(\boldsymbol{Y})) \, \mathrm{d}F(\boldsymbol{y})\right) \left[\frac{\partial \boldsymbol{\theta}_t}{\partial t}\right]^T \, .
\end{aligned} \tag{2.32}$$

Considering the term (2.31),

$$\int_{\mathcal{Y}} (\boldsymbol{y} - \boldsymbol{\mu}) \exp\big(b_t + \boldsymbol{\theta}_t^T \boldsymbol{y}\big) \frac{\partial}{\partial t} \, (1 + th(\boldsymbol{y})) \, \mathrm{d}F(\boldsymbol{y}) = \int_{\mathcal{Y}} (\boldsymbol{y} - \boldsymbol{\mu}) \exp\big(b_t + \boldsymbol{\theta}_t^T \boldsymbol{y}\big) h(\boldsymbol{y}) \, \mathrm{d}F(\boldsymbol{y}) \, . \tag{2.33}$$

Therefore,

$$\mathbf{0} = \int_{\mathcal{Y}} (\boldsymbol{y} - \boldsymbol{\mu}) \exp\big(b_t + \boldsymbol{\theta}_t^T \boldsymbol{y}\big) h(\boldsymbol{y}) \, \mathrm{d}F(\boldsymbol{y}) \tag{2.34}$$

$$+ \left(\int_{\mathcal{Y}} (\boldsymbol{y} - \boldsymbol{\mu}) \, \boldsymbol{y}^T \exp\big(b_t + \boldsymbol{\theta}_t^T \boldsymbol{y}\big)(1 + th(\boldsymbol{Y})) \, \mathrm{d}F(\boldsymbol{y})\right) \left[\frac{\partial \boldsymbol{\theta}_t}{\partial t}\right]^T \, . \tag{2.35}$$

Evaluating the expression above at $t = 0$ and simplifying the integrals,

$$\mathbf{0} = \int_{\mathcal{Y}} (\boldsymbol{y} - \boldsymbol{\mu}) \exp\big(b + \boldsymbol{\theta}^T \boldsymbol{y}\big) h(\boldsymbol{y}) \, \mathrm{d}F(\boldsymbol{y}) + \left(\int_{\mathcal{Y}} (\boldsymbol{y} - \boldsymbol{\mu}) \, \boldsymbol{y}^T \exp\big(b + \boldsymbol{\theta}^T \boldsymbol{y}\big) \, \mathrm{d}F(\boldsymbol{y})\right) \left[\left.\frac{\partial \boldsymbol{\theta}_t}{\partial t}\right|_{t=0}\right]^T$$

$$\mathbf{0} = \mathbb{E}_{\boldsymbol{\beta}, F} \left[h(\boldsymbol{Y}) \, (\boldsymbol{Y} - \boldsymbol{\mu}) \, | \boldsymbol{X}\right] + \Sigma_{\boldsymbol{Y}}(\boldsymbol{X}; \boldsymbol{\beta}, F) \left[\left.\frac{\partial \boldsymbol{\theta}_t}{\partial t}\right|_{t=0}\right]^T \, .$$

Re-arranging, we obtain

$$\begin{aligned}
\left[\left.\frac{\partial \boldsymbol{\theta}_t}{\partial t}\right|_{t=0}\right]^T &= -\Sigma_{\boldsymbol{Y}}^{-1}(\boldsymbol{X}; \boldsymbol{\beta}, F) \mathbb{E}_{\boldsymbol{\beta}, F} \left[h(\boldsymbol{Y}) \, (\boldsymbol{Y} - \boldsymbol{\mu}) \, | \boldsymbol{X}\right] \\
\left.\frac{\partial \boldsymbol{\theta}_t}{\partial t}\right|_{t=0} &= - \left(\mathbb{E}_{\boldsymbol{\beta}, F} \left[h(\boldsymbol{Y}) \, (\boldsymbol{Y} - \boldsymbol{\mu}) \, | \boldsymbol{X}\right]\right)^T \Sigma_{\boldsymbol{Y}}^{-1}(\boldsymbol{X}; \boldsymbol{\beta}, F) \, .
\end{aligned} \tag{2.36}$$

To find the partial derivative of $b$ with respect to $t$ we can differentiate the normalisation constraint

$$\begin{aligned}
b_t &= - \log \int_{\mathcal{Y}} \exp\big(\boldsymbol{\theta}_t^T \boldsymbol{y}\big)(1 + th(\boldsymbol{Y})) \, \mathrm{d}F(\boldsymbol{y}) \\
\frac{\partial b_t}{\partial t} &= - \frac{\frac{\partial}{\partial t} \int_{\mathcal{Y}} \exp\big(\boldsymbol{\theta}_t^T \boldsymbol{y}\big)(1 + th(\boldsymbol{Y})) \, \mathrm{d}F(\boldsymbol{y})}{\int_{\mathcal{Y}} \exp\big(\boldsymbol{\theta}_t^T \boldsymbol{y}\big)(1 + th(\boldsymbol{Y})) \, \mathrm{d}F(\boldsymbol{y})} \, .
\end{aligned} \tag{2.37}$$

Considering the numerator and bringing in the partial derivative by Leibniz integral rule,

$$\begin{aligned}\frac{\partial}{\partial t}\int_{\mathcal{Y}} \exp\big(\boldsymbol{\theta}_t^T \boldsymbol{y}\big)(1+th(\boldsymbol{y}))\,\mathrm{d}F(\boldsymbol{y}) &= \int_{\mathcal{Y}} \frac{\partial}{\partial t}\exp\big(\boldsymbol{\theta}_t^T \boldsymbol{y}\big)(1+th(\boldsymbol{y}))\,\mathrm{d}F(\boldsymbol{y}) \\ &\quad + \int_{\mathcal{Y}} \exp\big(\boldsymbol{\theta}_t^T \boldsymbol{y}\big)\frac{\partial}{\partial t}\left(1+th(\boldsymbol{y})\right)\,\mathrm{d}F(\boldsymbol{y}) \\ &= \int_{\mathcal{Y}} \frac{\partial \boldsymbol{\theta}_t}{\partial t}\frac{\partial \exp\big(\boldsymbol{\theta}_t^T \boldsymbol{y}\big)}{\partial \boldsymbol{\theta}_t}\,\mathrm{d}F(\boldsymbol{y}) + \int_{\mathcal{Y}} \exp\big(\boldsymbol{\theta}_t^T \boldsymbol{y}\big)h(\boldsymbol{y})\,\mathrm{d}F(\boldsymbol{y}) \\ &= \frac{\partial \boldsymbol{\theta}_t}{\partial t}\int_{\mathcal{Y}} \boldsymbol{y}\exp\big(\boldsymbol{\theta}_t^T \boldsymbol{y}\big)(1+th(\boldsymbol{y}))\,\mathrm{d}F(\boldsymbol{y}) + \int_{\mathcal{Y}} \exp\big(\boldsymbol{\theta}_t^T \boldsymbol{y}\big)h(\boldsymbol{y})\,\mathrm{d}F(\boldsymbol{y})\,.\end{aligned}$$

Evaluating the partial derivative (2.37) at $t=0$,

$$\left.\frac{\partial b_t}{\partial t}\right|_{t=0} = -\frac{\left(\left.\frac{\partial \boldsymbol{\theta}}{\partial t}\right|_{t=0}\right)\int_{\mathcal{Y}} \boldsymbol{y}\exp\big(\boldsymbol{\theta}^T \boldsymbol{y}\big)\,\mathrm{d}F(\boldsymbol{y}) + \int_{\mathcal{Y}} \exp\big(\boldsymbol{\theta}^T \boldsymbol{y}\big)h(\boldsymbol{y})\,\mathrm{d}F(\boldsymbol{y})}{\int_{\mathcal{Y}} \exp\big(\boldsymbol{\theta}^T \boldsymbol{y}\big)\,\mathrm{d}F(\boldsymbol{y})}\,.$$

Multiplying the numerator and denominator by $\exp(b) = \exp(b(\boldsymbol{X};\boldsymbol{\beta},F))$ to normalize the expression,

$$\begin{aligned}\left.\frac{\partial b_t}{\partial t}\right|_{t=0} &= -\frac{\left(\left.\frac{\partial \boldsymbol{\theta}}{\partial t}\right|_{t=0}\right)\int_{\mathcal{Y}} \boldsymbol{y}\exp\big(b+\boldsymbol{\theta}^T \boldsymbol{y}\big)\,\mathrm{d}F(\boldsymbol{y}) + \int_{\mathcal{Y}} \exp\big(b+\boldsymbol{\theta}^T \boldsymbol{y}\big)h(\boldsymbol{y})\,\mathrm{d}F(\boldsymbol{y})}{\int_{\mathcal{Y}} \exp\big(b+\boldsymbol{\theta}^T \boldsymbol{y}\big)\,\mathrm{d}F(\boldsymbol{y})} \\ &= -\left(\left.\frac{\partial \boldsymbol{\theta}}{\partial t}\right|_{t=0}\right)\int_{\mathcal{Y}} \boldsymbol{y}\exp\big(b+\boldsymbol{\theta}^T \boldsymbol{y}\big)\,\mathrm{d}F(\boldsymbol{y}) - \int_{\mathcal{Y}} \exp\big(b+\boldsymbol{\theta}^T \boldsymbol{y}\big)h(\boldsymbol{y})\,\mathrm{d}F(\boldsymbol{y}) \\ &= -\left(\left.\frac{\partial \boldsymbol{\theta}}{\partial t}\right|_{t=0}\right)\boldsymbol{\mu} - \mathbb{E}_{\boldsymbol{\beta},F}\left[h(\boldsymbol{Y})|\boldsymbol{X}\right]\,. \end{aligned} \tag{2.38}$$

Bringing together (2.36) and (2.38) into (2.28),

$$\begin{aligned}\left.\frac{\partial}{\partial t}\ell(\boldsymbol{\beta},F_t|\boldsymbol{X},\boldsymbol{Y})\right|_{t=0} &= h(\boldsymbol{Y}) + \left.\frac{\partial b_t}{\partial t}\right|_{t=0} + \left.\frac{\partial \boldsymbol{\theta}_t}{\partial t}\right|_{t=0}\boldsymbol{Y} \\ &= h(\boldsymbol{Y}) - \mathbb{E}_{\boldsymbol{\beta},F}\left[h(\boldsymbol{Y})|\boldsymbol{X}\right] + \left.\frac{\partial \boldsymbol{\theta}_t}{\partial t}\right|_{t=0}(\boldsymbol{Y}-\boldsymbol{\mu}) \\ &= h(\boldsymbol{Y}) - \mathbb{E}_{\boldsymbol{\beta},F}\left[h(\boldsymbol{Y})|\boldsymbol{X}\right] - \left(\mathbb{E}_{\boldsymbol{\beta},F}\left[h(\boldsymbol{Y})\left(\boldsymbol{Y}-\boldsymbol{\mu}\right)|\boldsymbol{X}\right]\right)^T \Sigma_{\boldsymbol{Y}}^{-1}(\boldsymbol{X};\boldsymbol{\beta},F)\left(\boldsymbol{Y}-\boldsymbol{\mu}\right)\,.\end{aligned}$$

To simplify notation and be consistent with Huang (2014), define operators $B_{\boldsymbol{\beta},F}$ and $C_{\boldsymbol{\beta},F}$ by

$$B_{\boldsymbol{\beta},F}h(\boldsymbol{X}) = \mathbb{E}_{\boldsymbol{\beta},F}[h(\boldsymbol{Y})|\boldsymbol{X}] \tag{2.39}$$

$$C_{\boldsymbol{\beta},F}h(\boldsymbol{X}) = \mathbb{E}_{\boldsymbol{\beta},F}\left[h(\boldsymbol{Y})\left(\boldsymbol{Y}-\boldsymbol{\mu}\right)^T|\boldsymbol{X}\right]\Sigma_{\boldsymbol{Y}}^{-\frac{1}{2}}(\boldsymbol{X};\boldsymbol{\beta},F)\,. \tag{2.40}$$

Here we note that $\Sigma_{\boldsymbol{Y}}^{-\frac{1}{2}}(\boldsymbol{X};\boldsymbol{\beta},F)$ is the matrix such that $\Sigma_{\boldsymbol{Y}}^{-\frac{1}{2}}(\boldsymbol{X};\boldsymbol{\beta},F)\Sigma_{\boldsymbol{Y}}^{-\frac{1}{2}}(\boldsymbol{X};\boldsymbol{\beta},F) = \Sigma_{\boldsymbol{Y}}^{-1}(\boldsymbol{X};\boldsymbol{\beta},F)$. This is because $\Sigma_{\boldsymbol{Y}}(\boldsymbol{X};\boldsymbol{\beta},F)$ is a symmetric, positive semi-definite matrix and invertible by assumption, which implies positive-definite and thus $\Sigma_{\boldsymbol{Y}}^{-1}(\boldsymbol{X};\boldsymbol{\beta},F)$ is positive-definite. Note we can bring the transpose into our expectation using our vector integral notation. The score operator for $F$ in the direction of $h$ is then given by

$$A_{\boldsymbol{\beta},F}h(\boldsymbol{X},\boldsymbol{Y}) = h(\boldsymbol{Y}) - B_{\boldsymbol{\beta},F}h(\boldsymbol{X}) - C_{\boldsymbol{\beta},F}h(\boldsymbol{X})\Sigma_{\boldsymbol{Y}}^{-\frac{1}{2}}(\boldsymbol{X};\boldsymbol{\beta},F)\left(\boldsymbol{Y}-\boldsymbol{\mu}\left(\boldsymbol{X}^T\boldsymbol{\beta}\right)\right)\,. \tag{2.41}$$

This is the vector analog of the score operator for $F$ given in Huang (2014). Furthermore, the expectation of the $A_{\boldsymbol{\beta},F}h$ for any $(\boldsymbol{\beta}, F)$ conditional on $\boldsymbol{X}$ is given by

$$\mathbb{E}_{\boldsymbol{\beta},F}\left[A_{\boldsymbol{\beta},F}h(\boldsymbol{X},\boldsymbol{Y})|\boldsymbol{X}\right] = \mathbb{E}_{\boldsymbol{\beta},F}\left[h(\boldsymbol{Y})\ |\boldsymbol{X}\right] - B_{\boldsymbol{\beta},F}h(\boldsymbol{X}) - C_{\boldsymbol{\beta},F}h(\boldsymbol{X})\Sigma_Y^{-\frac{1}{2}}\left(\mathbb{E}_{\beta,F}[\boldsymbol{Y}|\boldsymbol{X}] - \boldsymbol{\mu}\right) = 0\ .$$

Thus, for any $(\boldsymbol{\beta}, F)$, the score function for $F$ is unbiased,

$$\mathbb{E}_{\boldsymbol{\beta},F}\left[A_{\boldsymbol{\beta},F}h(\boldsymbol{X},\boldsymbol{Y})\right] = \mathbb{E}^{\mathcal{X}}\left[\mathbb{E}_{\boldsymbol{\beta},F}\left(A_{\boldsymbol{\beta},F}h(\boldsymbol{X},\boldsymbol{Y})|\boldsymbol{X}\right)\right] = 0\ . \tag{2.42}$$

## 2.4 Maximum Empirical Likelihood Estimation

Similar to Huang (2014), the mean model parameters $\boldsymbol{\beta}$ and reference distribution $F$ can be jointly estimated using maximum empirical likelihood estimation (MELE) by constructing an empirical likelihood (Owen 2001). This is done by replacing the densities $dF$ in the semiparametric log-likelihood with non-negative point probability masses $\boldsymbol{p} = \{p_i \geq 0 : \sum_{i=1}^n p_i = 1\}$ across the observed support $\{\boldsymbol{Y}_i \in \mathbb{R}^K | i = 1, 2, \ldots, n\}$. These probability masses $\boldsymbol{p}$ is known as Vardi's estimator or histogram estimators in nonparametric estimation of biased sampling or density ratio models (Gill et al. 1988). As a result, the empirical log-likelihood function is

$$\ell(\boldsymbol{\beta}, F|\boldsymbol{X},\boldsymbol{Y}) = \sum_{i=1}^n \left\{\log(p_i) + b(\boldsymbol{X}_i,\boldsymbol{\beta},F) + \boldsymbol{\theta}(\boldsymbol{X}_i,\boldsymbol{\beta},F)^T\boldsymbol{Y}_i\right\}\ . \tag{2.43}$$

The empirical log-likelihood is subject to the empirical analogous of the mean and normalisation constraints (2.4) and (2.7) for each $(b_i, \boldsymbol{\theta}_i) \in \mathbb{R} \times \mathbb{R}^K$ for $i = 1, 2, \ldots, n$. The normalisation constraint of each tilted nonparametric distribution becomes

$$1 = \sum_{j=1}^n p_j \exp\left(b_i + \boldsymbol{\theta}_i^T\boldsymbol{Y}_j\right), \qquad i = 1, \ldots, n \tag{2.44}$$

and the mean constraint to control the amount of tilt given to each observation is given by

$$\mu_{(k)}\left(\boldsymbol{X}_{i(k)}^T\boldsymbol{\beta}_{(k')}\right) = \sum_{j=1}^n p_j Y_{j(k)} \exp\left(b_i + \boldsymbol{\theta}_i^T\boldsymbol{Y}_j\right). \quad i = 1, \ldots, n,\ k = 1, \ldots, K \tag{2.45}$$

For equation (2.45) to be solvable, $\boldsymbol{\beta}$ must be such that the mean vector $\boldsymbol{\mu}\left(\boldsymbol{X}_i^T\boldsymbol{\beta}\right)$ lies in the convex hull of the observed support for $\boldsymbol{Y}$. For $\boldsymbol{\beta}$ which violates this condition, the convention in empirical likelihood literature is to set the empirical log-likelihood to $-\infty$ (Owen 2001).

**Definition 2.1 (Convex Hull)** *The series of points $\boldsymbol{Y}_i \in \mathbb{R}^K$ for $i = 1, 2, \ldots, n$ form the convex hull $\xi$ where*

$$\xi = \left\{\sum_{i=1}^n p_i\boldsymbol{Y}_i : p_i \in \mathbb{R}_+\ \forall i \text{ and } \sum_{i=1}^n p_i = 1\right\}\ .$$

We can also add identifiability constraints on the probability masses $\boldsymbol{p}$ by enforcing the mean constraint

$$\boldsymbol{\mu}_0 = \sum_{j=1}^n p_j\boldsymbol{Y}_j \tag{2.46}$$

for some $\boldsymbol{\mu}_0$ in the interior of the convex hull of the observed support for $\boldsymbol{Y}$. Again, note that the likelihood is invariant to the particular choice of $\boldsymbol{\mu}_0$, although Dennis (2021) discusses that enforcing a specific $\boldsymbol{\mu}_0$ is useful computationally by increasing numerical stability.

To estimate the mean model parameter $\boldsymbol{\beta}$, the profile empirical log-likelihood is constructed below which is defined to be the supremum of the empirical log-likelihood over all distribution on the observed support with mean $\boldsymbol{\mu}_0$.

$$pl(\boldsymbol{\beta}) = \sup_{\boldsymbol{p}} l(\boldsymbol{\beta}, \boldsymbol{p}) \ . \tag{2.47}$$

Our estimator for $\boldsymbol{\beta}$ is defined as the maximiser of the profile empirical log-likelihood

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} pl(\boldsymbol{\beta}) \ . \tag{2.48}$$

If $\hat{\boldsymbol{\beta}}$ is such that the mean vector $\boldsymbol{\mu}\left(\boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}\right)$ strictly lies in the interior of the convex hull of $\boldsymbol{Y}$, then the corresponding maximum empirical likelihood estimate of $\boldsymbol{p}$, $\hat{\boldsymbol{p}}$ exists and is unique (Theorem 1' Vardi 1985). Furthermore, $\hat{\boldsymbol{\beta}}$ corresponds with the $\boldsymbol{\beta}$ component of the joint maximiser $\left(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{p}}\right)$ of the empirical log-likelihood. We can also construct an empirical estimate for the reference distribution $F_{\boldsymbol{\mu}_0}$ given by

$$\hat{F}_{\boldsymbol{\mu}_0}(\boldsymbol{y}) = \sum_{j=1}^{n} \hat{p}_j \mathbb{1}(\boldsymbol{Y}_j \leq \boldsymbol{y}) \ . \tag{2.49}$$

As a result, we call $\left(\hat{\boldsymbol{\beta}}, \hat{F}_{\boldsymbol{\mu}_0}\right)$ the maximum empirical likelihood estimator (MELE) of $(\boldsymbol{\beta}, F_{\boldsymbol{\mu}_0})$. Often we omit the identifiability constraints and denote the MELE as $(\hat{\boldsymbol{\beta}}, \hat{F})$.

### 2.4.1 The MELE as the Solution to the Score Equations

We often express the MELE as the solution to the set of score equations derived in Section 2.3.1 and 2.3.2. Let $\mathbb{P}_n$ denote the empirical expectation operator and following notation given in Van Der Vaart and Wellner (1996) and Huang (2014), let $\Psi_n = (\Psi_{1n}, \Psi_{2n})^T$ be an element of $\mathbb{R}^q \times \ell^\infty(\mathcal{H}_L)$ where

$$\Psi_{1n}(\boldsymbol{\beta}, F) = \mathbb{P}_n S_{\boldsymbol{\beta},F} \tag{2.50}$$

$$\Psi_{2n}(\boldsymbol{\beta}, F)h = \mathbb{P}_n A_{\boldsymbol{\beta},F} h \ . \tag{2.51}$$

Because the conditional expectation operator retains boundedness, the mapping $h \mapsto \Psi_{2n}(\boldsymbol{\beta}, F)h$ is indeed uniformly bounded on $\mathcal{H}_L$ provided the response space $\mathcal{Y}$ has sufficiently light tails. In condition A.2.1 we make a stronger assumption that the response space $\mathcal{Y}$ is compact. The sequence of MELEs $(\hat{\beta}, \hat{F}_{\mu_0})$ are then the zeros of the maps $\Psi_n$,

$$\Psi_n(\hat{\boldsymbol{\beta}}, \hat{F}_{\boldsymbol{\mu}_0}) = \mathbf{0} \ . \tag{2.52}$$

To see why, note that $\Psi_{1n}(\boldsymbol{\beta}, F)$ is the derivative of the log-likelihood and $\hat{\boldsymbol{\beta}}$ by definition it is a zero of the score equation. If we consider $\Psi_{2n}h(\boldsymbol{\beta}, F)$, suppose a multivariate histogram with probability masses, and some cut-off point $\boldsymbol{r}$. Considering our class of functions $h$ which are left indicator functions, suppose we add a small mass $t$ to all probability masses to the left of $\boldsymbol{r}$ and re-normalize. If we optimize with respect to $t$ and we find that $t \neq 0$ is the optima, then the original histogram estimator was not the most optimal estimator.

Furthermore, let us define a centering function $\Psi = (\Psi_1, \Psi_2)^T$ which is a map from $\mathbb{R}^q \times \mathcal{H}_L$ to $\mathbb{R}^q \times \ell^\infty(\mathcal{H}_L)$ where

$$\Psi_1(\boldsymbol{\beta}, F) = P^* S_{\boldsymbol{\beta},F} \tag{2.53}$$

$$\Psi_2(\boldsymbol{\beta}, F)h = P^* A_{\boldsymbol{\beta},F} h \ . \tag{2.54}$$

Where $P^* = \mathbb{E}_{\boldsymbol{\beta}^*, F^*}[\cdot]$ denotes the expectation over $\boldsymbol{X}, \boldsymbol{Y}$ under the true parameter value $(\boldsymbol{\beta}^*, F^*)$. By the law of iterated expectations, we have that $P^* = \mathbb{E}^{\mathcal{X}}[\mathbb{E}_{\boldsymbol{\beta}^*, F^*}(\cdot|\boldsymbol{X})]$. As the conditional expectations of the score operators were found to be 0 for any $(\boldsymbol{\beta}, F)$, we also have

$$\Psi(\boldsymbol{\beta}^*, F^*) = \mathbf{0}\,. \tag{2.55}$$

## 2.5 Asymptotic Results

The asymptotic results of the MEL estimate $(\hat{\boldsymbol{\beta}}, \hat{F})$ are generalizations of the asymptotic results given in Huang (2014), derived under the regularity conditions in A.2. Similar to Huang (2014), we consider the weak topology on the infinite-dimensional parameter space $\mathcal{F}_{\boldsymbol{\mu}}$ with the distance function $||F_1 - F_2||_{\mathcal{H}_L} = \sup_{h \in \mathcal{H}_L} \int h(\mathrm{d}F_1 - \mathrm{d}F_2)$ where $\mathcal{H}_L := \{\mathbb{1}(\boldsymbol{y} \leq \boldsymbol{r}) : \boldsymbol{r} \in \mathcal{Y} \subset \mathbb{R}^K\}$ is the set of all left indicator functions on $\mathcal{Y}$. For the joint parameter space $\mathbb{R}^q \times \mathcal{F}_{\boldsymbol{\mu}}$, $||\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2|| + ||F_1 - F_2||_{\mathcal{H}_L}$ is equivalently written as $||(\boldsymbol{\beta}_1, F_1) - (\boldsymbol{\beta}_2, F_2)||$. The technical details of the results are provided in Chapter 5.

The first result is regarding the consistency of both $\hat{\boldsymbol{\beta}}$ and $\hat{F}$, assuming that $F^*(\boldsymbol{y}|\boldsymbol{x})$ lies within the multivariate exponential families.

**Lemma 2.2 (Consistency of MELE)** *As $n \to \infty$, $\hat{\boldsymbol{\beta}} \to \boldsymbol{\beta}^*$ and $\hat{F} \to F^*$ in probability, relative to the weak topology defined above.*

The joint asymptotic normality of $(\hat{\boldsymbol{\beta}}, \hat{F})$ can be established by adhering to Theorem 3.3.1 of Van Der Vaart and Wellner (1996) as done in Section 5.3. It should be noted that the asymptotic normality of $\hat{F}$ refers to the asymptotic normality of the stochastic process $\sqrt{n} \int h\, \mathrm{d}(\hat{F} - F^*)$ which is indexed by $h \in \mathcal{H}_L$.

**Proposition 2.1 (Joint Asymptotic Normality of the MELE)** *As $n \to \infty$*

$$\sqrt{n} \begin{pmatrix} \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \\ \hat{F} - F^* \end{pmatrix} \rightsquigarrow \begin{pmatrix} G_1 \\ G_2 \end{pmatrix} \tag{2.56}$$

*in $\mathbb{R}^q \times \ell^\infty(\mathcal{H}_L)$. Here $G_1$ is a mean zero multivariate normal random vector of dimension $q$ with covariance matrix $W_1$ given by*

$$W_1 = \left(\mathbb{E}^{\mathcal{X}}\left[D(\boldsymbol{X}; \boldsymbol{\beta}^*)\, \Sigma_{\boldsymbol{Y}}^{-1}(\boldsymbol{X}; \boldsymbol{\beta}^*, F^*)\, D(\boldsymbol{X}; \boldsymbol{\beta}^*)^T\right]\right)^{-1} \tag{2.57}$$

*$G_2$ is a mean zero Gaussian random process indexed by $h \in \mathcal{H}_L$ which is independent of $G_1$, with a covariance function $W_2(h_1, h_2) = W_F(\mathcal{D}h_1, \mathcal{D}h_2)$, where*

$$W_F(h_1, h_2) = \mathbb{E}^{\mathcal{X}}\left[Cov(h_1(\boldsymbol{Y}), h_2(\boldsymbol{Y})|\boldsymbol{X}) - C_{\boldsymbol{\beta}^*, F^*} h_1(\boldsymbol{X}) C_{\boldsymbol{\beta}^*, F^*} h_2(\boldsymbol{X})^T\right] \tag{2.58}$$

*and $\mathcal{D}h : \mathcal{H}_L \mapsto \ell^\infty(\mathcal{H}_L)$ is a continuous inverse operator.*

As stated by Rathouz and Gao (2009), an implication of $\boldsymbol{\beta}$ and $F$ being orthogonal is that estimate $\hat{\boldsymbol{\beta}}$ is generally consistent and asymptotically normal even when $F$ is misspecified or the exponential tilt model is invalid (Crowder 1986), although the standard errors will be biased. If $F^*(\boldsymbol{y}|\boldsymbol{x})$ lies outside of this class of distributions, we can find the asymptotic covariance matrix to be $W_{sandwich}$ given by

$$W_{sandwich} = \mathbf{A}(\boldsymbol{\beta}^*, F^*)^{-1} \mathbf{B}(\boldsymbol{\beta}^*, F^*) \mathbf{A}(\boldsymbol{\beta}^*, F^*)^{-1} \tag{2.59}$$

where

$$\mathbf{A}(\boldsymbol{\beta}^*, F^*) = -\mathbb{E}^{\mathcal{X}}\left[D(\boldsymbol{X}; \boldsymbol{\beta}^*)\, \Sigma_{\boldsymbol{Y}}^{-1}(\boldsymbol{X}; \boldsymbol{\beta}^*, F^*)\, D(\boldsymbol{X}; \boldsymbol{\beta}^*)^T\right] \tag{2.60}$$

and

$$\mathbf{B}(\boldsymbol{\beta}^*, F^*) = \mathbb{E}^{\mathcal{X}}\left[D(\boldsymbol{X};\boldsymbol{\beta}^*)\,\Sigma_{\boldsymbol{Y}}^{-1}(\boldsymbol{X};\boldsymbol{\beta}^*,F^*)\mathbb{E}_{\boldsymbol{\beta}^*}\left[(\boldsymbol{Y}-\boldsymbol{\mu})(\boldsymbol{Y}-\boldsymbol{\mu})^T|\boldsymbol{X}\right]\Sigma_{\boldsymbol{Y}}^{-1}(\boldsymbol{X};\boldsymbol{\beta}^*,F^*)\,D(\boldsymbol{X};\boldsymbol{\beta}^*)^T\right]. \quad (2.61)$$

The operator $\mathbb{E}_{\boldsymbol{\beta}^*}(\cdot)$ is with respect to the true distribution $F^*(\boldsymbol{y}|\boldsymbol{x})$ which is some multivariate distribution. We can see that under correct specification, we can recover $W_1$ from this expression as $-\mathbf{A}(\boldsymbol{\beta}^*, F^*) = \mathbf{B}(\boldsymbol{\beta}^*, F^*)$. Under correct specification, the asymptotic covariance matrix for $\boldsymbol{\beta}$ can be estimated by $\hat{\Sigma}_{\boldsymbol{\beta}}$ by plugging in our estimates $(\hat{\boldsymbol{\beta}}, \hat{F})$,

$$\hat{\Sigma}_{\boldsymbol{\beta}} := \left(\sum_{i=1}^{n} D(\boldsymbol{X}_i;\hat{\boldsymbol{\beta}})\left\{\sum_{j=1}^{n}(\boldsymbol{Y}_j-\hat{\boldsymbol{\mu}}_i)(\boldsymbol{Y}_j-\hat{\boldsymbol{\mu}}_i)^T\hat{p}_j\exp\{\hat{b}_i+\hat{\boldsymbol{\theta}}_i^T\boldsymbol{Y}_j\}\right\}^{-1} D(\boldsymbol{X}_i;\hat{\boldsymbol{\beta}})^T\right)^{-1}. \quad (2.62)$$

Furthermore, the asymptotically valid sandwich covariance estimator for $\boldsymbol{\beta}$ can be estimated by

$$\hat{\Sigma}_{sandwich} = \hat{\Sigma}_{\boldsymbol{\beta}}^{-1}\left(\textstyle\sum_{i=1}^{n} D(\boldsymbol{X}_i;\hat{\boldsymbol{\beta}})\hat{\Sigma}_Y^{-1}(\boldsymbol{X})\left\{\frac{1}{n}\sum_{j=1}^{n}(\boldsymbol{Y}_j-\hat{\mu}_j)(\boldsymbol{Y}_j-\hat{\mu}_j)^T\right\}\hat{\Sigma}_Y^{-1}(\boldsymbol{X})D(\boldsymbol{X}_i;\hat{\boldsymbol{\beta}})^T\right)\hat{\Sigma}_{\boldsymbol{\beta}}^{-1}. \quad (2.63)$$

It should be noted that when considering the asymptotic theory under the case of misspecification, although the sandwich estimate $W_{sandwich}$ can be used for the $\boldsymbol{\beta}$ component, it would induce a different covariance matrix for the $F$ component. The exploration of this was beyond the scope of the thesis, however, as we are generally only interested in inferences on $\boldsymbol{\beta}$, the sandwich estimator can still be used as an asymptotically valid covariance matrix in the case of misspecification.

Proposition 2.1 suggests to use Wald tests for inference on $\boldsymbol{\beta}$, testing for the null hypothesis $H_0 : \beta_{(k')j} = 0$ for any $k' = 1, 2, \ldots, K'$ and $j = 1, 2, \ldots q_k$. This was done in Huang and Rathouz (2012) where it's assumed under $H_0$ that

$$\frac{\hat{\beta}_{j(k')}}{\sqrt{\hat{\Sigma}_{\hat{\beta}_{(k')j}}}} \sim t_{n-q_k}, \quad (2.64)$$

where $q_k$ is the number of parameters in the $k'$-th marginal model. However, the asymptotic covariance of $\hat{\boldsymbol{\beta}}$ can be hard to numerically estimate accurately with numerical inverses and thus can limit the use of Wald tests. An alternative suggested by Huang and Rathouz (2012) and Huang (2014) is a likelihood-based inference on the $\boldsymbol{\beta}$ parameters using the profile empirical likelihood function $pl(\boldsymbol{\beta})$. Using the profile empirical likelihood allows for hypothesis testing and the construction of non-symmetric joint confidence intervals without needing to explicitly compute the covariance of $\hat{\boldsymbol{\beta}}$. Inference of this kind is done by profiling out the $F$ parameter from the empirical likelihood function for a fixed $\boldsymbol{\beta}$ so that the profile empirical log-likelihood ratio statistic can be used to provide inference on $\boldsymbol{\beta}$. This works because the profile empirical log-likelihood behaves asymptotically like a true log-likelihood, which is established in the following proposition based on results from Murphy and Van Der Vaart (2000).

**Proposition 2.2 (Point Hypotheses)** *Under the null hypothesis $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}^* \in \mathbb{R}^q$, the profile empirical log-likelihood ratio statistic.*

$$-2\left\{pl(\boldsymbol{\beta}^*) - pl(\hat{\boldsymbol{\beta}})\right\} \quad (2.65)$$

*is asymptotically $\chi^2_q$ distributed as $n \to \infty$.*

Following Proposition 2.2, we can generalise to considering composite hypothesis on $\boldsymbol{\beta}$ where $\boldsymbol{\beta}$ lies in the $r$ dimensional subspace of $\mathbb{R}^q, 1 \leq r \leq q$ and the null parameter is defined by

$$B_0 = \{\boldsymbol{\beta} \in \mathbb{R}^q : \boldsymbol{M}\boldsymbol{\beta} = \gamma\} \quad (2.66)$$

where $\boldsymbol{M}$ is a rank $r$ matrix and $\gamma$ is a vector. Denoting $\hat{\boldsymbol{\beta}}$ as the unconstrained MLE over $\mathbb{R}^q$ and $\hat{B}_0 = \arg\max_{\boldsymbol{\beta}\in\boldsymbol{\beta}_0} pl(\boldsymbol{\beta})$ as the MLE over $B_0$, we have the following corollary.

**Corollary 2.2 (Composite Hypotheses)** *Under the null hypothesis* $H_0 : \boldsymbol{\beta} \in B_0 = \{\boldsymbol{\beta} \in \mathbb{R}^q : \boldsymbol{M}\boldsymbol{\beta}^* = \gamma\}$*, the profile empirical log-likelihood ratio statistic*

$$-2\left\{pl(\hat{\boldsymbol{\beta}}_0) - pl(\hat{\boldsymbol{\beta}})\right\} \tag{2.67}$$

*is asymptotically* $\chi^2_r$ *distributed as* $n \to \infty$*.*

For finite samples, the empirical likelihood ratio can be compared to a $rF_{r,n-q}$ distribution for joint inference, as $rF_{r,n-q} = \chi^2_r + o_P(1)$ in distribution. As a result, hypothesis testing can be performed using an Empirical Likelihood Ratio Test (ELRT), and confidence regions for $\boldsymbol{\beta}$ can be constructed by inverting the empirical likelihood ratio test. In Chapter 4 Section 4.1, a simulation study is explored to examine the coverage rates of both the Wald test and ELRT.

Finally, we should note that if we have $n$ observations of $K$ independent response components, we could argue for there instead being $nK - q$ degrees of freedom. However, in cases where we share covariates and coefficients across the components, then it's appropriate to use $n - q$ degrees of freedom. The choice of degrees of freedom across VGLM frameworks is dependent on the structure of the problem, but in the current implementation for hypothesis testing, we are conservative in the choice of degrees of freedom.

# Chapter 3

# Fitting the Model Computationally

## 3.1 Introduction

This chapter provides details on the computational implementation of the estimation of the mean model parameters $\boldsymbol{\beta}$ and reference distribution parameters $\boldsymbol{p}$. The code for VSPGLM was originally implemented by Dennis (2021) in `MATLAB`, which is publicly available on GitHub [1]. The implementation and documentation required amendments to match the derivations in Chapter 2, which will be explored in this chapter. The joint estimation uses the function `fmincon` to solve the nonlinear constrained optimization problem with the mean and normalization constraints (2.4) and (2.7) respectively. All updated code, documentation and example usages for VSPGLM can be found on GitHub [2].

## 3.2 Model Interface

The VSPGLM is fitted using the wrapper function `fit_vspglm`, which has the following syntax

```
% fit_vspglm function definition
vspglmmodel = fit_vspglm(formula, tbl, links)
```

The main argument `formula` is a string object which follows similar to the syntax that is used in **R** functions such as **lm, glm** and **gldrm**. However, the implementation does have unique syntax to handle a variety of cases which we will detail below. The `tbl` argument is a `MATLAB` table with all response components and covariates of the model, and `links` is a cell array of strings that specify the link functions used for each component. The function `fit_vspglm` can currently handle the identity log, logit and inverse link functions given below respectively.

$$\mu,\ \log(\mu),\ \log\left(\frac{\mu}{1-\mu}\right),\ \frac{1}{\mu}\,. \tag{3.1}$$

The output of the function is a structure that contains information about the fitted model. This includes the log-likelihood, estimated parameters $(\boldsymbol{\beta}, \boldsymbol{p})$, parameter inference on $\boldsymbol{\beta}$, empirical covariance matrix of $\boldsymbol{Y}$ and $\boldsymbol{\beta}$ and the sandwich covariance matrix for $\boldsymbol{\beta}$.

Now let us detail how the syntax for the `formulas` argument changes to handle various problems. For the case where each component has separate marginal mean models,

$$g_1(\mu_{(1)}) = \boldsymbol{X}_{(1)}^T\boldsymbol{\beta}_{(1)},\ g_2(\mu_{(2)}) = \boldsymbol{X}_{(2)}^T\boldsymbol{\beta}_{(2)},\ \ldots,\ g_K(\mu_{(K)}) = \boldsymbol{X}_{(K)}^T\boldsymbol{\beta}_{(K)}\ ,$$

[1] https://github.com/gden173/vspglm
[2] https://github.com/Kyle-Macaskill/Honours-VSPGLM

where $g_k(\cdot)$ is the link function for component $k = 1, 2, \ldots, K$, the syntax is

```
% Separate marginal mean models
model = fit_vspglm(["y_1 ~ x_1", "y_2 ~ x_2", ..., "y_k ~ x_k"], tbl, links)
```

In the model above, each component has its own set of coefficients, however, coefficients can be shared across components as detailed in Chapter 2. These constraints on the coefficients are imposed through the `formula` argument. In the case where all components share coefficients (symmetry), but have different a different covariates

$$g_1(\mu_{(1)}) = \boldsymbol{X}_{(1)}^T\boldsymbol{\beta},\ g_2(\mu_{(2)}) = \boldsymbol{X}_{(2)}^T\boldsymbol{\beta},\ \ldots,\ g_K(\mu_{(K)}) = \boldsymbol{X}_{(K)}^T\boldsymbol{\beta}\ ,$$

the syntax is

```
% Shared marginal mean models, different covariates
model = fit_vspglm(["(y_1, y_2, ..., y_k) ~ ((x_1 & x_2 & ... & x_k))"], tbl, links)
```

Importantly, it's assumed that components share coefficients when they are a part of the same mean model and as a result, they must have the same link function $g$.

In the case where the coefficients are shared between all components but each component has the same set of covariates $\boldsymbol{X}$,

$$g_1(\mu_{(1)}) = \boldsymbol{X}^T\boldsymbol{\beta},\ g_2(\mu_{(2)}) = \boldsymbol{X}^T\boldsymbol{\beta},\ \ldots,\ g_K(\mu_{(K)}) = \boldsymbol{X}^T\boldsymbol{\beta}\ ,$$

the syntax is

```
% Shared marginal mean models, same covariates
model = fit_vspglm(["(y_1, y_2, ..., y_k) ~ x"], tbl, links)
```

Furthermore, we can consider hybrid cases where some components share covariates and others do not. Suppose the case where each component has its own set of covariates and the first two components share coefficients,

$$g_1(\mu_{(1)}) = \boldsymbol{X}_{(1)}^T\boldsymbol{\beta}_{(1)},\ g_2(\mu_{(2)}) = \boldsymbol{X}_{(2)}^T\boldsymbol{\beta}_{(1)},\ g_3(\mu_{(3)}) = \boldsymbol{X}_{(3)}^T\boldsymbol{\beta}_{(2)},,\ \ldots,\ g_K(\mu_{(K)}) = \boldsymbol{X}_{(K)}^T\boldsymbol{\beta}_{(K)}\ .$$

The syntax for the model above is given by

```
% Some mean models shared coefficients and others not, with different covariates
model = fit_vspglm(["(y_1, y_2) ~ ((x_1 & x_2))", "y_3 ~ x_3", ..., "y_k ~ x_k"], tbl,
    links)
```

We can also consider the case where components share coefficients but the dimensionality of the covariates is different. For this case, suppose the following model

$$\begin{aligned} g_1(\mu_{(1)}) &= \beta_{(1)1} + X_1\beta_{(1)2} + X_2\beta_{(1)3} + X_3\beta_{(1)4} + X_4\beta_{(1)5} \\ g_2(\mu_{(2)}) &= \beta_{(1)1} + X_1\beta_{(1)2} + X_2\beta_{(1)3} + X_5\beta_{(1)5}\ , \end{aligned}$$

where the second component is missing the term associated with $\boldsymbol{\beta}_{(1)4}$. To handle this case the syntax would be

```
% Some mean models shared coefficients and others not, a mix of shared and not shared
    covariates
model = fit_vspglm(["(y_1, y_2) ~ (x_1, x_2, (x_3&0), (x_4 & x_5))" ], tbl, links)
```

Above, the 0 represents the name of a covariate which only takes the value 0, which needs to be added to the argument `tbl`. Due to the initialisation of $\boldsymbol{\beta}$, this enforces the associated coefficient to be 0.

Point and compound hypothesis testing using the Empirical Likelihood Ratio Test (ELRT) can be done by fitting the model without the respective covariates when testing if the parameter is 0 or not. The function `fit_vspglm_constraint` fits the same VSPGLM as `fit_vspglm` but with the additional functionality of allowing for the specific parameters in $\boldsymbol{\beta}$ to be a fixed value. This allows for the likelihood under more general null hypotheses to be found, as well as allows for coverage rates using ELRT to be found for simulations.

```
% fit_vspglm_constraint function definition
vspglmmodel = fit_vspglm_constraint(formula, tbl, links, index, fixed)
```

The `index` argument is an array of the indices of $\boldsymbol{\beta}$ which will be fixed, and the `fixed` argument are the corresponding fixed values for the given $\beta$. As an example, suppose we wish to fit the following model

$$\begin{aligned} g_1(\mu_{(1)}) &= \beta_{(1)1} + X_{(1)2}\beta_{(1)2} \\ g_2(\mu_{(2)}) &= \beta_{(2)1} + X_{(2)2}\beta_{(2)2} + X_{(2)3}\beta_{(2)3} \ , \end{aligned}$$

where $\beta_{(1)1} = 1.5, \beta_{(2)2} = 0$. As there are 5 parameters in total being optimized, we can impose the constraints as follows

```
% Fitting model with certain parameter values fixed
model = fit_vspglm_constraint(["y_1 ~ x_11", "y_2 ~ x_21, x_22"], tbl, links, [2,5],
    [1.5, 0])
```

Once the model is fitted, we are able to predict the fitted means for a new set of covariates using the implemented predict function `predict_vspglm`. With the estimated means, the function backsolves for the tilt parameters $\boldsymbol{\theta}$ via the mean constraint to predict the joint distribution of each new observation, along with a covariance matrix. The syntax for doing so is

```
% Predicting fitted values and distribution for new observations
predict_output = predict_vspglm(model, X_new, Y_old);
```

where `model` is a fitted VSPGLM, `X_new` is a table of covariates for the new observations, and `Y_old` is a matrix of the original responses which is used for the calculating the estimated covariance matrix. In the current implementation, the function only works for the case of having separate mean models for each component, and the table `X_new` needs to have its columns ordered according to the covariates used for each mean model. However, this will be updated in future iterations of the code.

Finally, as VSPGLM estimates the reference density $\hat{\boldsymbol{p}}$ using the estimated tilt parameters $\hat{\boldsymbol{\theta}}$, the estimated joint density for each observation $i$ can be expressed as $\hat{\boldsymbol{p}}\exp\{b_i + \boldsymbol{\theta}_i^T\boldsymbol{Y}\}$. In the case where we have two components, we can visualise this estimated joint density for each observation in the data using the function `vspglmJointPlot`.

```
% Function to plot the joint density for all observations when k = 2
vspglmJointPlot(model, formulas,Y, stem_flag, play)
```

Here `model` is the fitted VSPGLM structure, `formulas` is the formula used to fit the VSPGLM, `Y` is a matrix of the response variables, `stem_flag` indicates whether a two-dimensional scatter plot or three-dimensional stem plot is shown and `play` for the option to play through the different plots one at a time. As tilting parameters $\hat{\boldsymbol{\theta}}$ satisfy the mean constraint, the plotted joint densities will have fitted means which align with the VSPGLM fitted model. However, given a new vector of estimated means $\boldsymbol{\mu}$ which lie outside of the fitted model and inside the convex hull of the $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$, the model is still able to provide an estimated joint distribution using the following syntax

```
% Function to plot the joint density for an arbitrary vector of estimated means
vspglmJointPredict(model,Y, mu, stem_flag)
```

where `mu` is a vector of new estimated means and the other parameters are the same as in `vspglmJointPlot`. Further documentation can be found on Github[3] and examples usages can be found in Chapter 4.

## 3.3 Optimization

There are many different optional arguments we can pass into `fmincon` to increase the computational speed and numerical stability of the estimation procedure. These arguments include gradients for the objective function, mean and normalisation constraints, initial parameter values and choices of optimization algorithm. Below we will detail the choices made in the implementation and compare them with the implementation by Dennis (2021).

To perform the non-linear constrained optimization, the optimization algorithm used in `fmincon` is either an interior-point algorithm or a sequential quadratic programming (SQP) algorithm. SQP was found to be the quickest algorithm for most problems, having the best speed and memory performance on small to medium-scale problems. However, for larger-scale problems, it can be slow as it performs linear algebra on full matrices which may be dense. `MATLAB`'s interior-point algorithm is slower for most problems, but it can handle larger scale problems better as it aims to store matrices as sparse matrices to perform linear algebra. However, the interior-point algorithm can have small inaccuracies in the log-likelihood as a result of iterating away from constraint boundaries during optimization. As this performance is relative to a particular problem, we interchange between two algorithms where appropriate. The optimization procedure currently solves for all parameters simultaneously for all $n$ observations, of which there are $q + n(K + 2)$. These are,

- $n$ normalising constants $\hat{b}_i,\ \ i = 1, \ldots, n$
- $nK$ tilt parameters $\hat{\boldsymbol{\theta}}_i,\ \ i = 1, \ldots, n$
- $\sum_{k=1}^{K} q_k = q$ regression coefficients $\hat{\beta}_{(k')j},\ \ j = 1, 2, \ldots, q_k,\ \ k' = 1, 2, \ldots, K'$,
- $n$ reference distribution parameters $\hat{\boldsymbol{p}}$.

Note that applying constraints on the mean models in the current implementation is done by passing a linear equality constraint matrix into `fmincon`, which is similar to the constraint matrices in Yee (2015). The objective function of the optimization procedure is the negative empirical log-likelihood

$$\ell(\boldsymbol{\beta}, \tilde{\boldsymbol{p}}, \boldsymbol{\theta}, \boldsymbol{b}) = -\sum_{i=1}^{n} \tilde{p}_i + b_i + \boldsymbol{\theta}_i^T \boldsymbol{Y}_i, \tag{3.2}$$

where $\tilde{p}_i = \log(p_i)$ to ensure that $p_i \geq 0, i = 1, \ldots n$ without needing any additional constraints. The minimization of (3.2) is subject to the normalization constraints and mean constraints for each observation $i = 1, 2, \ldots, n$ as given in (3.3) and (3.4) respectively.

$$0 = 1 - \sum_{j=1}^{n} \exp\left(\tilde{p}_j + b_i + \boldsymbol{\theta}_i^T \boldsymbol{Y}_j\right), \qquad i = 1, \ldots, n \tag{3.3}$$

$$0 = \boldsymbol{\mu}\left(\boldsymbol{X}_i^T \boldsymbol{\beta}\right) - \sum_{j=1}^{n} \boldsymbol{Y}_j \exp\left(\tilde{p}_j + b_i + \boldsymbol{\theta}_i^T \boldsymbol{Y}_j\right) \quad i = 1, \ldots, n. \tag{3.4}$$

[3] https://github.com/Kyle-Macaskill/Honours-VSPGLM

As a result of (3.3), we do not require normalization constraints on $\boldsymbol{p}$, ie.

$$\sum_{i=1}^{n} p_i = 1. \tag{3.5}$$

To improve the performance of the optimization procedure, gradients of the objective function and constraints can be passed into `fmincon`. The gradient of the objective function $\ell$ is a $q + n(K+2) \times 1$ vector of the form

$$\nabla \ell(\boldsymbol{\beta}, \tilde{\boldsymbol{p}}, \boldsymbol{\theta}, \boldsymbol{b}) = \begin{pmatrix} \frac{\partial l}{\partial \boldsymbol{\beta}} \\ \frac{\partial l}{\partial \tilde{\boldsymbol{p}}} \\ \frac{\partial l}{\partial \boldsymbol{b}} \\ \frac{\partial l}{\partial \boldsymbol{\theta}} \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ -\mathbf{1} \\ -\mathbf{1} \\ -\boldsymbol{Y}_i,\ i = 1, \ldots, n \end{pmatrix} \tag{3.6}$$

The normalization constraint (3.3) in implementation is a $n \times 1$ vector of constraints which we will denote as $\boldsymbol{bc}$,

$$\boldsymbol{bc} = \begin{bmatrix} bc_1 \\ \vdots \\ bc_n \end{bmatrix} = \mathbf{0} \in \mathbb{R}^{n \times 1}$$
$$bc_i = 1 - \sum_{j=1}^{n} \exp\left(\tilde{p}_j + b_i + \boldsymbol{\theta}_i^T \boldsymbol{Y}_j\right), \quad i = 1, 2, \ldots, n$$

Thus, the gradient is a $q + n(K+2) \times n$ matrix of the form

$$\nabla \boldsymbol{bc} = \begin{bmatrix} \frac{\partial bc_1}{\partial \boldsymbol{\beta}} & \frac{\partial bc_2}{\partial \boldsymbol{\beta}} & \cdots & \frac{\partial bc_n}{\partial \boldsymbol{\beta}} \\ \frac{\partial bc_1}{\partial \tilde{p}_1} & \frac{\partial bc_2}{\partial \tilde{p}_1} & \cdots & \frac{\partial bc_n}{\partial \tilde{p}_1} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial bc_1}{\partial \tilde{p}_n} & \frac{\partial bc_2}{\partial \tilde{p}_n} & \cdots & \frac{\partial bc_n}{\partial \tilde{p}_n} \\ \frac{\partial bc_1}{\partial b_1} & \frac{\partial bc_2}{\partial b_1} & \cdots & \frac{\partial bc_n}{\partial b_1} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial bc_1}{\partial b_n} & \frac{\partial bc_2}{\partial b_n} & \cdots & \frac{\partial bc_n}{\partial b_n} \\ \frac{\partial bc_1}{\partial \boldsymbol{\theta}_1} & \frac{\partial bc_2}{\partial \boldsymbol{\theta}_1} & \cdots & \frac{\partial bc_n}{\partial \boldsymbol{\theta}_1} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial bc_1}{\partial \boldsymbol{\theta}_n} & \frac{\partial bc_2}{\partial \boldsymbol{\theta}_n} & \cdots & \frac{\partial bc_n}{\partial \boldsymbol{\theta}_n} \end{bmatrix} \in \mathbb{R}^{q+n(K+2) \times n},$$

where for $i = 1, 2, \ldots, n$, $j_2 = 1, 2, \ldots, n$

$$\frac{\partial bc_i}{\partial \boldsymbol{\beta}} = \mathbf{0} \in \mathbb{R}^q,$$
$$\frac{\partial bc_i}{\partial \tilde{p}_{j_2}} = -\exp\left(\tilde{p}_{j_2} + b_i + \boldsymbol{\theta}_i^T \boldsymbol{Y}_{j_2}\right),$$
$$\frac{\partial bc_i}{\partial b_{j_2}} = \begin{cases} -\sum_{j=1}^{n} \exp\left(\tilde{p}_j + b_i + \boldsymbol{\theta}_i^T \boldsymbol{Y}_j\right) & i = j_2 \\ 0 & i \neq j_2 \end{cases},$$
$$\frac{\partial bc_i}{\partial \boldsymbol{\theta}_{j_2}} = \begin{cases} -\sum_{j=1}^{n} \boldsymbol{Y}_j \exp\left(\tilde{p}_j + b_i + \boldsymbol{\theta}_i^T \boldsymbol{Y}_j\right) & i = j_2 \\ \mathbf{0} & i \neq j_2 \end{cases}.$$

Furthermore, the mean constraint (3.4) in implementation is a $nK \times 1$ vector of constraints which we will denote as $\boldsymbol{\mu c}$,

$$\boldsymbol{\mu c} = \begin{bmatrix} \mu c_1 \\ \vdots \\ \mu c_n \end{bmatrix} = \mathbf{0} \in \mathbb{R}^{nK \times 1}$$

$$\mu c_i = \boldsymbol{\mu}\left(\boldsymbol{X}_i^T \boldsymbol{\beta}\right) - \sum_{j=1}^{n} \boldsymbol{Y}_j \exp\left(\tilde{p}_j + b_i + \boldsymbol{\theta}_i^T \boldsymbol{Y}_j\right) \in \mathbb{R}^{K \times 1},\ i = 1, 2, \ldots, n$$

Thus, the gradient is a $q + nK + 2n \times nK$ matrix of the form

$$\nabla \boldsymbol{\mu c} = \begin{bmatrix} \frac{\partial \boldsymbol{\mu} c_1}{\partial \boldsymbol{\beta}} & \frac{\partial \boldsymbol{\mu} c_2}{\partial \boldsymbol{\beta}} & \cdots & \frac{\partial \boldsymbol{\mu} c_n}{\partial \boldsymbol{\beta}} \\ \frac{\partial \boldsymbol{\mu} c_1}{\partial \tilde{p}_1} & \frac{\partial \boldsymbol{\mu} c_2}{\partial \tilde{p}_1} & \cdots & \frac{\partial \boldsymbol{\mu} c_n}{\partial \tilde{p}_1} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \boldsymbol{\mu} c_1}{\partial \tilde{p}_n} & \frac{\partial \boldsymbol{\mu} c_2}{\partial \tilde{p}_n} & \cdots & \frac{\partial \boldsymbol{\mu} c_n}{\partial \tilde{p}_n} \\ \frac{\partial \boldsymbol{\mu} c_1}{\partial b_1} & \frac{\partial \boldsymbol{\mu} c_2}{\partial b_1} & \cdots & \frac{\partial \boldsymbol{\mu} c_n}{\partial b_1} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \boldsymbol{\mu} c_1}{\partial b_n} & \frac{\partial \boldsymbol{\mu} c_2}{\partial b_n} & \cdots & \frac{\partial \boldsymbol{\mu} c_n}{\partial b_n} \\ \frac{\partial \boldsymbol{\mu} c_1}{\partial \boldsymbol{\theta}_1} & \frac{\partial \boldsymbol{\mu} c_2}{\partial \boldsymbol{\theta}_1} & \cdots & \frac{\partial \boldsymbol{\mu} c_n}{\partial \boldsymbol{\theta}_1} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \boldsymbol{\mu} c_1}{\partial \boldsymbol{\theta}_n} & \frac{\partial \boldsymbol{\mu} c_2}{\partial \boldsymbol{\theta}_n} & \cdots & \frac{\partial \boldsymbol{\mu} c_n}{\partial \boldsymbol{\theta}_n} \end{bmatrix} \in \mathbb{R}^{q+nK+2n \times nK},$$

where for $i = 1, 2, \ldots, n,\ j_2 = 1, 2, \ldots, n$

$$\frac{\partial \mu c_i}{\partial \boldsymbol{\beta}} = \frac{\partial \boldsymbol{\mu}(\boldsymbol{X}_i^T \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = T(\boldsymbol{X}_i; \boldsymbol{\beta}) = \mathrm{Diag}(\boldsymbol{X}_i')\mathrm{Diag}(\boldsymbol{\mu}_i'),$$

$$\frac{\partial \mu c_i}{\partial \tilde{p}_{j_2}} = -\boldsymbol{Y}_{j_2}^T \exp\left(\tilde{p}_{j_2} + b_i + \boldsymbol{\theta}_i^T \boldsymbol{Y}_{j_2}\right),$$

$$\frac{\partial \mu c_i}{\partial b_{j_2}} = \begin{cases} -\sum_{j=1}^{n} \boldsymbol{Y}_j \exp\left(\tilde{p}_j + b_i + \boldsymbol{\theta}_i^T \boldsymbol{Y}_j\right) & i = j_2 \\ 0 & i \neq j_2 \end{cases}$$

$$\frac{\partial \mu c_i}{\partial \boldsymbol{\theta}_j} = \begin{cases} -\sum_{j=1}^{n} \boldsymbol{Y}_j \boldsymbol{Y}_j^T \exp\left(\tilde{p}_j + b_i + \boldsymbol{\theta}_i^T \boldsymbol{Y}_j\right) & i = j_2 \\ \mathbf{0} & i \neq j_2 \end{cases}$$

Note that in the expression above we stack components for a particular observation together first, then stack all the observations. However, we can also stack the matrix by observations and then components to obtain an equivalent expression, just ordered differently.

We must also ensure that the initial parameter values are such that the algorithm is feasible and so that the initial mean vector $\boldsymbol{\mu}^{(0)}$ lies in the interior of the observed support's convex hull. The initial point

$$\boldsymbol{z}^{(0)} = \left(\boldsymbol{\beta}^{(0)^T}, \tilde{\boldsymbol{p}}^{(0)^T}, \boldsymbol{b}^{(0)^T}, \boldsymbol{\theta}^{(0)^T}\right)^T$$

is used where

$$\tilde{\boldsymbol{p}}^{(0)^T} = -\boldsymbol{ln}(\boldsymbol{n})^T,\ \boldsymbol{b}^{(0)^T} = \mathbf{0}^T,\ \boldsymbol{\theta}^{(0)^T} = \mathbf{0}^T,\ \beta_{(k')1}^{(0)^T} = g_{(k)}(\bar{\boldsymbol{Y}}_k),\ \beta_{(k')j}^{(0)^T} = 0,$$

for all $k = 1, \ldots, K,\ k' = 1, \ldots, K',\ j = 2, \ldots, q_k$. Note that the intercept coefficient is set to a sample mean from all the responses in the $k'$-th mean model and $p_i^{(0)} = \frac{1}{n}$. Thus, these initial parameters ensure that $\boldsymbol{\mu}^{(0)}$ satisfies the convex hull condition as

$$\boldsymbol{\mu}^{(0)} = \sum_{i=1}^{n} p_i^{(0)} \boldsymbol{Y}_i = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{Y}_i \tag{3.7}$$

To increase the numerical stability of the solution, Dennis (2021) implemented a centering and re-scaling of each response component $Y_{(k)}$ onto the interval [-1,1] as was done by Wurm and Rathouz (2018) in the **R** package **gldrm**. This is done via the invertible transformation

$$\tilde{Y}_{(k)} = \left(Y_{(k)} - \frac{M_{(k)} + m_{(k)}}{2}\right) \cdot \frac{2}{M_{(k)} - m_{(k)}} \tag{3.8}$$

$$Y_{(k)} = \frac{M_{(k)} + m_{(k)}}{2} + \left(\frac{M_{(k)} - m_{(k)}}{2}\right) \tilde{Y}_{(k)} \tag{3.9}$$

where $M_{(k)} = \max_{1 \le i \le n} \boldsymbol{Y}_{(k)}$, $m_{(k)} = \min_{1 \le i \le n} \boldsymbol{Y}_{(k)}$. The inverse link function then becomes

$$\tilde{\mu}_{(k)} = \left(\mu_{(k)} - \frac{M_{(k)} + m_{(k)}}{2}\right) \cdot \frac{2}{M_{(k)} - m_{(k)}} \tag{3.10}$$

This is useful as the log-likelihood is invariant under this transformation, so the parameter estimates for $\boldsymbol{\beta}, \boldsymbol{p}$ are the same as using the original variable. For a single observation $i$, as seen in Dennis (2021), the empirical log-likelihood, normalization constraint and mean constraint respectively become

$$\ell_i(\boldsymbol{\beta}, \tilde{\boldsymbol{p}}, \boldsymbol{\theta}, \boldsymbol{b}) = \tilde{p}_i + b_i + \boldsymbol{\theta}_i^T \boldsymbol{Y}_i = \tilde{p}_i + \tilde{b}_i + \tilde{\boldsymbol{\theta}}_i^T \tilde{\boldsymbol{Y}}_i, \tag{3.11}$$

$$\tilde{b}_i = -\log\left\{\sum_{j=1}^{n} p_j \exp\left(\tilde{\boldsymbol{\theta}}_i^T \tilde{\boldsymbol{Y}}_j\right)\right\} \tag{3.12}$$

$$\tilde{\boldsymbol{\mu}}\left(\boldsymbol{X}_i^T \boldsymbol{\beta}\right) = \sum_{j=1}^{n} \tilde{\boldsymbol{Y}}_j \exp\left(\tilde{p}_j + \tilde{b}_i + \tilde{\boldsymbol{\theta}}_i^T \tilde{\boldsymbol{Y}}_j\right) \tag{3.13}$$

for $i = 1, \ldots, n$ where

$$\tilde{\boldsymbol{\theta}} = \left[\left(\frac{M_{(1)} - m_{(1)}}{2}\right)\theta_{(1)}, \ldots, \left(\frac{M_{(K)} - m_{(K)}}{2}\right)\theta_{(K)}\right]^T.$$

Note that this produces equivalent normalization and mean constraints, so transforming our variables doesn't change our optimization procedure detailed prior and yields the same results. Furthermore, as similarly seen in Wurm and Rathouz (2018), to ensure that $\exp\left(\tilde{b} + \tilde{\boldsymbol{\theta}}^T \tilde{\boldsymbol{Y}}\right)$ stays within `MATLAB`'s numerical tolerance, the following constraint is placed on $\tilde{\boldsymbol{\theta}}$,

$$-\frac{500}{K} \le \tilde{\theta}_{(k)i} \le \frac{500}{K}, \quad k = 1, \ldots, K. \tag{3.14}$$

With these numerical optimizations and documentation fixes, the code when fitting a univariate response yields are the same as **gldrm**, which is expected as it implements the univariate SPGLM proposed by Huang (2014). However, it is now flexible to handle datasets with multivariate responses with any number of dimensions. However, a downside to the current computational implementation is that it optimizes $Q + n(K+2)$ parameters subject to $n(K+1)$ constraints, scaling with the sample size $n$. As a result, it's computationally slower to fit VSPGLM compared to **gldrm** and the MDRM implemented by Marchese and Diao (2017) for data with a large number of observations and response components (eg. $n > 1000$ and $K > 8$). Note that although the number of variables increases with $n$, note that we are still only estimating a single $F$, where this scaling is a part of the estimation procedure. Furthermore, variables such as $\boldsymbol{\theta}$ and $b$ are not estimated but rather implicitly defined by the constraints, which are not overfitted due to the independence between the estimation of $\boldsymbol{\beta}$ and $F$. Exploration for speeding up the code was beyond the scope of the thesis, but suggestions for future directions on how to are provided in Chapter 6.

# Chapter 4

# Simulation Studies and Examples

This chapter will showcase the flexibility and versatility of the proposed VSPGLM by exploring several example applications from VGLM literature. For each example, we will compare parameter estimates and inference to existing frameworks where possible and provide additional analysis available as a result of the estimation of the reference distribution $F$. Before doing so, a simulation study in the case of correct specification from the multivariate exponential family will be explored to assess the asymptotic properties of the estimates $\hat{\boldsymbol{\beta}}$ outlined in Chapter 2, and the performance of both Wald tests and empirical likelihood ratio tests (ELRT).

The first two applications focus on fitting the VSPGLM to datasets with two response components, including a World Countries dataset with continuous responses and the Burn Injury dataset (Fan and Gijbels 1996) with response components of mixed data types. The chapter then explores applications of VSPGLM to species modelling for both count and presence/absence data with a large number of response components. This will include the Butterfly dataset (Oliver et al. 2006) where the counts of butterfly species at various locations in Boulder, Colorado which was previously modelled by Hui et al. (2013), and a dataset of plant species in the Hunua mountain ranges explored by Huang (2017). The final two applications will explore fitting the VSPGLM with constraints on the coefficients across the components. This includes a dataset from a Sorbinil Retinopathy Trail (Rosner et al. 2006) which has bivariate ordinal responses, and a longitudinal study that looks at the impact of dietary intervention on the development of Kenyan school children (Neumann et al. 2003) which are typically fitted using GEEs.

## 4.1 Simulation Studies

To verify the consistency of $\hat{\boldsymbol{\beta}}$, the asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$ and proposed inference we conduct a series of simulation studies. We examine the VSPGLM in a correctly specified situation by considering analysis on multivariate normal data with correlated components, comparing with results obtained by the MDRM in Marchese and Diao (2017) and Dennis (2021). This is the only multivariate exponential family we can parametrically simulate from, but simulations of the VSPGLM for non-standard multivariate distributions using Generalized Linear Mixed Models can be found in Dennis (2021). In these situations where data is drawn from distributions that are not a part of the multivariate exponential family, the estimates are found to be unbiased which is in line with the result from Hiejima (1997). The unadjusted standard errors and coverage rates reported by Dennis (2021) were found to be relatively robust, however manual corrections aiming to reduce numerical errors when taking inverses were performed. Although not explored in this section, the standard errors in

misspecified simulations can be corrected with the proposed sandwich estimator which will produce asymptotically valid results. In each simulation the bias, average standard errors ($\hat{\sigma}$) from the empirical covariance matrix, and standard deviation of the $\boldsymbol{\beta}$ estimates are reported. Furthermore, 90%, 95% and 99% coverage rates for wald-confidence intervals against the $t_{n-q}$ distribution are given for each parameter in $\hat{\boldsymbol{\beta}}$.

### 4.1.1 Unconstrained Multivariate Normal Simulation

Let us consider a model with two responses generated from a bivariate normal distribution with correlated marginal distributions with coefficients not being shared between the components. The two response components will have covariates $\boldsymbol{X}_{(1)}, \boldsymbol{X}_{(2)}$ respectively, each being a vector of two independently sampled random variables drawn uniformly between -1 and 1. This covariate sampling procedure is the same as that implemented by Marchese and Diao (2017), as covariates can be centred and scaled without impacting the interpretation of the model. Therefore, the data is sampled from the following procedure

$$\begin{bmatrix} Y_{i(1)} \\ Y_{i(2)} \end{bmatrix} \sim \mathcal{N}\left( \begin{bmatrix} 1 + \boldsymbol{X}_{(1)}^T \boldsymbol{\beta}_{(1)} \\ 1 + \boldsymbol{X}_{(2)}^T \boldsymbol{\beta}_{(2)} \end{bmatrix}, \boldsymbol{\Sigma} \right)$$

where the covariance matrix $\Sigma$ is given by

$$\boldsymbol{\Sigma} = \begin{bmatrix} 0.8 & 0.4 \\ 0.4 & 1.2 \end{bmatrix}$$

and the mean-model parameters $\boldsymbol{\beta} = [\boldsymbol{\beta}_{(1)}, \boldsymbol{\beta}_{(2)}]^T$ are given by

$$\boldsymbol{\beta}_{(1)} = [-1.0, 0.0]^T, \ \boldsymbol{\beta}_{(2)} = [0.5, 2.2]^T \ .$$

The VSPGLM is fitted with two linear models for $N = 1000$ replications for each sample size ($n = 100, n = 200$), with the results reported in Table 4.1 and histograms of the estimates given in Appendix A.4, Figure A.1, A.2.

| $n$ | Param $\hat{\beta}_{(k')j}$ | Bias | $\hat{\sigma}$ | $\bar{se}(\hat{\beta})$ | 90% | 95% | 99% |
|---|---|---|---|---|---|---|---|
| 100 | $-1.0$ | 0.0062 | 0.1313 | 0.1396 | 87.1 | 94.1 | 98.4 |
| | 0.0 | 0.0045 | 0.1327 | 0.1477 | 86.2 | 90.9 | 97.6 |
| | 0.5 | -0.0155 | 0.1613 | 0.1782 | 86.4 | 91.4 | 97.7 |
| | 2.2 | -0.0094 | 0.1599 | 0.1799 | 85.8 | 92.2 | 97.3 |
| 200 | $-1.0$ | 0.0063 | 0.0967 | 0.1061 | 85.5 | 93.3 | 98.4 |
| | 0.0 | 0.0047 | 0.0974 | 0.1040 | 87.5 | 93.3 | 98.4 |
| | 0.5 | 0.0002 | 0.1191 | 0.1239 | 88.8 | 93.3 | 98.6 |
| | 2.2 | -0.0089 | 0.1182 | 0.1247 | 87.8 | 92.9 | 98.6 |

**Table 4.1:** $N = 1000$ simulations of the VSPGLM for a correlated unconstrained bivariate normal distribution with 2 responses components and 2 coefficients per component.

Similar to Dennis (2021), we find that the estimates are consistent with the bias on average decreasing as $n$ increases, with an average bias of order $10^{-3}$. Furthermore, from Figure A.1, A.2 we observe that the parameter estimates are approximately normally distributed with some skewness seen across the figures. This is likely due to a combination of simulation error the use of empirical likelihood and the convex hull restriction which constrains the estimates for $\boldsymbol{\beta}$ in small to moderate sample sizes. The calculated standard errors $\hat{\sigma}$ are found to be an underestimate of the asymptotic standard errors and as a result, lead to under-coverage of the confidence intervals across all parameters. This behaviour of undercoverage and underestimates of standard errors is also observed in simulation results in Huang (2014) and Wurm

and Rathouz (2018) which are univariate cases of the VSPGLM. For estimates derived from empirical likelihood methods, as a result of the convex hull restriction, it's known that coverage rates tend to converge from below (Diciccio et al. 1991). Thus, as $n$ increases we observe this convergence from below for the coverage rates, and overall indicates consistent estimates and asymptotically correct standard errors.

To correct for this under-coverage, as suggested by Huang (2014) we could instead invert the empirical likelihood ratio test against the asymptotic $\chi^2$-distribution, or for finite samples against a $rF_{r,n-q}$ distribution. This again has the advantage of giving non-symmetric and joint confidence intervals, and correct asymptotic properties without the computation of the empirical covariance matrix $\hat{\Sigma}$ which may introduce numerical error. However, this is computationally expensive because it requires the model to be refit for each parameter in the model. We will explore coverage rates using LRT in Section 4.1.2 for a smaller-scale simulation.

Furthermore, we can explore how the asymptotic efficiency changes as the number of parameters estimated increases for moderate sample sizes by performing the simulations performed by Marchese and Diao (2017). Here the covariates $\boldsymbol{X}_{(1)}, \boldsymbol{X}_{(2)}$ are vectors of three independently sampled uniform random variables between -1 and 1, and the sampling procedure is the same as above except with mean model parameters $\boldsymbol{\beta} = [\boldsymbol{\beta}_{(1)}, \boldsymbol{\beta}_{(2)}]^T$ are now given by

$$\boldsymbol{\beta}_{(1)} = [-1.0, 0.0, 0.67]^T,\ \boldsymbol{\beta}_{(2)} = [0.5, 2.2, -1.25]^T\ .$$

The VSPGLM is similarly fitted with two linear models for $N = 1000$ replications for each sample size ($n = 100, 200$), with the results reported in Table 4.2 and histograms of the estimates given in Appendix A.4, Figure A.3, A.4.

| $n$ | Param $\hat{\beta}_{(k')j}$ | Bias | $\hat{\sigma}$ | $\bar{se}(\hat{\beta})$ | 90% | 95% | 99% |
|---|---|---|---|---|---|---|---|
| 100 | $-1.00$ | 0.0277 | 0.1267 | 0.1555 | 82.4 | 89.0 | 95.3 |
| | 0.00 | -0.0000 | 0.1284 | 0.1548 | 83.5 | 90.5 | 95.9 |
| | 0.67 | -0.0200 | 0.1280 | 0.1590 | 81.5 | 87.1 | 95.8 |
| | 0.50 | -0.0114 | 0.1568 | 0.1861 | 84.3 | 90.2 | 96.1 |
| | 2.20 | -0.0496 | 0.1544 | 0.1944 | 80.1 | 88.1 | 94.9 |
| | $-1.25$ | 0.0529 | 0.1565 | 0.2022 | 78.2 | 87.0 | 94.1 |
| 200 | $-1.00$ | 0.0092 | 0.0948 | 0.1068 | 85.2 | 91.2 | 98.2 |
| | 0.00 | 0.0012 | 0.0959 | 0.1043 | 86.9 | 92.9 | 98.3 |
| | 0.67 | -0.0044 | 0.0954 | 0.1049 | 86.9 | 92.1 | 97.5 |
| | 0.50 | -0.0075 | 0.1171 | 0.1269 | 86.7 | 92.8 | 97.9 |
| | 2.20 | -0.0214 | 0.1157 | 0.1343 | 84.5 | 90.9 | 97.2 |
| | $-1.25$ | 0.0251 | 0.1163 | 0.1347 | 83.6 | 90.5 | 97.1 |

**Table 4.2:** $N = 1000$ simulations of the VSPGLM for a correlated unconstrained bivariate normal distribution with 2 responses components and 3 coefficients per component.

In Table 4.2 we observe that the bias of each parameter is generally higher and standard errors more underestimated than in Table 4.1, but show signs of aligning as $n$ increases. The bias and standard errors for the parameters associated with the second component of the bivariate tend to be more likely as a result of having higher variance compared to the first component. We do find that estimating more parameters results in more undercoverage for the same sample size $n$, but we still exhibit convergence from below as the sample size increases.

Compared with Marchese and Diao (2017), the MDRM for the sample size $n$ has slightly less bias and asymptotically correct standard errors. The smaller bias is a result of modelling on a canonical scale which doesn't have any convex hull restriction, but doing so results in a loss of interpretability of

$\boldsymbol{\beta}$. Furthermore, as highlighted before the $\boldsymbol{\beta}$ needed to be scaled back to the mean scale, but finding this calling can be challenging for more complex structures. Thus, in the standard correctly specified case VSPGLM's estimates are still asymptotically valid and interpretable, approximately normally distributed (Figure A.3, A.4) and consistent as the convex hull issue diminishes as $n$ increases.

Generally, the standard errors of the estimates are difficult to analytically express in the MDRM so Marchese and Diao (2017) standard errors are calculated by bootstrapping with 500 replications. This is computationally expensive even when compared to inverting the likelihood ratio test, but as a result, they achieve asymptotically correct standard errors and coverage rates for the same sample sizes $n$. Similar bootstrapping methods could be employed for VSPGLM, but from Table 4.2, standard errors from the empirical covariance matrix converge to asymptotic standard errors and still act as a good estimate for moderate sample sizes if it's understood that they typically converge from below due to the convex hull.

### 4.1.2 Constrained Multivariate Normal Simulation

Next, we will simulate the data from a bivariate normal distribution with a common set of coefficients $\boldsymbol{\beta}$ across the two components. Using the same covariate sampling procedure, the data is sampled from the following procedure

$$\begin{bmatrix} Y_{i(1)} \\ Y_{i(2)} \end{bmatrix} \sim \mathcal{N}\left( \begin{bmatrix} 1 + \boldsymbol{X}_{(1)}^T \boldsymbol{\beta} \\ 1 + \boldsymbol{X}_{(2)}^T \boldsymbol{\beta} \end{bmatrix}, \boldsymbol{\Sigma} \right)$$

where the covariance matrix $\Sigma$ is given by

$$\boldsymbol{\Sigma} = \begin{bmatrix} 0.8 & 0.4 \\ 0.4 & 1.2 \end{bmatrix}$$

and with the mean-model parameters $\boldsymbol{\beta} = [-1, 0, 0.67]^T$. For this simulation, we will also find coverage rates by inverting the ELRT calibrated against a $F_{1,n-3}$ distribution and compare its performance with the coverage rates found using Wald confidence intervals. The VSPGLM is fitted with two linear models for $N = 1000$ replications for each sample size ($n = 50, 100, 200$) with the results reported in Table 4.3 and histograms of the estimates for $n = 100, 200$ given in Appendix A.4, Figure A.5, A.6.

| | | | | | Wald | | | LRT | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $n$ | Param $\hat{\beta}_j$ | Bias | $\hat{\sigma}$ | $\bar{se}(\hat{\beta})$ | 90% | 95% | 99% | 90% | 95% | 99% |
| 50 | −1.00 | 0.0395 | 0.1311 | 0.1668 | 79.8 | 86.1 | 94.7 | 80.2 | 87.2 | 95.3 |
| | 0.00 | -0.0051 | 0.1388 | 0.1599 | 84.5 | 90.4 | 96.6 | 86.7 | 92.7 | 97.8 |
| | 0.67 | -0.0227 | 0.1350 | 0.1711 | 79.9 | 86.4 | 94.2 | 82.1 | 87.9 | 96.9 |
| 100 | −1.00 | 0.0153 | 0.1007 | 0.1159 | 84.6 | 91.3 | 97.4 | 85.1 | 91.1 | 97.4 |
| | 0.00 | -0.0031 | 0.1041 | 0.1146 | 86.9 | 92.8 | 98.1 | 87.8 | 93.6 | 98.8 |
| | 0.67 | -0.0066 | 0.1024 | 0.1177 | 84.6 | 91.9 | 97.4 | 85.6 | 92.1 | 98.4 |
| 200 | −1.00 | 0.0088 | 0.0743 | 0.0777 | 89.1 | 93.7 | 98.3 | 89.4 | 94.4 | 98.3 |
| | 0.00 | -0.0033 | 0.0757 | 0.0800 | 88.0 | 94.0 | 99.1 | 89.1 | 94.4 | 99.3 |
| | 0.67 | -0.0026 | 0.0749 | 0.0817 | 86.7 | 92.9 | 99.1 | 86.8 | 93.3 | 99.1 |

**Table 4.3:** $N = 1000$ simulations of the VSPGLM for a correlated constrained bivariate normal distribution sharing coefficients between 2 response components.

We observe that similar to the unconstrained multivariate normal simulation, the standard errors underestimate the asymptotic standard errors leading to undercoverage for both types of confidence intervals. However, we note that inverting the ELRT provides an improvement in coverage compared

to the Wald confidence intervals. The coverage improvement is larger for smaller $n$, but as $n$ grows both methods tend towards the expected coverage rates. The estimates are approximately normally distributed with similar skewness observed as in the unconstrained multivariate normal simulations. An interesting observation is that for the parameter $\beta_3 = 0.67$, for $n = 200$ there is a slight right skewness to the histogram in Figure A.6 resulting in slightly more under coverage across the methods compared to the other two parameters, which was not observed for $n = 50$ and $n = 100$.

## 4.2 GDP, Fertility and Urban Percentage

The first application we will consider is a standard example looking at a country statistics dataset from 2017, with two continuous response variables. The dataset obtained from UNData [4] consists of Gross Domestic Product (GDP) per capita in USD, the fertility rate given by number of live births per woman fertility rate and the percentage of population living in urban areas for 229 countries/states in 2017. For this example, we are interested in how the percentage of urban population jointly impacts both GDP per capita and fertility rates (number of live births per woman). The proposed VSPGLM cannot handle missing covariates and responses so after cleaning the data, 198 countries/states remain.

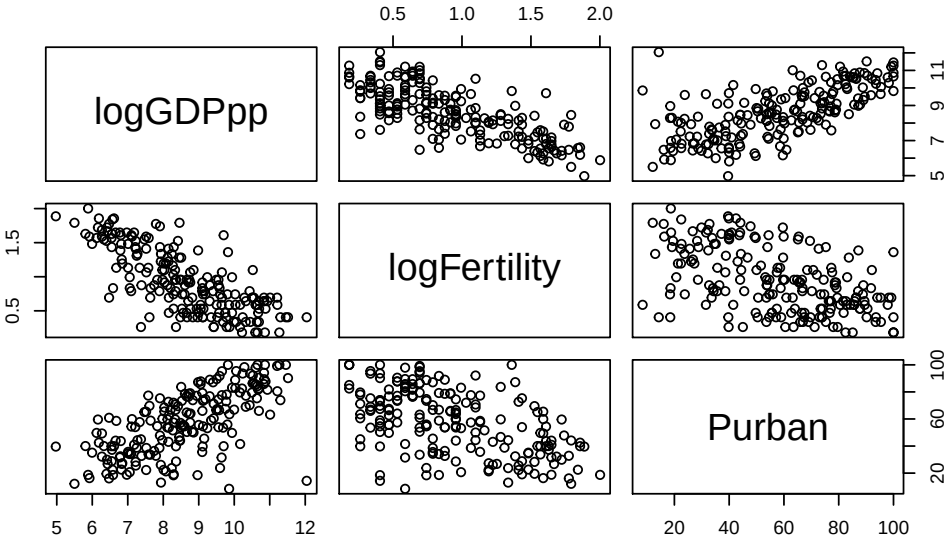


**Figure 4.1:** Pairwise plot of the log of GDP per capita (logGDPpp), log of fertility rates (logFertility) and the percentage of the population living in urban areas (Purban)

Observing an approximate linear relationship in each component in Figure 4.1, we propose a marginal linear mean model for both responses,

$$\mathbb{E}\left[\,\text{logGDPpp} \mid \text{Purban}\,\right] = \mu_{(1)}(\boldsymbol{X}^T\boldsymbol{\beta}) = \beta_{(1)1} + \beta_{(1)2}\text{Purban}\,,$$
$$\mathbb{E}\left[\,\text{logFertility} \mid \text{Purban}\,\right] = \mu_{(2)}(\boldsymbol{X}^T\boldsymbol{\beta}) = \beta_{(2)1} + \beta_{(2)2}\text{Purban}\,.$$

We can fit a joint model with no distributional assumptions using VSPGLM as follows, which produces the model output in Table 4.4

```
% World Countries / States Example
country_model = fit_vspglm(["logGDPpp ~ Purban", "logFertility ~ Purban"],...
                data, {"id", "id"});
```

[4] https://www.kaggle.com/datasets/sudalairajkumar/undata-country-profiles

The parameter estimates and individual Wald tests suggest that there is significantly strong evidence for an increasing relationship between the urban percentage and the log GDP per capita and a decreasing relationship between the urban percentage and the log fertility rates. The fitted VSPGLM is given by (4.1) and is visualised in Figure 4.2.

$$\begin{aligned}\hat{\mathbb{E}}\left[\,\text{logGDPpp} \mid \text{Purban}\,\right] &= \hat{\mu}_{(1)} = 6.250352 + 0.041478\ \text{Purban}\ , \\ \hat{\mathbb{E}}\left[\,\text{logFertility} \mid \text{Purban}\,\right] &= \hat{\mu}_{(2)} = 1.529563 - 0.010363\ \text{Purban}\ . \end{aligned} \tag{4.1}$$

| Component | Coefficient | Estimate | S.E | $T$ | $p$ |
|---|---|---|---|---|---|
| logGDPpp | Intercept | 6.250352 | 0.19969 | 31.300 | $< 2.22 \times 10^{-16}$ |
| | PUrban | 0.041478 | 0.00297 | 13.969 | $< 2.22 \times 10^{-16}$ |
| logFertility | Intercept | 1.529563 | 0.07578 | 20.184 | $< 2.22 \times 10^{-16}$ |
| | PUrban | -0.010363 | 0.00107 | -9.659 | $< 2.22 \times 10^{-16}$ |
| **Log-Likelihood** ($\ell$) | -990.0390 | | | | |

**Table 4.4:** Coefficient summary for GDP-Fertility model

The parameter estimates are similar to those from fitting two univariate normal linear models ($\hat{\boldsymbol{\beta}} = [6.163, 0.0424, 1.5374, -0.01044]^T$), but the advantage of the VSPGLM is that it accounts for the correlation between components. Furthermore, we can also estimate the correlation between the components and how this changes as the fitted values change in the model. In Figure 4.2 we observe that for high fertility rates and low GDP per capita, we have a strong negative correlation, but as Purban increases the negative correlation becomes weaker. Using VSPGLM we can estimate a correlation surface on the entire convex hull of $\boldsymbol{Y}$, but the current implementation of VSPGLM only estimates a correlation surface on the convex hull of the fitted values $\hat{\boldsymbol{\mu}}$.

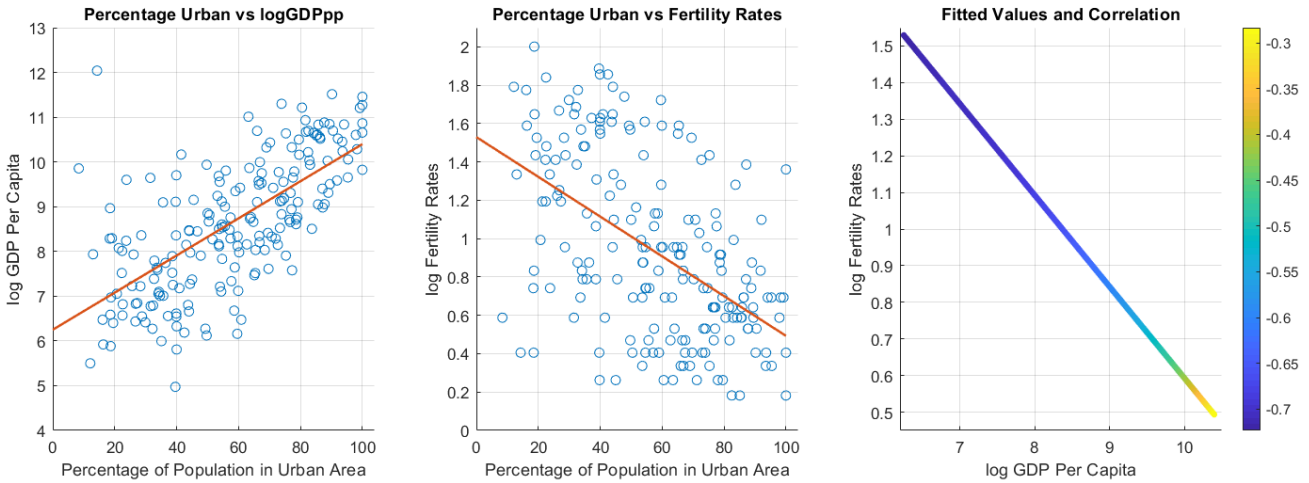


**Figure 4.2:** Fitted marginal components (red) and estimated correlation between components.

The output in Table 4.4 performs single parameter hypothesis tests, but we could also perform the compound hypothesis that the urban percentage has no relationship with logFertility and logGDPpp, $H_0 : \beta^*_{(1)2} = \beta^*_{(2)2} = 0$. Fitting an intercept-only model with the code below finds $\ell_{intercept} = -1047.1$.

```
% Compound hypothesis for world countries example
country_model_comp = fit_vspglm_constraint(["logGDPpp ~ Purban", "logFertility ~
    Purban"],data, {"id", "id"}, [2,4], [0,0]);
```

Calibrating the ELRT against a $F_{2,194}$ distribution we find that the percentage of the population in urban areas is associated with fertility rates and GDP per capita.

$$-2(\ell_{intercept} - \ell) = -2(-1047.1 + 990.0390) = 114.1220$$
$$p\text{-value} = \mathbb{P}\left(F_{2,194} \geq \frac{114.1220}{2}\right) < 2.22 \times 10^{-16} .$$

As the VSPGLM estimates the joint distribution of (logGDPpp, logFertility), we can also visualise this joint distribution for any urban percentage by plotting the estimated probability masses on the observed support. In Figure 4.3 the estimated joint density is shown for 3 values of urban percentage (8.4%, 64%, 100%), where on repeated response values the probability mass is accumulated.

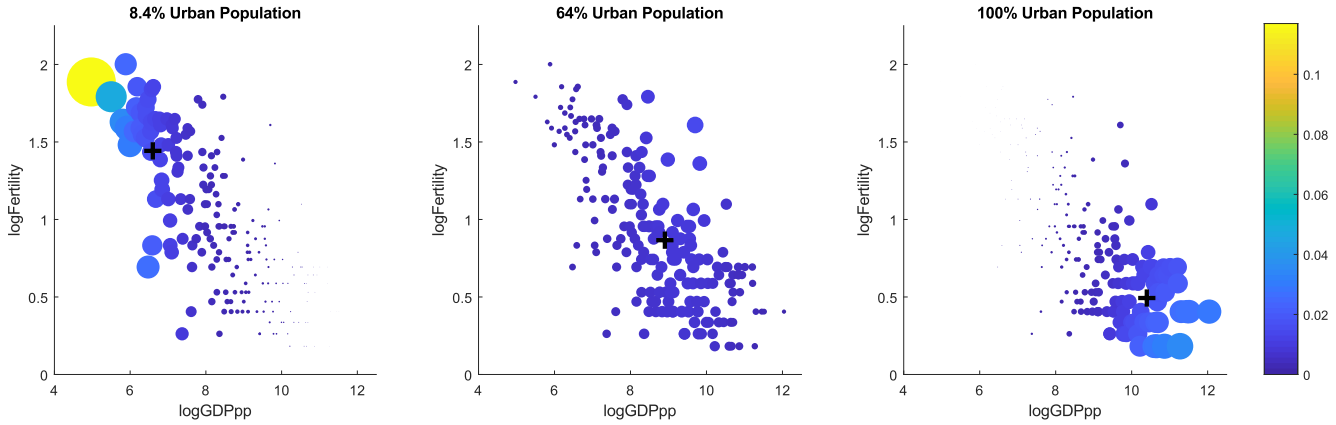


**Figure 4.3:** Estimated joint probability density function for different percentages of urban population. The estimated mean given by the black plus sign and probability masses area is sized according to its value.

Figure 4.3 shows that for lower urban percentages, the probability distribution has greater density for lower GDP per capita and higher fertility rates on the log scale, with essentially no mass allocated to high values of logGDPpp. As the urban percentage increases the tails of the density become fatter and more density accumulates symmetrically about the fitted mean. For higher urban percentages, the density accumulates at regions of high GDP per capita and low fertility rates as similarly predicted by the fitted model, with thinner left tails. Here we see the advantage of utilising exponential tilting as the probability density shifts together with the estimated mean.

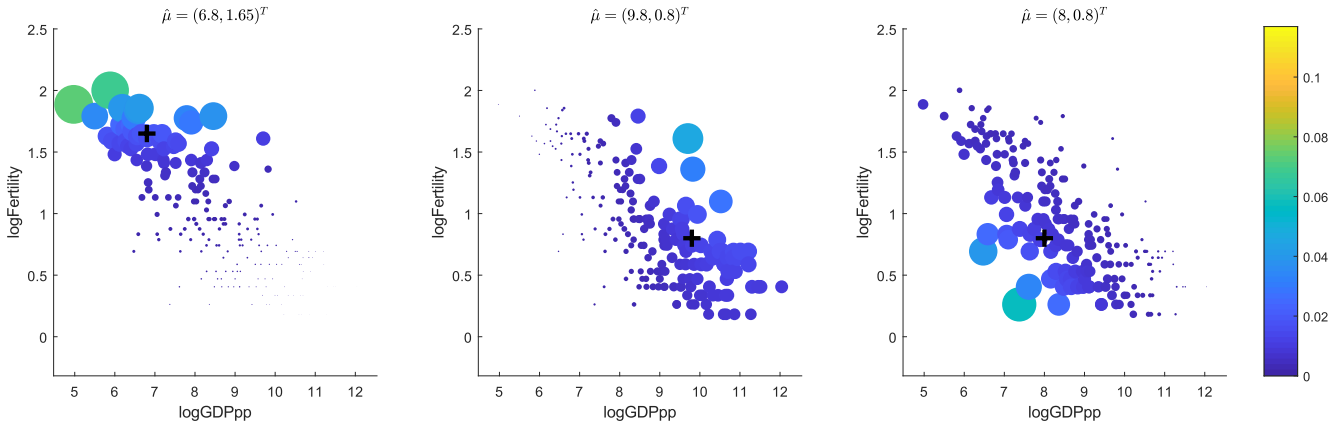


**Figure 4.4:** Estimated joint probability density function for new estimated values of the response

Another feature of VSPGLM is that post-fitting the model given a vector of new estimated means $\hat{\boldsymbol{\mu}}$ which was originally not observed in the dataset, it's able to estimate a vector of tilts $\hat{\boldsymbol{\theta}}$ and use

the estimated reference distribution $\hat{\boldsymbol{p}}$ to predict a joint density about this fitted mean. In Figure 4.4 we present the joint density for different responses to visualise how the probability density shifts as the mean of the distribution is tilted outside the fitted linear mean model. Similar to Figure 4.3, as the mean shifts towards the boundary of the convex hull, the joint distribution becomes skewed as probability mass accumulates at points on the boundary. Here VSPGLM is able to predict probability mass beyond the observed support but as the support of $\boldsymbol{Y}$ is truncated to the observed support due to the empirical likelihood approach, probability masses can only be placed on observations.

## 4.3 Burn Injury

Next, to illustrate the flexibility of the proposed VSPGLM, let us consider fitting and analyzing the VSPGLM to the Burn Injury dataset. The dataset first reported by Fan and Gijbels (1996) contains 981 observations of hospital patients admitted with some amount of 3rd-degree burns. There are two response variables, the relative burn severity $Y_{(1)} = \log(\text{burn area} + 1)$ which is a continuous variable and the disposition of death $Y_{(2)}$ is a binary variable with 1 for death and 0 for survival. The burn area is the percentage of the patient's body that was covered in 3rd degree burns at the time of admittance to the hospital. This data has been analysed in VGLM literature Song (2007) using Gaussian Couplas and by Huang (2017) using Generalized Estimating Equations (GEEs) to investigate how age (in weeks) affects patients burn severity and probability of death. As burn severity is continuous and disposition of death is binary, the following set of default marginal mean models is considered by both Song (2007) and Huang (2017)

$$\mathbb{E}\left[Y_{(1)}|\text{age}\right] = \mu_{(1)} = \beta_{(1)1} + \beta_{(1)2}\text{age} \tag{4.2}$$

$$\mathbb{E}\left[Y_{(2)}|\text{age}\right] = \mu_{(2)} = \frac{\exp\{\beta_{(2)1} + \beta_{(2)2}\text{age}\}}{1 + \exp\{\beta_{(2)1} + \beta_{(2)2}\text{age}\}}, \tag{4.3}$$

with corresponding marginal variances

$$\text{Var}(Y_{(1)|\text{age}}) = \phi_{(1)} = \sigma^2, \ \text{Var}(Y_{(2)|\text{age}}) = \phi_{(2)}\mu_{(2)}(1-\mu_{(2)}) = \mu_{(2)}(1-\mu_{(2)}) \ . \tag{4.4}$$

Here $\mu_{(1)}$ is the expected log-burn area given the age of the patient and $\mu_{(2)}$ is the probability of death from burn injury given the age of the patient. To handle continuous-binary response data, hierarchical continuous-binary models such as the method proposed by Fitzmaurice and Laird (1995) could be used. However, the model relies on the continuous response $Y_{(1)}$ being conditional on the binary response $Y_{(2)}$ and as outlined by Huang (2017), for this example it's more reasonable to assume that instead, the disposition of death is conditional on burn severity.

For the Gaussian Coupla approach explored by Song (2007), the marginal distributions using the mean-variance functions above are first formalized as

$$Y_{(1)}|\text{age} \sim \mathcal{N}(\mu_{(1)}, \phi_{(1)}^2), \quad Y_{(2)}|\text{age} \sim \text{Bernoulli}(\mu_{(2)}) \ .$$

Then, using Gaussian Couplas the joint density can be expressed as

$$f(\boldsymbol{y}) = \begin{cases} \varphi(y_{(1)}; \mu_{(1)}, \sigma^2)(1 - C_1^*(\mu_{(2)}, z_1|\alpha), & y_2 = 0 \\ \varphi(y_{(1)}; \mu_{(1)}, \sigma^2)C_1^*(\mu_{(2)}, z_1|\alpha), & y_2 = 1 \end{cases} \tag{4.5}$$

where $\varphi(\cdot\ ; \mu_{(1)}, \sigma^2)$ is the density of a $\mathcal{N}(\mu_{(1)}, \sigma^2)$, $z_1 = (y_{(1)} - \mu_{(1)}/\sqrt{\sigma^2}\ )$ and $C_1^*(a, b|\alpha) = \Phi\left(\dfrac{\Phi^{-1}(a) - \alpha b}{\sqrt{1-\alpha^2}}\right)$. Thus, the log-likelihood can be expressed by

$$\ell(\boldsymbol{\beta}, \sigma^2, \alpha) = \sum_{i=1}^{n} \log \varphi(y_{i(1)}; \mu_{i(1)}, \sigma^2) + \sum_{i \in I_0} \log\left\{(1 - C_1^*(\mu_{i(2)}, z_{i1}|\alpha)\right\} + \sum_{i \in I_0^c} \log\left\{(C_1^*(\mu_{i(2)}, z_{i1}|\alpha)\right\} \ , \tag{4.6}$$

where $I_0 = \{i : y_{i(2)} = 0\}$. The model is fitted using the Gauss-Newton algorithm outlined in Chapter 1 and the results are reported in Table 4.5. Song (2007) estimates the association parameter to be $\alpha = 0.8$, indicating a strong association between the components, and finds age to have significant marginal relationships with both burn severity and risk of death.

Huang (2017) considers fitting the GEEs with the working independence correlation model, and then adjusting for within-subject correlation afterwards using the sandwich estimator (1.35) for the covariance matrix. As a result, the parameter estimates are the same as fitting the data with two univariate models. As seen in Table 4.5, the model similarly finds the marginal effect of age on both a patient's burn severity and probability of death was significant.

The proposed VSPGLM is fitted using the same marginal mean models with the identity and logit link respectively with no additional assumptions, using the code

```
% Burn Injury Model
burns_model= fit_vspglm(["burn_severity ~ age", "death ~ age"],data,{"id", "logit"});
```

| Model | Component | Coefficient | Estimate | S.E | $T$ | $p$ |
|---|---|---|---|---|---|---|
| VSPGLM | Burn Severity | Intercept | 6.7318 | 0.0646 | 104.19 | $< 2.22 \times 10^{-16}$ |
| | | Age | 0.0027 | 0.0019 | 1.3875 | 0.1656 |
| | Death | Intercept | -3.6570 | 0.2199 | -16.627 | $< 2.22 \times 10^{-16}$ |
| | | Age | 0.0501 | 0.0042 | 11.86 | $< 2.22 \times 10^{-16}$ |
| G-Coupla | Burn Severity | Intercept | 6.6980 | 0.0479 | 139.73 | $< 2.22 \times 10^{-16}$ |
| | | Age | 0.0039 | 0.0012 | 3.16 | 0.0016 |
| | Death | Intercept | -4.0521 | 0.1658 | -24.44 | $< 2.22 \times 10^{-16}$ |
| | | Age | 0.0527 | 0.0028 | 19.13 | $< 2.22 \times 10^{-16}$ |
| GEE | Burn Severity | Intercept | 6.7118 | 0.0640 | 104.845 | $< 2.22 \times 10^{-16}$ |
| | | Age | 0.0035 | 0.0017 | 2.105 | 0.0353 |
| | Death | Intercept | -3.6891 | 0.2643 | -13.959 | $< 2.22 \times 10^{-16}$ |
| | | Age | 0.0509 | 0.0051 | 10.035 | $< 2.22 \times 10^{-16}$ |
| Univariate | Burn Severity | Intercept | 6.7118 | 0.0690 | 97.24 | $< 2.22 \times 10^{-16}$ |
| | | Age | 0.0035 | 0.0018 | 1.97 | 0.0488 |
| | Death | Intercept | -3.6891 | 0.2342 | -17.78 | $< 2.22 \times 10^{-16}$ |
| | | Age | 0.0509 | 0.0046 | 11.07 | $< 2.22 \times 10^{-16}$ |
| **VSPGLM Log-Likelihood ($\ell_{full}$)** | | | -6668.4 | | | |

**Table 4.5:** Coefficient summary for different fitted models on Burns Injury dataset

The fitted VSPGLM is given by

$$\hat{\mathbb{E}}\left[Y_{(1)}|\text{age}\right] = \hat{\mu}_{(1)} = 6.7318 + 0.0027\text{age}\,. \tag{4.7}$$

$$\hat{\mathbb{E}}\left[Y_{(2)}|\text{age}\right] = \hat{\mu}_{(2)} = \frac{\exp\{-3.6570 + 0.0501\text{age}\}}{1 + \exp\{-3.6570 + 0.0501\text{age}\}}\,. \tag{4.8}$$

The parameter estimates and fitted model is similar to the GEE and univariate model, which we can visualise in Figure 4.5. The model also estimates a positive correlation between the components where the correlation increases as both burn severity and probability of death increase, as seen on the correlation surface plot. Similar to Song (2007) and Huang (2017), we can test whether age is associated with burn severity, $H_0 : \beta_{(1)2} = 0$ using an ELRT calibrated against a $F_{1,981-4}$ distribution. Fitting the reduced model with the code below finds a log-likelihood of $\ell_{interceptburn} = -6669.3$.

```
% Burn Injury Model intercept only for burn_severity
```

```
burns_model_reduced = fit_vspglm_constraint(["burn_severity ~ age", "death ~
    age"],data,{"id", "logit"},2,0);
```

Thus, the p-value of the test is can found as

$$-2(\ell_{interceptburn} - \ell_{full}) = -2(-6669.3 + 6668.4) = 1.8$$
$$p\text{-value} = \mathbb{P}\left(F_{1,977} \geq 1.8\right) = 0.1800242\,.$$

Furthermore, using the ELRT we can also test for the compound hypothesis that age has no relationship with both burn severity and death $H_0 : \beta^*_{(1)2} = \beta^*_{(2)2} = 0$. Running the code below finds a log-likelihood of $\ell_{intercept} = -6757.7$.

```
% Burn Injury Model intercept only model for each component
burns_model_intercept = fit_vspglm_constraint(["burn_severity ~ age", "death ~
    age"],data,{"id", "logit"},[2,4],[0,0]);
```

Calibrating against a $2F_{2,981-4}$ distribution, we find the following p-value.

$$-2(\ell_{intercept} - \ell_{full}) = -2(-6757.7 + 6668.4) = 178.6$$
$$p\text{-value} = \mathbb{P}\left(F_{2,977} \geq \frac{178.6}{2}\right) < 2.22 \times 10^{-16}\,.$$

Therefore, by the ELRT we find that age has a significant marginal effect on the disposition of death, but not on the burn severity of the patient. This is interesting as the conclusion is different to that found by fitting univariate GLMs, GEEs or G-Coupla.

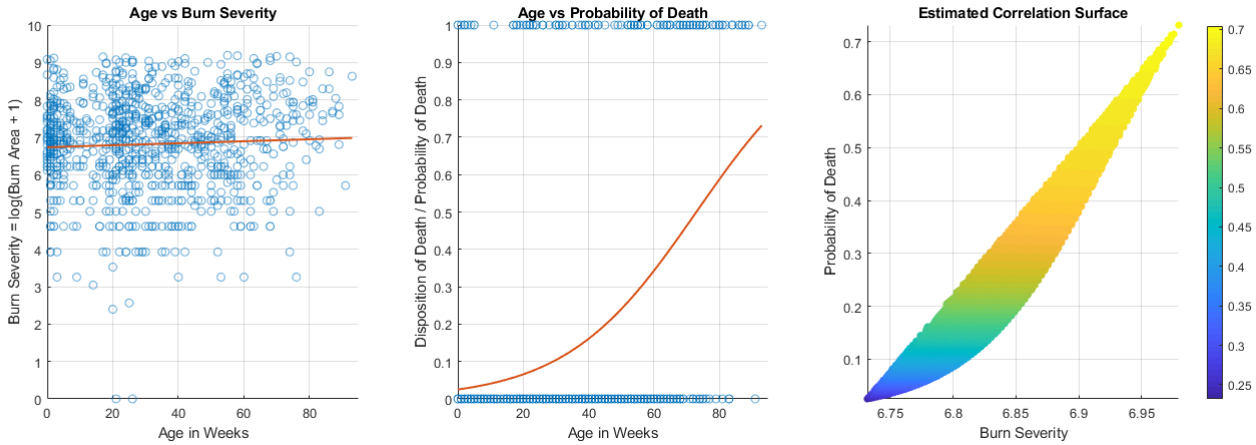


**Figure 4.5:** Fitted marginal mean models (red) and estimated correlation surface between components.

## 4.4 Butterfly Abundance in Boulder County

Now let us focus on some applications of VSPGLM in species modelling by considering the dataset of butterfly species counts reported in Oliver et al. (2006). Although the sampling of the dataset is limited, the example is included to show the performance of VSPGLM on an arbitrary number of response components which include positive and negative correlation (Appendix A.5, Figure A.7) as some generalizations of Negative Binomial and Poisson distributions don't allow for both positive and negative correlation between components. The dataset contains the abundance/counts of 33 butterfly species observed at 66 locations in Boulder County, Colorado as well as the type of habitat (Hayfield, Mixed, Short, Tall), the height of surrounding buildings and the concentration of vegetation at each location (Table 4.6).

| Colias Philodice | Pieris Rapae | ... | Habitat | Building | Urban Vegetation |
|---|---|---|---|---|---|
| 19 | 11 | ... | Hayfield | 0.44 | 26.59 |
| 9 | 17 | ... | Hayfield | 0.44 | 26.30 |
| 0 | 68 | ... | Tall | 6.36 | 20.82 |
| ⋮ | ⋮ | ... | ⋮ | ⋮ | ⋮ |

**Table 4.6:** Sample data from Butterfly dataset

The butterfly dataset was previously analysed by Hui et al. (2013) who compared modelling the 14 butterfly species with more than 10 total observations with separate Species Distribution Models (SDMs) and Species Archetypes Models (SAMs). Separate SDMs are modelled as an independent GLM for each species, where for presence-abense data a Bernoulli distribution with logit link is assumed, and for abundance data, a Negative Binomial distribution with a log link is assumed to account for any potential overdispersion. For the Butterfly dataset, Hui et al. (2013) verifies graphically a quadratic form of overdispersion $\text{Var}(Y_{(k)i}) = \mu_{(k)i} + \phi_{(k)}\mu^2_{(k)i}$ where $\phi_{(k)}$ is a dispersion parameter. When considering multiple species, separate SDMs don't account for the correlation between the species counts and have trouble modelling rare species with a low number of observations.

The Species Archetypes Models (SAMs) proposed by Dunstan et al. (2011) use a finite mixture of regression models (McLachlan and Peel 2000) to cluster species together into archetypes to borrow information across species, allowing for better prediction of rarer species. The SAM supposes $G < K$ archetype species and has a likelihood function of the form

$$L(\boldsymbol{\beta}|\boldsymbol{Y}, \boldsymbol{X}) = \sum_{g=1}^{G} \pi_g \prod_{i=1}^{N} f(Y_{(k)i}; \mu_{(k)i}(\boldsymbol{X}^T_{(k)i}\boldsymbol{\beta}_{(k)}), \phi_{(k)}) \tag{4.9}$$

with proportions $\sum_{g=1}^{G} \pi_g = 1$ and the mean model with link function $g(\cdot)$ given by $g(\mu_{(k)}) = \boldsymbol{X}^T_{(k)}\boldsymbol{\beta}_{(k)}$. From this, the fitted mean of each species can be found by taking a linear combination of the group's fitted means, weighted by the posterior probability of being in each group. It was found that for the Butterfly dataset, there was not much difference between the two models, however, there was evidence that SAMs ($G = 2$) provided improvements in predicting rarer species which was the same conclusion found in other datasets.

As an alternative, we can jointly model the 14 same species using VSPGLM to account for the correlation between species, with separate mean models (4.10) and mean model parameters $\boldsymbol{\beta}_{(k)}$ and the same set of covariates $\boldsymbol{X}_i \in \mathbb{R}^6$ across the 66 locations.

$$\mu_{(k)}\left(\boldsymbol{X}^T_{(k)}\boldsymbol{\beta}_{(k)}\right) = \mathbb{E}\left[\boldsymbol{Y}_{(k)}|\boldsymbol{X}\right] = \exp\left\{\boldsymbol{X}^T_{(k)}\boldsymbol{\beta}_{(k)}\right\}, \quad k = 1, \ldots, 14. \tag{4.10}$$

The fitted model coefficients for each species are given in Appendix A.5 Table A.2, noting that Hayfield is taken as the baseline habitat. An advantage of VSPGLM is that with the estimated reference distribution $\hat{F}$ and the estimated tilt parameters $\hat{\boldsymbol{\theta}}$, we are able to estimate the covariance matrix of $\boldsymbol{Y}$ at each observation,

$$\hat{\Sigma}_{\boldsymbol{Y}_i} = \sum_{j=1}^{n} (\boldsymbol{Y}_j - \hat{\boldsymbol{\mu}}_i)(\boldsymbol{Y}_j - \hat{\boldsymbol{\mu}}_i)^T \hat{p}_j \exp\{\hat{\boldsymbol{\theta}}_i^T \boldsymbol{Y}_j + \hat{b}_i\}, \quad i = 1, 2, \ldots, n\,. \tag{4.11}$$

In the context of the Butterfly dataset, we are able to use $\hat{\Sigma}_{\boldsymbol{Y}_i}$ to construct correlation matrices at each location to both, allowing us to estimate the correlation between the species and how this changes across the locations. In the context of non-normal, non-linear data it may not be appropriate to assess correlation quantitatively, but we can still interpret the relationships qualitatively. For this example,

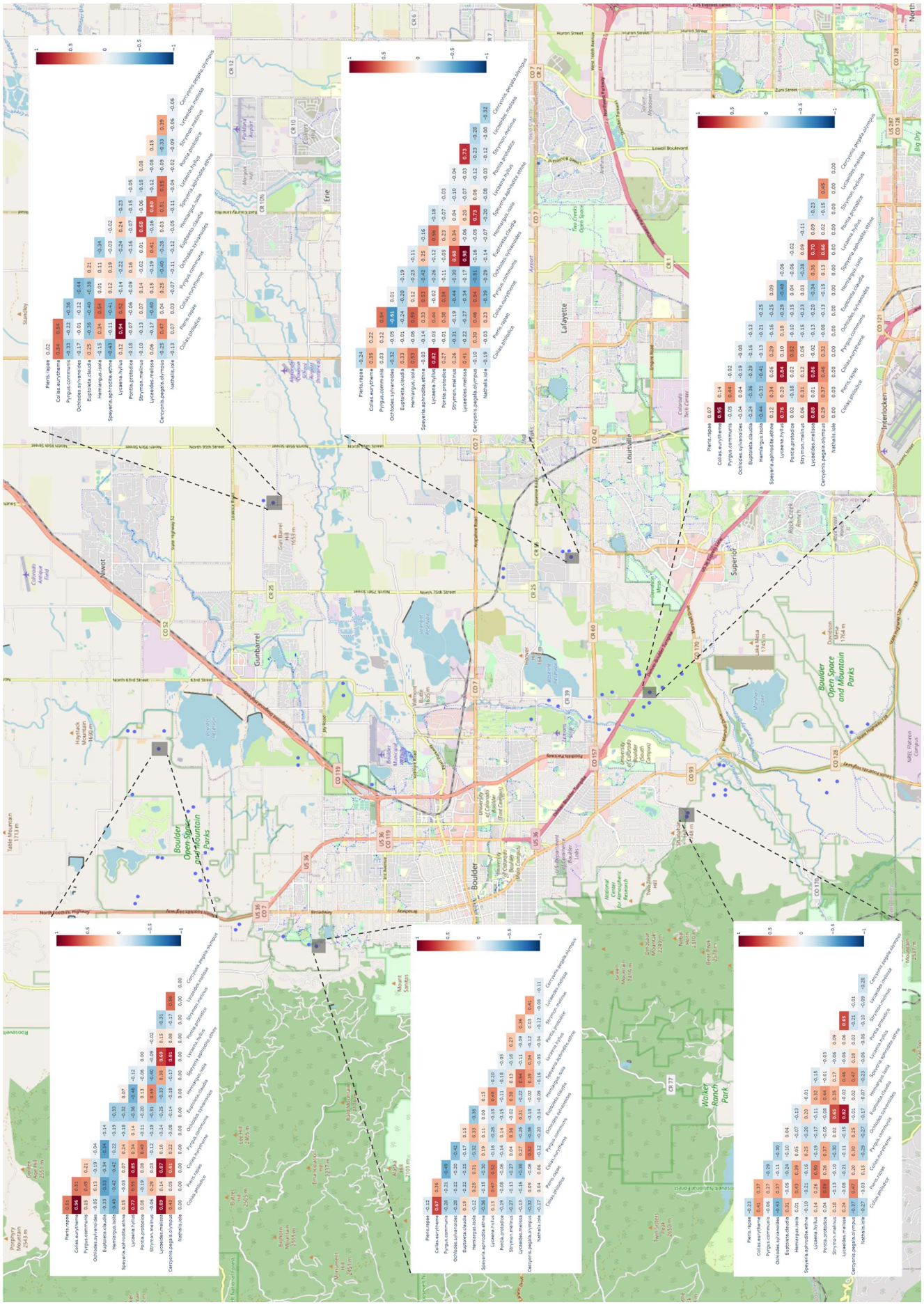

**Figure 4.6:** Map of Boulder County, Colorado with the observed locations and predicted correlations between butterfly species

an interactive map [5] which displays the estimated correlation between the 14 butterflies at all the locations was created. Example correlation matrices from the interactive map are displayed in Figure 4.6.

In Figure 4.6 we see that as the various locations have different habitat, building and vegetation covariates, the predicted correlations between the species changes. For locations close to each other, they generally have similar covariates resulting in similar predicted correlations, and this change in correlation between locations is smooth as the covariates and fitted means as a result of the exponential tilting. Generally, we observe both positive and negative correlations similar to the sample correlation matrix, where species with a higher number of counts tend to have higher correlations with the other species. The negative correlations indicate potential competition for resources between the species, whereas positive correlations indicate cooperation between the species.

Highlighting a few interesting relationships, across the location there are examples of pairs of species that are positively correlated at one location and become negatively correlated at another despite their estimated counts both increasing. For example, we see that at site 3, Colias Philodice and Hemiargus Isola have a moderately strong positive correlation with estimated counts of 2.82 and 0.13 respectively, but at site 18 they have a moderately strong negative correlation with estimated counts of 22.73 and 0.97. The same behaviour can also be seen when considering Pieris Rapae and Euptoieta Claudia which at site 62 have a strong positive correlation, but at other sites where their counts increase such as at site 18 and 27, they have an estimated negative correlation. Generally, this large shift in correlation occurs when the estimated counts of one species have a much larger estimated increase in counts, and highlights how a species of butterfly in an environment can be more abundant, but also more competitive with other butterfly species.

For species such as Nathalis iole which only has 11 sightings or Lycaena hyllus which has 0 sightings in short habitats, the correlation between other species is 0 and the fitted coefficients are very large indicating overfitting. Therefore, VSPGLM similar to SDMs tends to overfit coefficients for species with a low number of observations across the locations and habitat types.

Within the VSPGLM framework, there are no explicit ways for overcoming this, but fitting a SAMs model finds $G = 2$ optimal archetypes and the posterior probability for each species being in a particular is very close to 1. As a naive alternative, we could group the butterflies into the estimated archetypes treating each species in an archetype as the same, and then fitting a two-response VSPGLM. This places more assumptions on the underlying model to enforce this sharing of information, but it allows for an estimate of shared effects of the covariates for a species archetype and by reducing the dimensionality down allows for an estimated correlation surface between the archetypes. Fitting VSPGLM on these two grouped archetypes across the 66 locations produces the following fitted model

$$\begin{aligned}\hat{\mu}_{(1)} &= \exp\left(3.761 + 0.028X_{building} + 0.006X_{vege} - 2.484X_{mixed} - 3.237X_{short} - 0.861X_{tall}\right)\\ \hat{\mu}_{(2)} &= \exp\left(0.922 - 0.001 + X_{building} + 0.009X_{vege}0.965X_{mixed} - 0.297X_{short} + 0.193X_{tall}\right)\ .\end{aligned} \quad (4.12)$$

However, a disadvantage of this approach is that VSPGLM doesn't have notions of the posterior probability of belonging to a group, so unlike a SAM we cannot back out estimated means for each species. Figure 4.7 presents the estimated correlation surface of the two archetypes across the convex hull of the fitted values.

[5] https://github.com/Kyle-Macaskill/Honours-VSPGLM

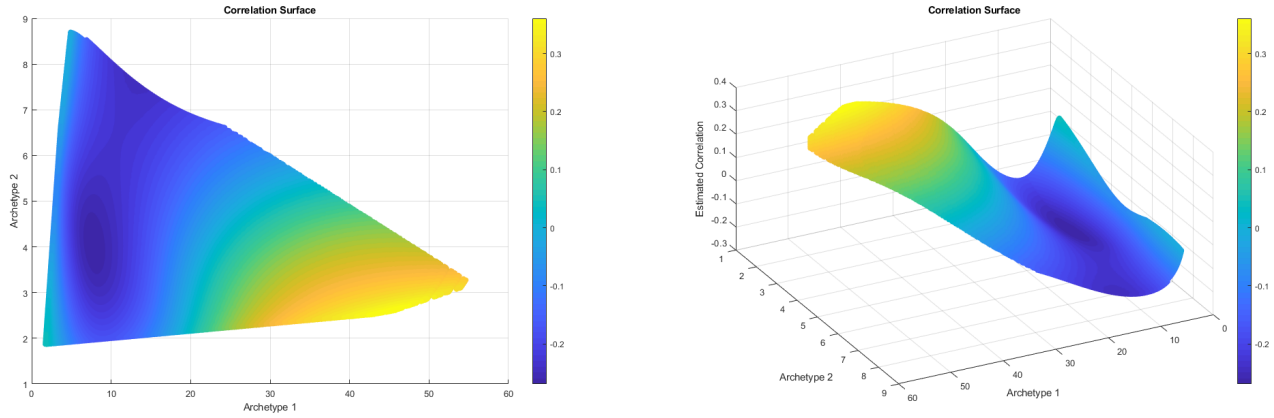


**Figure 4.7:** Fitted correlation surface for the two butterfly species archetypes

We observe that Archetype 1 is generally more abundant and as its estimated counts increase the counts for Archetype 2 remain low. Interestingly, as Archetype 1 estimated counts increase, the correlation between the species increases and is positive. As the estimated counts for Archetype 2 increase, the estimated counts for Archetype 1 remain low and the correlation between the species is negative. We interestingly note that the fitted correlation surface is smoothed as the fitted counts change and the shape is non-linear.

Although we can continue to consider the 14-species joint model, for the remaining analysis we consider modelling the 3 most abundant butterfly species (*Colias philodice, Pieris rapae, Colias eurytheme*) as each has a count higher than the number of locations. It should be noted that by considering a subset of species in this context, the joint distribution is marginalising over the other species observed species. Thus, with the same log-link mean models for each species, we can fit VSPGLM with the following code. Firstly, we are interested in performing model selection between the full model and a model without building and vegetation covariates.

```
% Three Species Butterfly Model
butterfly_model= fit_vspglm(["colias_philodice ~ (building, vegetation, habitat)",...
                             "pieris_rapae ~ (building, vegetation, habitat)",...
                             "colias_eurytheme ~ (building, vegetation, habitat)"],...
                             tbl, {"log","log","log"});
```

The fitted coefficients are presented in Appendix A.5 Table A.3 but we are interested in testing the the compound hypothesis $H_0 : \boldsymbol{\beta}^*_{building,urban.vege} = 0$. The fitted coefficients of the habitat-only model can be found in Table 4.7, so noting that $\ell_{habitat} = -234.4406$ we can perform an ELRT calibrated against a $F_{6,48}$ distribution as follows

$$-2(\ell_{habitat} - \ell_{full}) = -2(-234.4406 + 228.8709) = 11.1394.$$
$$\implies p\text{-value} = \mathbb{P}\left(F_{6,66-18} \geq \frac{11.1394}{6}\right) = 0.1079691.$$

As a result, we find no evidence that building sizes and levels of vegetation have an impact on the 3 butterfly species and opt to select the habitat-only model using a parsimony argument.

| Species | Coefficient | Estimate | S.E | $T$ | $p$ |
|---|---|---|---|---|---|
| *Colias Philodice* | Intercept | 3.17 | 0.22 | 14.4 | $< 2.22 \times 10^{-16}$ |
| | Habitat.Mixed | -2.33 | 0.38 | -6.21 | $4.81 \times 10^{-8}$ |
| | Habitat.Short | -4.27 | 0.53 | -7.99 | $4.15 \times 10^{-11}$ |
| | Habitat.Tall | -1.94 | 0.38 | -5.17 | $2.70 \times 10^{-6}$ |
| *Pieris Rapae* | Intercept | 2.22 | 0.28 | 7.96 | $4.67 \times 10^{-11}$ |
| | Habitat.Mixed | -2.86 | 0.71 | -4.07 | $1.37 \times 10^{-4}$ |
| | Habitat.Short | -2.73 | 0.98 | -2.78 | $7.17 \times 10^{-3}$ |
| | Habitat.Tall | 0.02 | 0.44 | 0.05 | 0.9615 |
| *Colias Eurytheme* | Intercept | 2.46 | 0.26 | 9.53 | $9.28 \times 10^{-14}$ |
| | Habitat.Mixed | -2.23 | 0.38 | -5.92 | $1.53 \times 10^{-7}$ |
| | Habitat.Short | -3.09 | 0.47 | -6.62 | $9.21 \times 10^{-9}$ |
| | Habitat.Tall | -0.95 | 0.30 | -3.17 | 0.0024 |
| **Log-Likelihood** ($\ell_{habitat}$) | | -234.4406 | | | |

**Table 4.7:** Coefficient summary for the 3 species habitat only model

Furthermore, given that only habitat covariates are left we can investigate whether or not the impact of the habitat on the species counts is shared across the species. We are able to do this by fitting a constrained model that enforces the same habitat coefficients across the marginal mean models in the VSPGLM,

$$\mu_{(k)} = \exp\{\beta_0 + \beta_1 X_{mixed} + \beta_2 X_{short} + \beta_3 X_{tall}\},\ k = 1, 2, 3\ .$$

The syntax for fitting the constrained model is given by the code below, with the output presented in Table 4.8

```
% Butterfly Constrained Habitat Model 3 Species
butterfly_model_contraint = fit_vspglm( "(Colias_philodice, Pieris_rapae,
    Colias_eurytheme) ~ (Habitat_Mixed,Habitat_Short,Habitat_Tall)", tbl,
    {"log","log","log"})
```

| Coefficient | Estimate | S.E | $T$ | $p$ |
|---|---|---|---|---|
| Intercept | 2.02 | 0.18 | 11.5 | $< 2.22 \times 10^{-16}$ |
| Habitat.Mixed | -1.66 | 0.28 | -5.89 | $1.73 \times 10^{-7}$ |
| Habitat.Short | -2.85 | 0.36 | -7.94 | $4.96 \times 10^{-11}$ |
| Habitat.Tall | -0.41 | 0.22 | -1.81 | 0.07389 |
| **Log-Likelihood** ($\ell_{constrained}$) | -254.6481 | | | |

**Table 4.8:** Coefficient summary for the 3 species constrained habitat Model

Therefore, as the constrained model is a nested model of the full model, taking the constrained model as the null model $H_0$ we can perform an ELRT calibrated against a $F_{6,66-12}$ distribution to test whether there is a significant difference between the models.

$$-2(\ell_{habitat} - \ell_{constrained}) = -2(-254.6481 + 234.4406) = 40.415$$
$$\implies p\text{-value} = \mathbb{P}\left(F_{6,66-12} \geq \frac{40.415}{6}\right) = 2.30 \times 10^{-5}\ .$$

Therefore, there is significantly strong evidence to reject the null hypothesis, suggesting that there are the habitats impact the butterfly species counts differently.

## 4.5 Hunua Plant Species

To further highlight the flexibility of the proposed VSPGLM and compare it with existing frameworks in VGLM literature, we will consider a simple application with a trivariate binary response where the specification of joint distributions may be challenging. Huang (2017) explores an application of a dataset from Yee (2015), which records the presence/absence of 17 plant species at 392 locations in the Hunua ranges in Auckland, with altitude as the only predictor variable. Within the **VGAM** framework proposed by Yee (2015), we can consider modelling 3 common plant species (Kniexc, Beitaw and Cyadea) using a loglinear trivariate binomial regression.

| Species | Counts |
|---|---|
| Kniexc | 230 |
| Beitaw | 161 |
| Cyadea | 129 |

**Table 4.9:** Plant Species Counts on Hunua Mountain Ranges

The parametric joint distribution for this case is given by

$$\begin{aligned}\log\left\{\mathbb{P}(Y_{(1)}=y_{(1)},Y_{(2)}=y_{(2)},Y_{(3)}=y_{(3)}|\ \boldsymbol{x})\right\} &= u_0(\boldsymbol{x})+u_1(\boldsymbol{x})y_{(1)}+u_2(\boldsymbol{x})y_{(2)}+u_3(\boldsymbol{x})y_{(3)}\\ &\quad+u_{12}(\boldsymbol{x})y_{(1)}y_{(2)}+u_{13}(\boldsymbol{x})y_{(1)}y_{(3)}+u_{23}(\boldsymbol{x})y_{(2)}y_{(3)}\end{aligned} \tag{4.13}$$

where $y_j = 0$ or 1, $\boldsymbol{\beta} = (u_1, u_2, u_3, u_{12}, u_{23}, u_{13})^T$ are parameters estimated in the model. A set of logistic mean models is specified for each component

$$\mathbb{P}(Y_{(k)}=1|\ \text{altitude}) = \mu_{(k)} = \frac{\exp\{\beta_{(k)0}+\beta_{(k)1}\ \text{altitude}\}}{1+\exp\{\beta_{(k)0}+\beta_{(k)1}\ \text{altitude}\}} \tag{4.14}$$

for $k \in \{\text{Cyadea, Beitaw and Kniexc}\}$, with a corresponding set of variance functions given by

$$\text{Var}(Y_{(k)}=1|\ \text{altitude}) = \mu_{(k)}(1-\mu_{(k)})\ . \tag{4.15}$$

Huang (2017) explored using the plug-in sandwich estimator (1.35) from the GEE framework in the **VGAM** framework to obtain asymptotically correct standard errors. Here the within-vector correlations are dependent on the marginal means and $\{\gamma_{k_1,k_2}\}$ parameters, by assuming a constant log-odds ratio model to represent the dependence structure between plant species at each location

$$\frac{\mathbb{P}(Y_{(k_1)}=1,Y_{(k_2)}=1)\mathbb{P}(Y_{(k_1)}=0,Y_{(k_2)}=0)}{\mathbb{P}(Y_{(k_1)}=1,Y_{(k_2)}=0)\mathbb{P}(Y_{(k_1)}=0,Y_{(k_2)}=1)} = e^{\gamma_{k_1,k_2}},\ k_1 \neq k_2 \tag{4.16}$$

The proposed VSPGLM can be fitted to the dataset using the same log-linear marginal mean models (4.14) and no further assumptions using the code below, with the results being reported in Table 4.10.

```
% Hunua Plant Species Model
plant_model = fit_vspglm(["cyadea ~ altitude", "beitaw ~ altitude", "kniexc ~
    altitude"],data,{"logit", "logit", "logit"});
```

| Model | Component | Coefficient | Estimate | S.E (Adjusted) | $T$ (Adjusted) |
|---|---|---|---|---|---|
| VSPGLM | Cyadea | Intercept | -0.755 | 0.179 (0.179) | -4.21 (-4.21) |
| | | Altitude ($\times 10^3$) | -0.253 | 0.878 (0.873) | 0.29 ( 0.29) |
| | Beitaw | Intercept | -1.047 | 0.184 (0.213) | -5.68 (-4.93) |
| | | Altitude ($\times 10^3$) | 4.083 | 0.920 (1.113) | 4.44 ( 3.60) |
| | Kniexc | Intercept | -0.069 | 0.175 (0.186) | -0.39 (-0.37) |
| | | Altitude ($\times 10^3$) | 2.654 | 0.925 (1.075) | 2.87 ( 2.47) |
| VGLM | Cyadea | Intercept | -0.977 | 0.222 (0.170) | -4.40 (-5.74) |
| | | Altitude ($\times 10^3$) | -0.570 | 0.921 (0.813) | -0.62 (-0.70) |
| | Beitaw | Intercept | -1.890 | 0.265 (0.181) | -7.14 (-10.5) |
| | | Altitude ($\times 10^3$) | 3.850 | 0.962 (0.870) | 4.00 ( 4.42) |
| | Kniexc | Intercept | -0.377 | 0.197 (0.197) | -1.92 (-1.92) |
| | | Altitude ($\times 10^3$) | 1.611 | 0.970 (1.188) | 1.66 ( 1.36) |
| **VSPGLM Log-Likelihood ($\ell$)** | | | -2327.7 | | |

**Table 4.10:** Coefficient summary for Hunua Plant Species Model

Note that the estimates and standard errors for Altitude are multiplied by $10^3$, so the relative difference in magnitude between the models is quite small and the VSPGLM finds similar estimates to those from loglinear trivariate binomial regression in Yee (2015). Huang (2017) indicates that for the VGLM, the adjusted and unadjusted standard errors are very similar indicating that the working variance and dependence models seem reasonable, although the adjusted standard errors are asymptotically valid. The unadjusted standard errors for VSPGLM are also found to be similar to the adjusted standard errors from the sandwich estimator in VGLM, although with less assumptions. The sandwich-adjusted standard errors are also provided for VSPGLM, which are mostly similar except for the altitude covariate for Beitaw, where the adjusted standard error is larger than the others. This observed difference in the finite sample setting may be appropriate as the adjustments are only asymptotically valid and each has different rates of convergence.

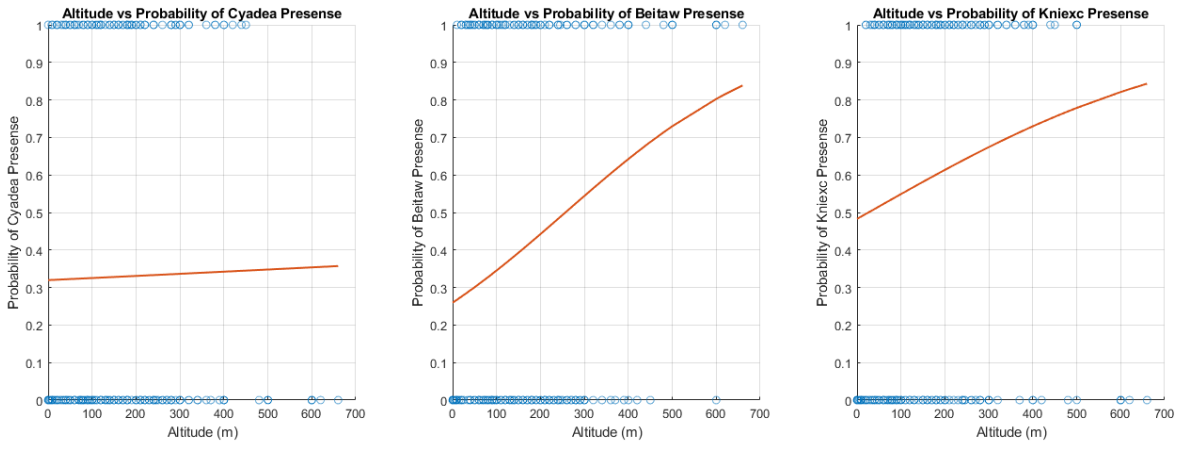


**Figure 4.8:** Fitted marginal mean models for Hunua Plant Species Model

Assuming VSPGLM is correctly specified, we can test the hypothesis for each coefficient for altitude $H_0 : \beta^*_{(k)1} = 0, H_1 : \beta^*_{(k)1} \neq 0$ for $k = 1, 2, 3$ using an ELRT calibrated against a $F_{1,392-6}$ distribution we find in Table 4.11 that altitude has a significant marginal effect on the presence of Beitaw and Kniexc.

| Component | LRTS | $p$-value |
|---|---|---|
| Cyadea | 0.0804 | 0.7769 |
| Beitaw | 22.738 | $2.6\times10^{-6}$ |
| Kniexc | 8.7303 | 0.003 |

**Table 4.11:** Empirical Likelihood Ratio Test for Altitude Coefficient

Note that on the Hunua ranges, 17 plant species were recorded so fitting the data across only 3 plant species marginalises over the unaccounted plant species. Although the log-linear binomial regression can be extended to further higher dimensions, the implementation in Yee (2015) currently only considers the bivariate and trivariate cases. In contrast, the VSPGLM and its implementation can handle an arbitrary number of vector dimensions. To highlight this advantage of the VSPGLM implementation, we will introduce the plant species Kuneri, Daccup and Cyamed with 136, 120 and 104 counts respectively and consider a 6-dimensional plant species model, with a set of logistic mean models and no coefficients shared between components. The resulting model coefficients and inference from the fitted VSPGLM on these 6 plant species can be found in Appendix A.6 Table A.4.

Similar to Section 4.4, as the VSPGLM estimates a joint distribution at each observation, it allows for the calculation of an estimated correlation matrix for $\boldsymbol{Y}$ at each altitude in the dataset and we can explore how the correlation changes between the presence of plant species changes as altitude increases. An interactive script to explore the correlation matrices can be found on Github [6], and unadjusted correlation matrices at 4 particular altitudes can be found in Figure 4.9.

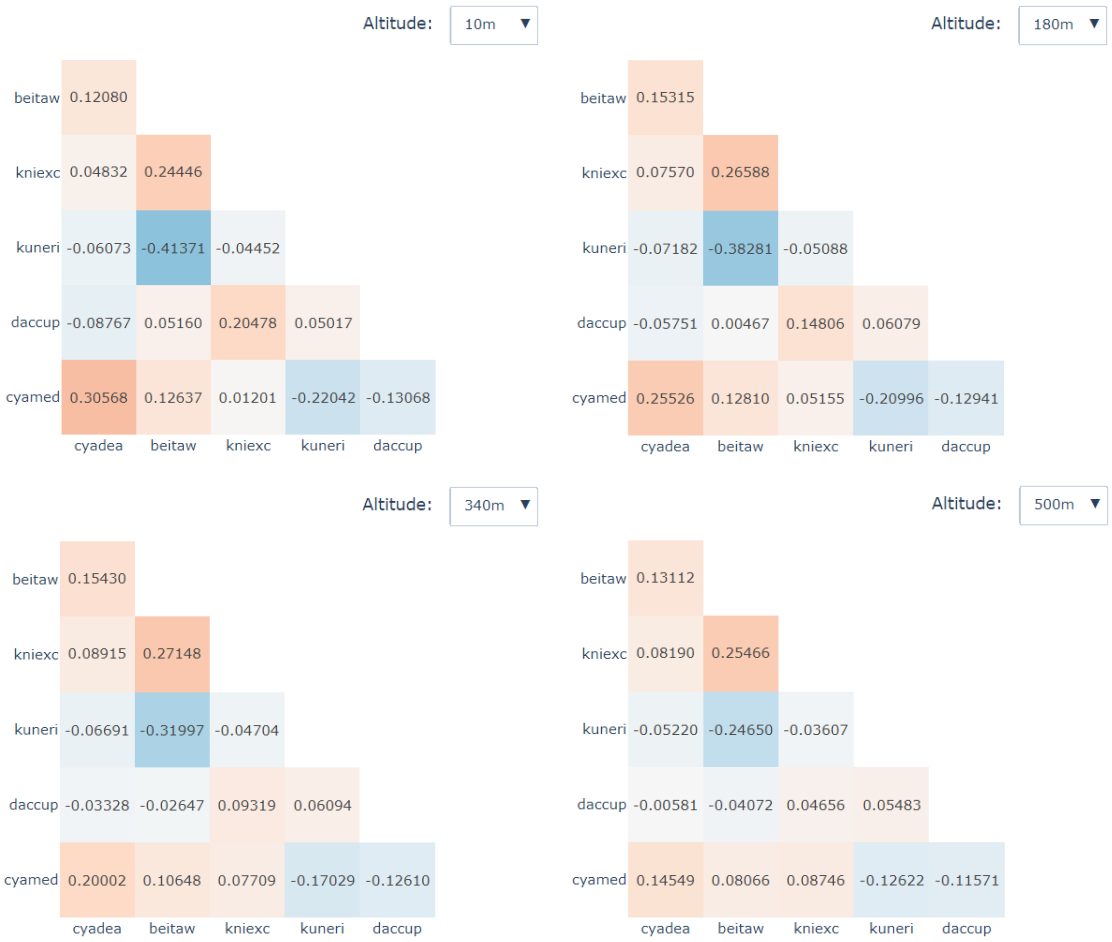


**Figure 4.9:** Estimated correlation matrices at different altitudes between 6 Hunua plant species

[6] https://github.com/Kyle-Macaskill/Honours-VSPGLM

Although considering the linear correlation values between presence/absence data may not be directly interpretable, if analysed qualitatively it can still provide insight on whether there is any competition between plant species. Highlighting a few interesting relationships in Figure 4.9, we find that across all altitudes Kniexc and Beitaw are the two species with the most amount of sightings across all the altitude levels and VSPGLM predicts them to have a consistent positive correlation across the altitudes. We find that Kuneri is negatively correlated with most plant species, but mainly with Beitaw which becomes more negatively correlated as the altitude decreases, indicating potential competition between plant species at lower altitudes.

## 4.6 Sorbinil Retinopathy Trial

For the next application, we consider analysing a dataset of $n = 41$ subjects in a Sorbinil Retinopathy Trial (1990) explored by Rosner et al. (2006). First subjects were exposed to an allergen for two qualifying visits and asked to use an ordinal itching scale from 0 (no itch) to 4 (severe incapacitating itch) in increments of 0.5, and any subjects with itching scores of +2 or higher in each eye were randomized into 4 treatment groups. Upon a third visit, 16 hours after an exposure to the same allergen the subjects were administered to their left and right eyes a combination of a treatment with sorbinil and a placebo, with their itching scores being assessed 5 minutes after the treatment. The full dataset can be seen in Table 4.12, where the response variable is a pair of itching scores $\boldsymbol{Y} = (Y_L, Y_R)^T$ for the left and right eye respectively.

| Treatment Combinations | | | | | | | |
|---|---|---|---|---|---|---|---|
| $n = 6$ | | $n = 14$ | | $n = 14$ | | $n = 7$ | |
| sorbinil Left | sorbinil Right | sorbinil Left | Placebo Right | Placebo Left | sorbinil Right | Placebo Left | Placebo Right |
| 2.0 | 2.0 | 1.0 | 1.5 | 2.5 | 2.0 | 3.0 | 3.0 |
| 1.0 | 1.0 | 2.0 | 2.5 | 2.5 | 2.5 | 2.0 | 3.0 |
| 0.5 | 2.0 | 3.0 | 1.0 | 3.0 | 3.0 | 2.5 | 2.5 |
| 2.5 | 1.0 | 2.0 | 3.0 | 2.5 | 2.0 | 1.0 | 3.0 |
| 3.0 | 2.5 | 3.0 | 2.5 | 1.0 | 0.5 | 2.0 | 2.5 |
| 2.0 | 2.5 | 2.0 | 3.0 | 2.0 | 0.0 | 2.0 | 1.0 |
| | | 3.0 | 3.0 | 3.0 | 2.5 | 2.0 | 2.0 |
| | | 0.5 | 1.5 | 3.0 | 1.0 | | |
| | | 3.0 | 3.0 | 2.0 | 1.5 | | |
| | | 3.0 | 3.0 | 0.5 | 0.0 | | |
| | | 3.0 | 3.0 | 2.5 | 1.5 | | |
| | | 1.0 | 2.0 | 2.0 | 2.0 | | |
| | | 1.0 | 2.0 | 2.5 | 2.5 | | |
| | | 1.5 | 2.5 | 2.5 | 2.5 | | |

**Table 4.12:** Itching scores for each of the four treatment groups from 0 to 4 in 0.5 increments.

The ordinal nature of the data means that the predicted means must be constrained on the support $[0, 4] \times [0, 4]$ for any value of the covariates, which poses a challenge for standard VGLM methods. One way to enforce these constraints is to transform each response $Y_L, Y_R$ to the interval [0,1] so that the transformed responses are pseudo proportions (McCullagh and Nelder (1989) pp 328-332). Then with these transformed variables $\tilde{Y}_L, \tilde{Y}_R$, the following logistic mean models can be used

$$\mathbb{E}\left[\tilde{Y}_L | \mathbb{1}_L(\text{Sorbinil})\right] = \mu_{(1)} = \frac{\exp\left\{\beta_{(1)0} + \beta_{(1)1}\mathbb{1}_L(\text{Sorbinil})\right\}}{1 + \exp\left\{\beta_{(1)0} + \beta_{(1)1}\mathbb{1}_L(\text{Sorbinil})\right\}}, \tag{4.17}$$

$$\mathbb{E}\left[\tilde{Y}_R | \mathbb{1}_R(\text{Sorbinil})\right] = \mu_{(2)} = \frac{\exp\left\{\beta_{(2)0} + \beta_{(2)1}\mathbb{1}_R(\text{Sorbinil})\right\}}{1 + \exp\left\{\beta_{(2)0} + \beta_{(2)1}\mathbb{1}_R(\text{Sorbinil})\right\}}, \tag{4.18}$$

where $\mathbb{1}_L(\text{Sorbinil}), \mathbb{1}_R(\text{Sorbinil})$ is an indicator for the sorbinil treatment in the left and right eye respectively. This transformation is considered appropriate as it is reasonable to assume that the variance of $Y_L, Y_R$ approaches 0 as the expected itching score tends to either 0 or 4, which can be represented by the following working variance functions where $\phi$ is some constant dispersion parameter.

$$\text{Var}\left[\tilde{Y}_L|\mathbb{1}_L(\text{Sorbinil})\right] = \phi\,\mu_{(1)}(1-\mu_{(1)}), \tag{4.19}$$

$$\text{Var}\left[\tilde{Y}_R|\mathbb{1}_R(\text{Sorbinil})\right] = \phi\,\mu_{(2)}(1-\mu_{(2)})\,. \tag{4.20}$$

However, in standard VGLM approaches such as those based on the use of Couplas, it can be challenging to construct marginal distributions while satisfying the mean and variance constraints given above. As a result, semiparametric VGLM methods are better suited to the problem, which is what Huang (2017) explores by using Generalized Estimating Equations with the logistic mean models and a within-subject correlation given by

$$\text{Corr}(\tilde{Y}_L, \tilde{Y}_R|\boldsymbol{X}) = \rho(\gamma; \mu_{(1)}, \mu_{(2)})\,. \tag{4.21}$$

As an alternative modelling approach, we can use the proposed VSPGLM which doesn't require any variance or correlation assumptions. Furthermore, VSPGLM implicitly enforces the mean constraints through the convex hull, meaning that it can be fit on the original variables $(Y_L, Y_R)$ with identity link functions to increase the interpretability of the estimated parameters.

First, let us consider fitting VSPGLM with separate linear mean models (4.22) which allows for a different treatment effect across the left and right eye

$$\begin{aligned}\mathbb{E}\left[Y_L|\mathbb{1}_L(\text{Sorbinil})\right] &= \mu_{(1)} = \beta_{(1)0} + \beta_{(1)1}\mathbb{1}_L(\text{Sorbinil})\\ \mathbb{E}\left[Y_R|\mathbb{1}_R(\text{Sorbinil})\right] &= \mu_{(2)} = \beta_{(2)0} + \beta_{(2)1}\mathbb{1}_R(\text{Sorbinil})\,.\end{aligned} \tag{4.22}$$

| Component | Coefficient | Estimate | S.E | $T$ | $p$ |
|---|---|---|---|---|---|
| Left Eye | Intercept | 2.196 | 0.1571 | 13.976 | $< 2.22 \times 10^{-16}$ |
| | Sorbinil Treatment | -0.203 | 0.2198 | -0.928 | 0.35895 |
| Right Eye | Intercept | 2.401 | 0.1266 | 18.960 | $< 2.22 \times 10^{-16}$ |
| | Sorbinil Treatment | -0.667 | 0.2171 | -3.073 | 0.00385 |
| **Log-Likelihood** ($\ell_{separate}$) | | -146.9701 | | | |

**Table 4.13:** Coefficient summary for separate sorbinil models

The model results are presented in Table 4.13 and interestingly we find strong evidence that the sorbinil treatment was only effective in the right eye and not the left eye. The same result was found using GEEs but given the nature of the data, Huang (2017) was interested in testing whether there was symmetry between the the two eyes. To do this, we fit a symmetric model (4.23), where the coefficients for the left and right eye are constrained to be equal

$$\begin{aligned}\mathbb{E}\left[Y_L|\mathbb{1}_L(\text{Sorbinil})\right] &= \mu_{(1)} = \beta_0 + \beta_1\mathbb{1}_L(\text{Sorbinil})\\ \mathbb{E}\left[Y_R|\mathbb{1}_R(\text{Sorbinil})\right] &= \mu_{(2)} = \beta_0 + \beta_1\mathbb{1}_R(\text{Sorbinil})\,.\end{aligned} \tag{4.23}$$

The syntax for fitting a symmetric model using VSPGLM is as follows

```
% Symmetric Sorbinil Model
[rossner_model, formulas] = fit_vspglm(['(yL, yR) ~ ((xL & xR))'], tbl, {'id', 'id'});
```

| Coefficient | Estimate | S.E | $T$ | $p$ |
|---|---|---|---|---|
| Intercept | 2.303 | 0.1110 | 20.754 | $< 2.22 \times 10^{-16}$ |
| Sorbinil Treatment | -0.434 | 0.1414 | -3.0683 | 0.0039043 |
| **Log-Likelihood** ($\ell_{symmetric}$) | -147.9416 | | | |

**Table 4.14:** Coefficient summary for symmetric sorbinil models

Our null hypothesis is given by $H_0 : \boldsymbol{\beta}^*_{(1)} = \boldsymbol{\beta}^*_{(2)}$, so calibrating the ELRT against a $F_{2,41-4}$ distribution,

$$-2(\ell_{symmetric} - \ell_{separate}) = -2\left(-147.9416 + 146.9701\right) = 1.943$$
$$\implies p\text{-value} = \mathbb{P}\left(F_{2,37} \geq \frac{1.943}{2}\right) = 0.3879609\ .$$

Thus, we find no evidence against the null hypothesis, so by a parsimony argument we find the symmetric model to be an adequate model for the data. To continue the testing performed by Huang (2017), we can consider testing whether applying the sorbinil treatment to one eye impacts the itchiness score of the other eye. To do this, we fit the following model with a common additive inference term

$$\mathbb{E}\left[Y_L | \mathbb{1}_L(\text{Sorbinil})\right] = \mu_{(1)} = \beta_0 + \beta_1 \mathbb{1}_L(\text{Sorbinil}) + \beta_2 \mathbb{1}_R(\text{Sorbinil})$$
$$\mathbb{E}\left[Y_R | \mathbb{1}_R(\text{Sorbinil})\right] = \mu_{(2)} = \beta_0 + \beta_1 \mathbb{1}_R(\text{Sorbinil}) + \beta_2 \mathbb{1}_L(\text{Sorbinil})\ .$$

which can be fitted with the following syntax

```
% Additive Inference Sorbinil Model
[rossner_model, formulas] = fit_vspglm(["(yL, yR) ~ ((xL & xR), (xR & xL))"], tbl,
    {"id", "id"});
```

This produces the results in Table 4.15, where we find no evidence to suggest there is an additive interference effect.

| Coefficient | Estimate | S.E | $T$ | $p$ |
|---|---|---|---|---|
| Intercept | 2.288 | 0.1996 | 11.459 | $6.79 \times 10^{-14}$ |
| Sorbinil Treatment | -0.420 | 0.2029 | -2.072 | 0.045009 |
| Inference Term | 0.019 | 0.1987 | 0.093 | 0.92652 |
| **Log-Likelihood** ($\ell_{additive}$) | -147.9372 | | | |

**Table 4.15:** Coefficient summary for additive inference sorbinil models

Thus, similar to Huang (2017) we find the final fitted model to be the symmetric model which is summarized in Table 4.14. The interpretation of the symmetric model is that the sorbinil treatment is associated with an estimated reduction of 0.434 in the itchiness score from 2.303 to 1.867, which is very similar to the estimated reduction of 0.444 found by Huang (2017) using GEEs with an arbitrary correlation on the transformed variables.

Finally, with the estimated $\hat{F}$ and tilts $\hat{\boldsymbol{\theta}}$ on the final fitted model, we can visualise a joint distribution of $\boldsymbol{Y}$ for each treatment group on the discrete lattice between 0 to 4. The estimated distributions are given in Figure 4.10, where the area of probability masses is proportional to weight and the scaling is consistent across the distributions. Stepping through the treatment groups, when sorbinal is applied to both eyes we see that the distribution is relatively symmetric about the diagonal and relatively consistent weightings across the domain. However, when only one eye receives the sorbinil treatment, the distribution tilts and places a greater probability of masses to higher itching scores on the eye that received the placebo. This effect is symmetric whether we are considering applying the sorbinal treatment to the left or right eye because the model enforces symmetry between

the eyes. Finally, in the placebo treatment group, we observe that the probability mass is tilted to higher itching scores for both the left and right eye on the diagonal.

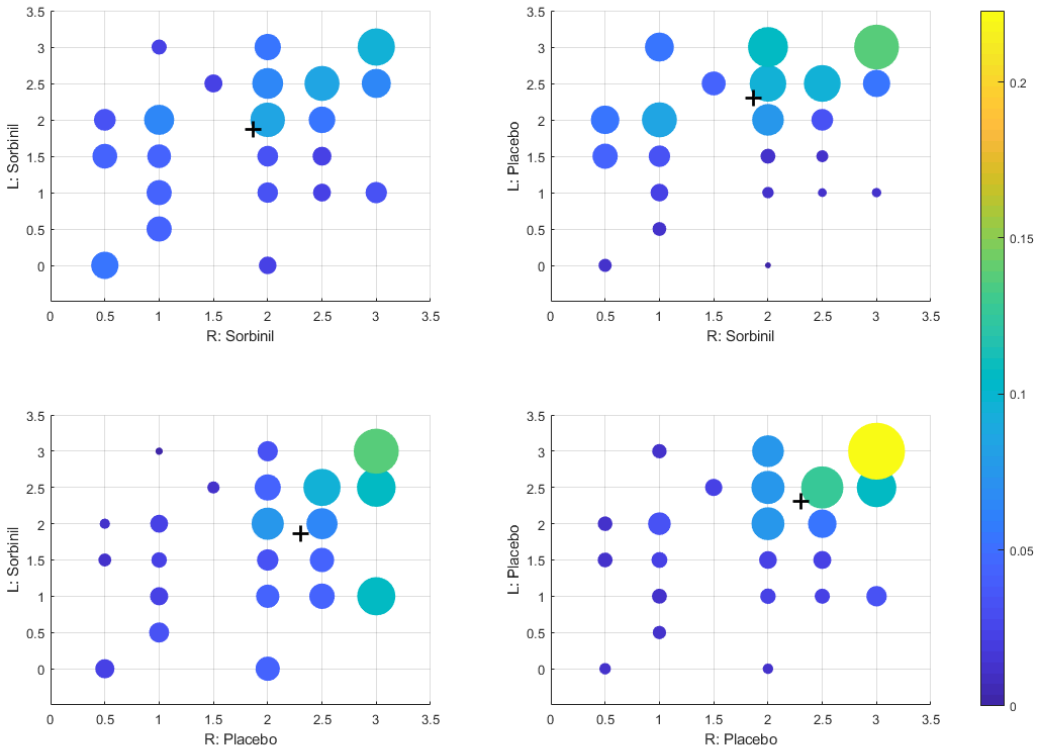


**Figure 4.10:** Estimated joint distribution for the symmetric sorbinil model for each treatment group. The itching scores for the right and left eye are on the $x$ and $y$ axis respectively. The estimated mean for each treatment group is indicated by the black cross and the probability masses area is scaled according to its value.

Observe again that probability masses are only placed on the observed support, which means that no probability masses are placed above a score of 3 for either eye. However, VSPGLM can still estimate probability beyond what was observed and in the event where this occurs, the probability mass accumulates at the boundary of the convex hull, which occurs in the placebo treatment group for an itchiness score of 3.

## 4.7 Kenyan School Children Dietary Intervention

For the final application, we showcase the flexibility of the VSPGLM framework by considering a longitudinal study with repeated measures across time. Neumann et al. (2003) conducted a study to examine the relationship between intake of animal-source foods and growth, cognitive development and physical activity cognitive development, which was done through a randomised controlled school feeding intervention study in Kenyan primary school children. A total of 554 school children from 12 schools in Kenya were randomized into four nutritional intervention groups, Meat, Milk, Energy (Calorie) and Control. For meat, milk and energy groups, the foods were added to a local plant-based dish Githeri which served as the vehicle. The study was carried out over 21 months where various measurements were taken at a baseline visit and 4 proceeding visits.

We will consider a subset of 524 observations without missing covariates presented in Weiss (2005)

which is unbalanced and concerns cognitive and arithmetic outcomes of the study[7]. The cognitive assessments were carried out using Raven's Colored Progressive Matrices (Raven et al. 1995) and the arithmetic test was adapted from the Wechsler Intelligence Scales for Children (Wechsler 2003). Initially, the dataset was analysed by Whaley et al. (2003) using hierarchical linear random effects models, but more recently has been explored by Han et al. (2014) using the Conditional Empirical Likelihood (CEL) method to analyse the cognitive component of the data.

Providing a brief overview of the CEL method, suppose $\boldsymbol{g}_i(\boldsymbol{\beta})$ is a residual vector for samples $i = 1, \ldots, n$ and to handle an unbalanced structure let $\tau_i$ denote the follow-up pattern for subject $i$ and $S_i = \{j : 1 \leq j \leq N, \tau_j = \tau_i\}$ be the stratum that belongs to subject $i$. Letting $\Omega$ denote the collection of distinctive follow-up patterns, then we have that for any $\omega \in \Omega$,

$$\mathbb{E}\left[g_u(\boldsymbol{\beta})|\boldsymbol{X}_i\right] = 0, \quad i \in S^{\omega}$$

Using the notation above, we have that if the data is balanced then all subjects are in a single stratum. Then empirical probabilities $p_{ij}$ are placed on the observed support of the residuals $\{g_j(\boldsymbol{\beta}) : j \in S_i\}$. This can be seen as a conditional empirical probability as it's the probability of observing a residual for $j$ conditional on the covariates of subject $i$. From this a weighted smoothed empirical log-likelihood likelihood can be obtained, so that the estimate for $\boldsymbol{\beta}$ is so that it maximizes the empirical likelihood,

$$\hat{\boldsymbol{\beta}}_{CEL} = \arg\max_{\boldsymbol{\beta}} \max_{p_{ij}} \sum_{i=1}^{n} \sum_{j \in S_i} w_{ij} \log\{p_{ij}\} \tag{4.24}$$

subject to $p_{ij} \geq 0, \sum_{j \in S_i} p_{ij} = 1$ and $\sum_{j \in S_i} p_{ij} \boldsymbol{g}_j(\boldsymbol{\beta}) = 0$ for all $i = 1, \ldots, n$. The weights $w_{ij}$ are estimated using a kernel-based weight calculation, where the complexity of this calculation increases as the complexity of the problem increases.

The key advantage of CEL similar to VSPGLM is that it doesn't require an explicit variance-covariance structure, as it is instead consistently estimated in a nonparametric manner. Thus, CEL is similar to VSPGLM in that it uses empirical likelihood, but is a framework is suited for longitudinal data where it can handle missing data by stratifying subjects according to their follow-up patterns. VSPGLM, in contrast, is a more general framework for handling vector responses, so to model this longitudinal study we have that the response is given by $\boldsymbol{Y}_i = (Y_{i(1)}, Y_{i(2)}, Y_{i(3)}, Y_{i(4)})^T$ where $Y_{i(k)}$ is the Raven score for the subject $i$ at the $k$-th visits after the baseline, $k = 1, 2, 3, 4$. However, unlike GEEs and CEL a drawback of VSPGLM can only be applied to a balanced study meaning that observations with missing responses had to be removed, resulting in 469 observations remaining.

Firstly, we are interested in modelling the cognitive outcomes of the children for the different treatments, where subjects at the baseline visit are all taken to belong to the control group and as a result, the baseline Raven score (braven) is treated as a covariate in the model. Other covariates include the age of the child at the baseline (age), the baseline social-economic status (ses) determined by a survey, an indicator for whether the child is a male, the treatment group (meat, milk, energy) where the baseline is the control group, and the time of visit relative to a baseline time across all observations.

As highlighted by Whaley et al. (2003), there is no treatment difference at time zero so any treatment effect observed will be presented as a difference in the rate of increase in the child's cognitive performance. Therefore, the model should include the treatment by time interaction rather than just a treatment effect. Han et al. (2014) handles this by using the average relative time of each visit

[7]Dataset available at https://robweiss.faculty.biostat.ucla.edu/book-data-sets

(avg_time) as the iteration term, resulting in the following marginal mean model

$$\begin{aligned}\mathbb{E}\left[\text{Raven}_{i(k)}|\boldsymbol{X}_i\right] &= \beta_0 + \beta_1\ \text{age}_i + \beta_2\ \text{ses}_i + \beta_3\ \text{braven}_i + \beta_4\ \mathbb{1}(\text{male}_i) \\ &\quad + (\beta_5 + \beta_6\ \text{milk}_i + \beta_7\ \text{meat}_i + \beta_8\ \text{energy}_i) \times \text{avg\_time}_k, \end{aligned} \tag{4.25}$$

for $i = 1, \ldots, n$ and $k = 1, 2, 3, 4$. Note that across the time points, the coefficients are shared meaning that we can fit the model using VSPGLM by constraining the components to have the same coefficients using the following syntax

```
% Cognitive model using average time interaction. Note variables are named by their
    round, so they are from 2-5 as 1 is the baseline
formula_raven = "(raven_r2,raven_r3,raven_r4,raven_r5) ~ (age_at_time0, ses, braven,
    male, (avg_time_2 & avg_time_3 & avg_time_4 & avg_time_5), (milk_2 & milk_3 &
    milk_4 & milk_5), (meat_2 & meat_3 & meat_4 & meat_5), (calorie_2 & calorie_3 &
    calorie_4 & calorie_5))";
cognitive_model = fit_vspglm(formula_raven, data, {"id", "id", "id", "id"});
```

Fitting VSPGLM to the data, the results are presented in Table 4.16 where it is compared with the results presented in Han et al. (2014) of the CEL method and GEEs with various correlation structures. It should be noted that as VSPGLM is fit on $n = 469$ instead of $n = 524$, the estimates and standard errors are not directly comparable, but we can still consider which factors the models found to be significant.

| | VSPGLM ($n = 469$) | | | CEL ($n = 524$) | | | GEE.AR ($n = 524$) | | | GEE.UN ($n = 524$) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Est | SE | $p$-value | Est | SE | $p$-value | Est | SE | $p$-value | Est | SE | $p$-value |
| Intercept | 11.88 | 0.992 | $< 0.01$ | 11.53 | 0.955 | $< 0.01$ | 11.67 | 1.055 | $< 0.01$ | 11.47 | 1.061 | $< 0.01$ |
| Age | 0.107 | 0.075 | 0.15 | 0.119 | 0.075 | 0.11 | 0.111 | 0.082 | 0.18 | 0.137 | 0.082 | 0.09 |
| SES | 0.006 | 0.004 | 0.13 | 0.009 | 0.004 | 0.02 | 0.006 | 0.004 | 0.12 | 0.006 | 0.004 | 0.08 |
| Braven | 0.236 | 0.038 | $< 0.01$ | 0.241 | 0.037 | $< 0.01$ | 0.251 | 0.044 | $< 0.01$ | 0.250 | 0.044 | $< 0.01$ |
| Male | 0.613 | 0.178 | $< 0.01$ | 0.511 | 0.178 | $< 0.01$ | 0.636 | 0.178 | $< 0.01$ | 0.596 | 0.179 | $< 0.01$ |
| Avg Time | 0.901 | 0.151 | $< 0.01$ | 0.885 | 0.149 | $< 0.01$ | 1.010 | 0.142 | $< 0.01$ | 0.958 | 0.141 | $< 0.01$ |
| Energy × time | 0.039 | 0.192 | 0.84 | 0.119 | 0.193 | 0.54 | -0.127 | 0.189 | 0.50 | -0.089 | 0.186 | 0.63 |
| Meat × time | 0.511 | 0.209 | 0.01 | 0.538 | 0.199 | 0.01 | 0.354 | 0.203 | 0.08 | 0.392 | 0.201 | 0.05 |
| Milk × time | -0.126 | 0.192 | 0.51 | -0.019 | 0.191 | 0.92 | -0.273 | 0.188 | 0.15 | -0.234 | 0.186 | 0.21 |

**Table 4.16:** Fitted Models on cognitive data with average time interaction. CEL: Conditional Empirical Likelihood, GEE.AR: First order autoregressive GEE, GEE.UN: Unstructured correlation GEE. Note that GEE with compound symmetry is omitted, but can be found in Han et al. (2014)

VSPGLM, CEL and GEE methods all find that cognitive ability significantly improves over time and that this improvement is significantly higher for males. VSPGLM and GEEs don't find that social economic status is a significant factor whereas the CEL method does. Furthermore, both VSPGLM and CEL at a 5% significance level find that only the meat treatment significantly improves cognitive development, aligning with the results found by Whaley et al. (2003). In contrast, GEE methods find no significance for any of the treatments at a 5% level. Using VSPGLM we can predict the correlation between the time points for an average male in each treatment group (averaged for each covariate) which is given in Appendix A.7 Figure A.9. We observe that raven scores which are closer together in time are more correlated and that the scores are all positively correlated with one another. There does seem to be some form of time dependence in the correlation but it isn't particularly an autoregressive relationship and seems more unstructured.

Note that the models above are dependent on the assumption that the time effect at each measurement for each observation can be grouped as having the same effect at the average time. However,

when looking at the time points in which observations are measured, we can see that for each observation there is some variance as to when in the round they are measured. This is visualised in Appendix A.7 Figure A.8 where we see that rounds are a few months long and each observation doesn't have the same difference between measurements. Thus, it may be more appropriate to model the time interaction with the actual time of the measurements of each observation $i$ ($\text{rel\_time}_i$) relative to the baseline, as given in the following marginal mean model

$$\begin{aligned}\mathbb{E}\left[\text{Raven}_{i(k)}|\boldsymbol{X}_i\right] &= \beta_0 + \beta_1\ \text{age}_i + \beta_2\ \text{ses}_i + \beta_3\ \text{braven}_i + \beta_4\ \mathbb{1}(\text{male}_i) \\ &\quad + (\beta_5 + \beta_6\ \text{milk}_i + \beta_7\ \text{meat}_i + \beta_8\ \text{energy}_i) \times \text{rel\_time}_{ik},\end{aligned} \tag{4.26}$$

for $i = 1, \ldots, n$ and $k = 1, 2, 3, 4$. The syntax to fit the model is the same but with the appropriate variable changes, and the fitted VSPGLM to the marginal mean model (4.26) with shared coefficients across components is given in Table 4.17. In this example, there is no explicit justification for needing to adjust the standard errors, but they are included for the sake of interest.

| Coefficient | Estimate | S.E (Adjusted) | T (Adjusted) | $p$ | $p$ Adjusted |
|---|---|---|---|---|---|
| Intercept | 11.89 | 0.993 (1.151) | 11.98 ( 10.32) | $< 2.2 \times 10^{-16}$ | $< 2.2 \times 10^{-16}$ |
| Age | 0.108 | 0.075 (0.089) | 1.433 ( 1.209) | 0.1523 | 0.2274 |
| SES | 0.006 | 0.004 (0.005) | 1.526 ( 1.284) | 0.1278 | 0.1997 |
| Braven | 0.236 | 0.038 (0.044) | 6.211 ( 5.309) | $1.2 \times 10^{-9}$ | $1.7 \times 10^{-7}$ |
| Male | 0.615 | 0.178 (0.208) | 3.469 ( 2.954) | $5.7 \times 10^{-4}$ | 0.0033 |
| Rel_time | 0.946 | 0.153 (0.164) | 6.177 ( 5.775) | $1.4 \times 10^{-9}$ | $1.4 \times 10^{-8}$ |
| Energy × time | -0.016 | 0.197 (0.216) | -0.082 (-0.075) | 0.9345 | 0.9402 |
| Meat × time | 0.450 | 0.206 (0.222) | 2.168 ( 2.026) | 0.0307 | 0.0433 |
| Milk × time | -0.191 | 0.191 (0.212) | -0.998 (-0.900) | 0.3188 | 0.3684 |
| **Log-Likelihood ($\ell$)** | | -2854.4 | | | |

**Table 4.17:** Fitted VSPGLM on Raven scores with relative time interaction

Comparing with the average time interaction model, we observe that most of the coefficient estimates are similar, with notable changes being that the Meat treatment coefficient decreased and Energy treatment has a small negative coefficient. We still find that cognitive ability significantly improves over time and that the improvement is larger for males. We also at a 5% significance level still find that meat is the only treatment to significantly improve cognitive development, even after applying the sandwich correction in the case of misspecification. In Figure 4.11, we visualise the fitted model for an average male student in each of the control groups. Furthermore, the estimated correlation for an average male in each group is the same as the average time interaction model presented in Appendix A.7 Figure A.8.

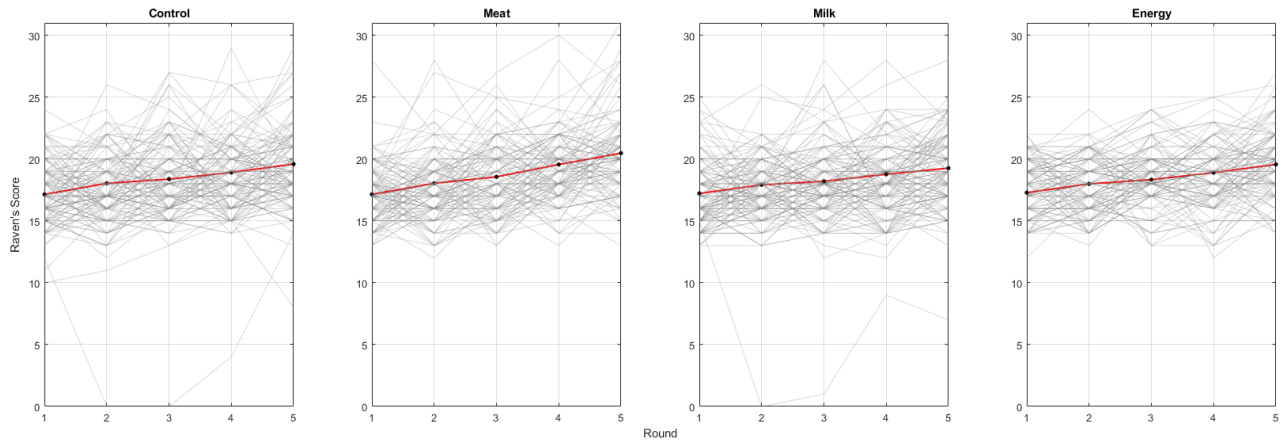


**Figure 4.11:** Fitted VSPGLM for Raven Score using a relative time interaction with treatment

Given that the arithmetic scores were also measured we can extend this analysis to test for whether dietary intervention is associated with improved arithmetic development. Here the same fitted marginal mean models are proposed, but now the response vector is the arithmetic scores of children at the 4 visits after the baseline, as given in (4.27)

$$\begin{aligned}\mathbb{E}\left[\text{Arithmetic}_{i(k)}|\boldsymbol{X}_i\right] &= \beta_0 + \beta_1\ \text{age}_i + \beta_2\ \text{ses}_i + \beta_3\ \text{barithmetic}_i + \beta_4\ \mathbb{1}(\text{male}_i) \\ &\quad + (\beta_5 + \beta_6\ \text{milk}_i + \beta_7\ \text{meat}_i + \beta_8\ \text{energy}_i) \times \text{rel\_time}_{ik}, \end{aligned} \tag{4.27}$$

for $i = 1, \ldots, n$ and $k = 1, 2, 3, 4$. Here the baseline arithmetic score is a covariate in the model rather than the baseline raven score, and we adopt the treatment-relative time interaction covariate model proposed prior. The dataset of observations with complete arithmetic scores was the same as for the raven scores, so fitting VSPGLM to this yields the results presented in Table 4.18 along with sandwich corrections.

| Coefficient | Estimate | S.E (Adjusted) | T (Adjusted) | $p$ | $p$ Adjusted |
|---|---|---|---|---|---|
| Intercept | 3.899 | 0.441 (0.647) | 8.851 ( 6.031) | $< 2.2 \times 10^{-16}$ | $3.4 \times 10^{-9}$ |
| Age | 0.008 | 0.041 (0.059) | 0.201 ( 0.139) | 0.8411 | 0.8899 |
| SES | 0.000 | 0.002 (0.003) | 0.158 ( 0.109) | 0.8741 | 0.9130 |
| Barithmetic | 0.443 | 0.035 (0.054) | 12.63 ( 8.265) | $< 2.2 \times 10^{-16}$ | $1.5 \times 10^{-15}$ |
| Male | -0.039 | 0.097 (0.141) | -0.397 (-0.274) | 0.6914 | 0.7844 |
| Rel_time | 0.816 | 0.071 (0.083) | 11.45 ( 9.792) | $< 2.2 \times 10^{-16}$ | $< 2.2 \times 10^{-16}$ |
| Energy × time | 0.337 | 0.096 (0.115) | 3.493 ( 2.934) | $5.2 \times 10^{-4}$ | 0.003 |
| Meat × time | 0.224 | 0.097 (0.115) | 2.304 ( 1.940) | 0.0217 | 0.053 |
| Milk × time | 0.056 | 0.091 (0.110) | 0.611 ( 0.509) | 0.5417 | 0.611 |
| **Log-Likelihood** ($\ell$) | | -2788.0 | | | |

**Table 4.18:** Fitted VSPGLM on Arithmetic scores with relative time interaction

Similar to the cognitive model, we observe that arithmetic ability improves over time, but interestingly unlike the cognitive ability, there is no significant difference in the arithmetic ability between males and females. Furthermore, at a 5% significance level, both meat and energy treatments were found to improve arithmetic ability in children, which is the same conclusion found by Whaley et al. (2003) using hierarchical linear random effects models. For the analysis, there is no particular justification for the need to adjust the standard errors, but for interest sake, if a sandwich correction is performed then the meat treatment is no longer significant. Similar to the VSPGLM cognitive model, as evident in Appendix A.7 Figure A.10 there is no apparent correlation structure between the components, other than measurements closer to each other in time are more correlated than those that are further apart.

As an interesting extension, given that arithmetic and cognitive ability may be correlated, we can perform a joint longitudinal study between the two sets of responses across the time periods using the proposed VSPGLM. The joint longitudinal study in this case involves combining the marginal mean models (4.26) and (4.27) together given in (4.28) where the coefficients are shared across components associated with the same underlying measurement.

$$\begin{aligned}\mathbb{E}\left[\text{Raven}_{i(k)}|\boldsymbol{X}_i\right] &= \beta_{(1)0} + \beta_{(1)1}\ \text{age}_i + \beta_{(1)2}\ \text{ses}_i + \beta_{(1)3}\ \text{braven}_i + \beta_{(1)4}\ \mathbb{1}(\text{male}_i) \\ &\quad + \left(\beta_{(1)5} + \beta_{(1)6}\ \text{milk}_i + \beta_{(1)7}\ \text{meat}_i + \beta_{(1)8}\ \text{energy}_i\right) \times \text{rel\_time}_{ik}, \\ \mathbb{E}\left[\text{Arithmetic}_{i(k+4)}|\boldsymbol{X}_i\right] &= \beta_{(2)0} + \beta_{(2)1}\ \text{age}_i + \beta_{(2)2}\ \text{ses}_i + \beta_{(2)3}\ \text{barithmetic}_i + \beta_{(2)4}\ \mathbb{1}(\text{male}_i) \\ &\quad + \left(\beta_{(2)5} + \beta_{(2)6}\ \text{milk}_i + \beta_{(2)7}\ \text{meat}_i + \beta_{(2)8}\ \text{energy}_i\right) \times \text{rel\_time}_{ik}\ , \end{aligned} \tag{4.28}$$

for $i = 1, \ldots, n$ and $k = 1, 2, 3, 4$. Fitting VSPGLM with marginal means given by (4.28) done using the following syntax

```
% Joint model with cognitive and arithmetic outcomes
formulas_raven = "(raven_r2,raven_r3,raven_r4,raven_r5) ~ (age_at_time0, ses, raven_r1,
    male, (rn2_year & rn3_year & rn4_year & rn5_year), (milk_2 & milk_3 & milk_4 &
    milk_5), (meat_2 & meat_3 & meat_4 & meat_5), (calorie_2 & calorie_3 & calorie_4 &
    calorie_5))";
formulas_arithmetic = "(arithmetic_r2,arithmetic_r3,arithmetic_r4,arithmetic_r5) ~
    (age_at_time0, ses, arithmetic_r1, male, (rn2_year & rn3_year & rn4_year &
    rn5_year), (milk_2 & milk_3 & milk_4 & milk_5), (meat_2 & meat_3 & meat_4 &
    meat_5), (calorie_2 & calorie_3 & calorie_4 & calorie_5))";
cognitive_model_joint = fit_vspglm([formulas_raven, formulas_arithmetic], data, links);
```

Although the model is computationally expensive to fit, VSPGLM does converge and produces the results presented in Table 4.19.

| Component | Coefficient | Estimate | S.E | T | $p$ |
|---|---|---|---|---|---|
| Raven | Intercept | 13.03 | 0.875 | 14.90 | $< 2.2 \times 10^{-16}$ |
| | Age | 0.049 | 0.071 | 0.684 | 0.4941 |
| | SES | 0.004 | 0.004 | 1.293 | 0.1965 |
| | Braven | 0.208 | 0.033 | 6.292 | $7.3 \times 10^{-10}$ |
| | Male | 0.613 | 0.166 | 3.684 | $2.5 \times 10^{-10}$ |
| | Rel_time | 0.895 | 0.149 | 6.015 | $3.7 \times 10^{-9}$ |
| | Energy × time | -0.003 | 0.190 | -0.019 | 0.9854 |
| | Meat × time | 0.432 | 0.199 | 2.173 | 0.0303 |
| | Milk × time | -0.197 | 0.183 | -1.073 | 0.2837 |
| Arithmetic | Intercept | 4.181 | 0.416 | 10.04 | $< 2.2 \times 10^{-16}$ |
| | Age | 0.030 | 0.039 | 0.767 | 0.4437 |
| | SES | 0.001 | 0.002 | 0.486 | 0.6269 |
| | Barithmetic | 0.374 | 0.028 | 13.21 | $< 2.2 \times 10^{-16}$ |
| | Male | -0.055 | 0.094 | -0.587 | 0.5575 |
| | Rel_time | 0.833 | 0.069 | 12.03 | $< 2.2 \times 10^{-16}$ |
| | Energy × time | 0.313 | 0.094 | 3.347 | $8.8 \times 10^{-4}$ |
| | Meat × time | 0.198 | 0.095 | 2.094 | 0.0369 |
| | Milk × time | 0.034 | 0.088 | 0.389 | 0.6971 |
| **Log-Likelihood ($\ell$)** | | -2776.3 | | | |

**Table 4.19:** VSPGLM fitted model jointly using a relative time interaction with treatment

Performing a joint analysis does lead to slight changes in the coefficients, but they are generally similar to those in the independent model. As the model has become more complex, we expect the slower asymptotic convergence of the standard errors and given standard errors were found to converge from below in simulation studies, we similarly find that they are either the same or slightly smaller than those given in the univariate analysis. Despite this, all factors that are considered significant above are the same as performing a longitudinal analysis on each component independently.

Similar to the prior examples, VSPGLM can be used to predict the underlying correlation structure between all of the responses across the various time points, which is visualised in Figure 4.12 for an average male observation within each treatment group. When considering the correlation between the raven scores, the positive correlation between successive time points now seems constant across the groups, and again measurements that are further away in time are less correlated. Arithmetic scores are found to generally be more positively correlated compared to raven scores but generally have the same structure as the independent model. We also observe a small positive correlation between the raven and arithmetic scores at the various time points, with measurements closer in time being more positively correlated. The correlation of the average observation is taken to be Male, but the outcomes

are the same for an average female observation with each treatment.

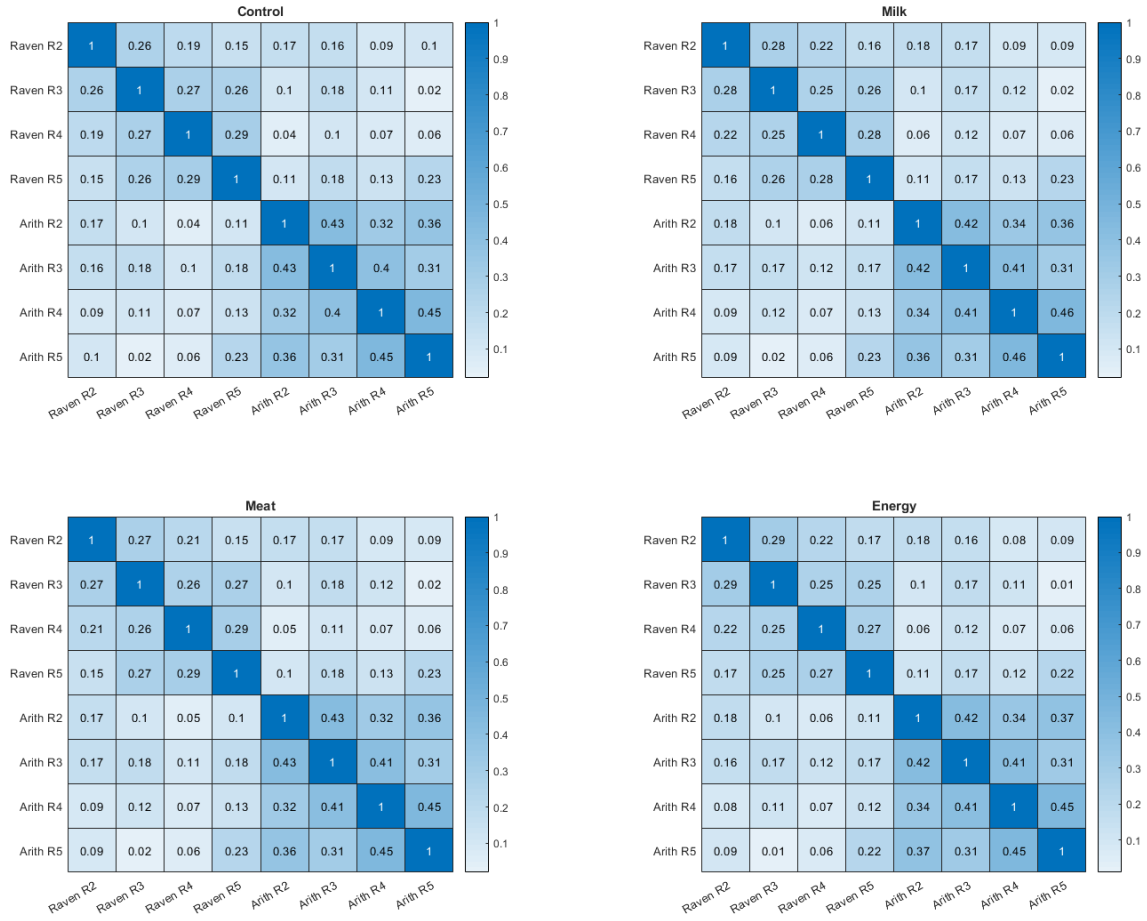


**Figure 4.12:** Estimated Correlation of Raven and Arithmetic Scores at Rounds 2-5 for an average male observation in each treatment group (averaging over the covariates) using a relative time interaction.

# Chapter 5

# Technical Details for Asymptotic Results

## 5.1 Introduction

The following chapter derives the technical details for the Joint Asymptotic Normality of $(\hat{\boldsymbol{\beta}}, \hat{F})$ claimed in Proposition 2.1 and the asymptotic $\chi^2$ convergence of the profile empirical likelihood claimed in Proposition 2.2. The chapter generalizes the derivations published in the supplementary material of Huang (2014) to the Vector Semiparametric Generalized Linear Model (VSPGLM). We will also establish the orthogonality of the parameters $\boldsymbol{\beta}, F$ claimed in Lemma 2.1 and the consistency of $(\hat{\boldsymbol{\beta}}, \hat{F})$ claimed in Lemma 2.2. In situations where the proof is the same as the Semiparametric Generalized Linear Model (SPGLM), the results will be stated and the details can be found in Huang (2011) and Huang (2014). Across the chapter, we assume that $F$ and $F^*$ unless stated otherwise lie within the class of distributions which are considered to be correctly specified for VSPGLM.

## 5.2 Technical Details of Lemmas

### 5.2.1 Lemma 2.1: Orthogonality Between Parameters

To show Lemma 2.1, we follow the argument presented by Huang and Rathouz (2017) and make appropriate changes for the case of multivariate responses. Orthogonality in semiparametric models between a finite-dimensional parameter $\boldsymbol{\beta}$ and an infinite-dimensional parameter $F$ is characterized by the score function for $\boldsymbol{\beta}$ being orthogonal to the nuisance tangent space for $F$. To do so, let us consider semiparametric restricted moment models as seen in Tsiatis (2006), which is characterized by

$$\boldsymbol{Y} = \boldsymbol{\mu}(\boldsymbol{X}, \boldsymbol{\beta}) + \boldsymbol{\epsilon} \,,$$

where $\boldsymbol{\epsilon}$ has the moment condition $\mathbb{E}\left[\boldsymbol{\epsilon}|\boldsymbol{X}\right] = 0$. Thus, by setting $\boldsymbol{\mu}(\boldsymbol{X}, \boldsymbol{\beta}) = \boldsymbol{\mu}(\boldsymbol{X}^T\boldsymbol{\beta})$ and noting the mean constraint $\mathbb{E}\left[\boldsymbol{Y} - \boldsymbol{\mu}|\boldsymbol{X}\right] = 0$, by construction the vector semiparametric exponential tilt regression model is a special case of the semiparametric restricted moment model. As a result, the nuisance tangent space for $F$ in the exponential tilt regression model is necessarily a subspace of the nuisance tangent space for the restricted moment model.

Following the derivations in Tsiatis (2006), pp 73-83, the nuisance tangent space $\Lambda$ for the semiparametric restricted moment model is given by

$$\Lambda = \{g^{q\times 1}(\boldsymbol{X}, \boldsymbol{Y}) : \mathbb{E}\left[h(\boldsymbol{X}, \boldsymbol{Y})(\boldsymbol{Y} - \boldsymbol{\mu})^T|\boldsymbol{X}\right] = \boldsymbol{0}^{q\times K}\} \,. \tag{5.1}$$

Then, by equation (4.44) in Tsiatis (2006), for any $g = g(\boldsymbol{X}, \boldsymbol{Y})$ in some Hilbert space $\mathcal{H}$ with $\mathbb{E}[g^T g] < \infty$, the projection of $g$ onto the nuisance tangent space is given by

$$\Pi(g|\Lambda) = g - \mathbb{E}[g(\boldsymbol{Y} - \boldsymbol{\mu})^T|\boldsymbol{X}]\ \Sigma_{\boldsymbol{Y}}^{-1}\ (\boldsymbol{Y} - \boldsymbol{\mu})\ , \tag{5.2}$$

where $\Sigma_{\boldsymbol{Y}}^{-1} = \mathbb{E}_{\boldsymbol{\beta},F}\left[(\boldsymbol{Y} - \boldsymbol{\mu})\,(\boldsymbol{Y} - \boldsymbol{\mu})^T\,|\boldsymbol{X}\right]$.

Applying the projection operator (5.2) to the score function for $\boldsymbol{\beta}$, we have for all $(\boldsymbol{\beta}, F)$ that

$$\begin{aligned}
\Pi\left(S_{\boldsymbol{\beta},F}(\boldsymbol{X},\boldsymbol{Y})|\Lambda\right) &= S_{\boldsymbol{\beta},F}(\boldsymbol{X},\boldsymbol{Y}) - \mathbb{E}\left[S_{\boldsymbol{\beta},F}(\boldsymbol{X},\boldsymbol{Y})(\boldsymbol{Y}-\boldsymbol{\mu})^T|\boldsymbol{X}\right]\ \Sigma_{\boldsymbol{Y}}^{-1}\ (\boldsymbol{Y}-\boldsymbol{\mu}) \\
&= D(\boldsymbol{X};\boldsymbol{\beta})\Sigma_{\boldsymbol{Y}}^{-1}\ (\boldsymbol{Y}-\boldsymbol{\mu}) - D(\boldsymbol{X};\boldsymbol{\beta})\Sigma_{\boldsymbol{Y}}^{-1}\mathbb{E}\left[(\boldsymbol{Y}-\boldsymbol{\mu})(\boldsymbol{Y}-\boldsymbol{\mu})^T|\boldsymbol{X}\right]\Sigma_{\boldsymbol{Y}}^{-1}\ (\boldsymbol{Y}-\boldsymbol{\mu}) \\
&= D(\boldsymbol{X};\boldsymbol{\beta})\Sigma_{\boldsymbol{Y}}^{-1}\ (\boldsymbol{Y}-\boldsymbol{\mu}) - D(\boldsymbol{X};\boldsymbol{\beta})\Sigma_{\boldsymbol{Y}}^{-1}\ (\boldsymbol{Y}-\boldsymbol{\mu}) \\
&= 0\ .
\end{aligned}$$

Therefore, the score function for $\boldsymbol{\beta}$ is orthogonal to the nuisance tangent space for the restricted moment model. As a result, the score function for $\boldsymbol{\beta}$ is necessarily orthogonal to the nuisance tangent space for the exponential tilt regression model which proves the claim of orthogonality between $\boldsymbol{\beta}$ and $F$ given in Lemma 2.1.

Intuitively, as the projection of the score function for $\boldsymbol{\beta}$ is a measure of the loss of information about $\boldsymbol{\beta}$ due to the presence of the nuisance parameter $F$, this shows that asymptotically there is no information lost about $\boldsymbol{\beta}$ by having to jointly estimate $F$. Furthermore, as a result the efficient score for $\boldsymbol{\beta}$ for all $(\boldsymbol{\beta}, F^*)$ is given by the naive score for $\boldsymbol{\beta}$,

$$\begin{aligned}
\tilde{S}_{\boldsymbol{\beta},F}(\boldsymbol{X},\boldsymbol{Y}) &= S_{\boldsymbol{\beta},F}(\boldsymbol{X},\boldsymbol{Y}) - \Pi(S_{\boldsymbol{\beta},F}(\boldsymbol{X},\boldsymbol{Y})|\Lambda) \\
\tilde{S}_{\boldsymbol{\beta},F}(\boldsymbol{X},\boldsymbol{Y}) &= D(\boldsymbol{X};\boldsymbol{\beta})\Sigma_{\boldsymbol{Y}}^{-1}(\boldsymbol{X};\boldsymbol{\beta},F)\,(\boldsymbol{Y}-\boldsymbol{\mu})\ ,
\end{aligned} \tag{5.3}$$

showing that the score equation (2.22) is efficient. As a result, we have that the optimal estimator for $\boldsymbol{\beta}$ is obtained by solving the estimating equations

$$\sum_{i=1}^{n} D(\boldsymbol{X}_i;\boldsymbol{\beta})\Sigma_{\boldsymbol{Y}}^{-1}(\boldsymbol{X}_i;\boldsymbol{\beta},F)\left(\boldsymbol{Y}_i - \boldsymbol{\mu}(\boldsymbol{X}_i^T\boldsymbol{\beta})\right)\ ,$$

which shows that the maximum empirical likelihood estimator $\hat{\boldsymbol{\beta}}$ by construction is optimal. Furthermore, as the semiparametric efficiency bound is equal to the inverse covariance matrix of the efficient score (Theorem 4.1 Tsiatis 2006), then we have that the inverse covariance matrix of $S_{\boldsymbol{\beta}^*,F^*}(\boldsymbol{X},\boldsymbol{Y})$, $W_1$ defined in Proposition 2.1, also attains the semiparametric efficiency bound. The application of the results above comes from the theory developed on pp 87-93 of Tsiatis (2006) which ensures these concepts hold in arbitrary multivariate response spaces and that a parametric submodel for arbitrary restricted moment models exists.

### 5.2.2 Lemma 2.2: Consistency of MELE

To establish the consistency of $(\hat{\boldsymbol{\beta}}, \hat{F})$ in Lemma 2.2, as the parameter $\boldsymbol{\beta}$ are stacked in a single vector similar to the univariate case, the argument will be similar to that presented in Huang (2014). For the sake of completeness, we will step through the working, following the steps of the consistency proof in Murphy and Van Der Vaart (2001). This is done by first proving Lemma 2.2 under the additional assumption that the likelihood is maximized with respect to $\boldsymbol{\beta}$ over a compact subset $\mathcal{B} \subset \mathbb{R}^q$. Then at the end of the proof the assumption, adhering to the assumption that the response space $\mathcal{Y}$ is compact in $\mathbb{R}^K$.

The proof similar to Huang (2014) replies on the existence of a consistent estimate of $F$ of the same form as $\hat{F}$. The VSPGLM analogous estimate which is noncomputable is given by

$$\tilde{F} = \sum_{i=1}^{n} \left( \frac{\mathbb{1}(\boldsymbol{Y}_i \leq \boldsymbol{y})}{\sum_{j=1}^{n} \exp\left(b_i^* + \boldsymbol{\theta}_i^{*T} \boldsymbol{Y}_j\right)} \right) , \tag{5.4}$$

where $(b_i^*, \boldsymbol{\theta}_i^*) = (b(\boldsymbol{X}_i; \boldsymbol{\beta}^*, F^*), \boldsymbol{\theta}(\boldsymbol{X}_i; \boldsymbol{\beta}^*, F^*))$ are the true normalization and tilt values for $i = 1, \ldots, n$. Note that the log-likelihood function is invariant to the choice of tilt or normalisation, so no tilting or renormalising is required for $\tilde{F}$. Here $\tilde{F}$ is unbiased for $F^*$ and is also the exponential tilt analog of the empirical distribution for iid data by the defined constraints, implying it is consistent. However, it can be shown more formally by applying the following theorem from Wellner (1981) for independent but non-identically distributed random variables.

**Theorem 5.1** *Let $(\mathcal{Z}, d)$ be a separable metric space, and $\mathcal{P}(\mathcal{Z})$ be the set of all Borel probability measures on $\mathcal{Z}$. Let $Z_1, \ldots, Z_n$ be independent random variables on $\mathcal{Z}$ with measures $P_1, \ldots, P_n$. Furthermore, define an average measure*

$$\bar{P}_n = \frac{1}{n} \sum_{i=1}^{n} P_i \tag{5.5}$$

*Let $\rho_1$ be the Prohorov metric on $\mathcal{P}(\mathcal{Z})$ for $P, Q \in \mathcal{P}(\mathcal{Z})$,*

$$\rho_1(P, Q) = \inf \{\epsilon > 0 : P(A) \leq \epsilon + Q(A^\epsilon) \text{ for all Borel sets } A\}$$

*where $A^\epsilon \equiv \{y \in \mathcal{Z} : d(x, y) < \epsilon \text{ for some } x \in A\}$. Let $\rho_2$ denote the dual-bounded-Lipschitz metrics on $\mathcal{P}(\mathcal{Z})$,*

$$\rho_2(P, Q) = \sup \left\{ \int_{\mathcal{Z}} f \mathrm{d}(P - Q) : ||f||_\infty + ||f||_L \leq 1 \right\}$$

*where $||f||_\infty = \sup_x |f(x)|$, $||f||_L = \sup_{x \neq y} |f(x) - f(y)|/d(x, y)$.*

*If $\{\bar{P}_n\}_{n \geq 1}$ is tight, then as $n \to \infty$,*

$$\rho_1(\mathbb{P}_n, \bar{P}_n) \xrightarrow{a.s} 0 \tag{5.6}$$

$$\rho_2(\mathbb{P}_n, \bar{P}_n) \xrightarrow{a.s} 0 \tag{5.7}$$

The tightness of the sequence of average measures holds as we assume $\mathcal{Y}$ is compact in $\mathbb{R}^K$, implying that $\tilde{F}$ is consistent. However, as the true values $b^*$ and $\boldsymbol{\theta}^*$ are unknown, $\tilde{F}$ is non-computable, but it's used to sandwich the estimator $\hat{F}$ to the true value $F^*$ which we will formalise below.

Following the consistency proof from Murphy and Van Der Vaart (2001), define the function

$$m_{(\beta, F_1), F_2}(\boldsymbol{x}, \boldsymbol{y}) = \ell(\boldsymbol{\beta}, F_1 | \boldsymbol{x}, \boldsymbol{y}) - \ell(\boldsymbol{\beta}^*, F_2 | \boldsymbol{x}, \boldsymbol{y}) \tag{5.8}$$

where $\ell(\boldsymbol{\beta}, F | \boldsymbol{x}, \boldsymbol{y})$ is the empirical log likelihood. Then, by the definition of the MELE $(\hat{\boldsymbol{\beta}}, \hat{F})$, we have that

$$\mathbb{P}_n m_{(\hat{\boldsymbol{\beta}}, \hat{F})\tilde{F}} = \frac{1}{n} \sum_{i=1}^{n} \ell(\hat{\boldsymbol{\beta}}, \hat{F} | \boldsymbol{x}_i, \boldsymbol{y}_i) - \ell(\boldsymbol{\beta}^*, \tilde{F} | \boldsymbol{x}_i, \boldsymbol{y}_i) \geq 0 \tag{5.9}$$

as $\ell(\hat{\boldsymbol{\beta}}, \hat{F} | \boldsymbol{x}_i, \boldsymbol{y}_i) \geq \ell(\boldsymbol{\beta}^*, \tilde{F} | \boldsymbol{x}_i, \boldsymbol{y}_i)$ for $i = 1, 2, \ldots, n$. The above holds under the empirical measure, but the expected log-likelihood is maximised at the true value $(\boldsymbol{\beta}^*, F^*)$ which implies that for every $(\boldsymbol{\beta}, F) \neq (\boldsymbol{\beta}^*, F^*)$

$$P^* m_{(\boldsymbol{\beta}, F) F^*} < 0. \tag{5.10}$$

The functions $m_{(\boldsymbol{\beta},F_1)F_2}(\boldsymbol{x},\boldsymbol{y})$ are continuous in $(\boldsymbol{\beta},F_1)$ and $F_2$ for every $(\boldsymbol{x},y)$ where continuity in $F_1,F_2$ is with respect to the weak topology defined on $\mathcal{Y}$. Following Murphy and Van Der Vaart (2001), for every fixed $(\boldsymbol{\beta},F_1)$ and sufficiently small neighbourhoods $U$ of $(\boldsymbol{\beta},F_1)$ and $V_{(\boldsymbol{\beta},F_1)}$ of $F^*$, we have that

$$P^* \sup_{(\boldsymbol{\beta}',F_1')\in U_{(\boldsymbol{\beta},F_1)},F_2\in V_{(\boldsymbol{\beta},F_1)}} m_{(\boldsymbol{\beta}',F_1'),F_2)} < \infty. \tag{5.11}$$

Similar to the consistency proof in Wald (1949), the above allows for the monotone convergence theorem to be invoked along sequences of neighbourhoods shrinking to $(\boldsymbol{\beta},F_1)$ and $F^*$. This constructs for every $(\boldsymbol{\beta},F_1)\neq(\boldsymbol{\beta^*},F^*)$ a neighbourhood $U_{(\boldsymbol{\beta},F_1)}$ and $V_{(\boldsymbol{\beta},F_1)}$ of $F^*$ such that

$$P^* \sup_{(\boldsymbol{\beta}',F_1')\in U_{(\boldsymbol{\beta},F_1)},F_2\in V_{(\boldsymbol{\beta},F_1)}} m_{(\boldsymbol{\beta}',F_1'),F_2)} < 0. \tag{5.12}$$

Define the distance function on the product space $\mathcal{B}\times\mathcal{F}_{\boldsymbol{\mu}}$ as $||(\boldsymbol{\beta}_1,F_1)-(\boldsymbol{\beta}_2,F_2)|| = ||\boldsymbol{\beta}_1-\boldsymbol{\beta}_2|| + ||F_1-F_2||_{\mathcal{H}_L}$. For a fixed $\epsilon>0$, the event $\{||(\hat{\boldsymbol{\beta}},\hat{F})-(\boldsymbol{\beta^*},F^*)||\geq\epsilon\}$ implies that the supremum of $\mathbb{P}_n m_{(\beta,F_1),\tilde{F}}$ over $\{||(\hat{\boldsymbol{\beta}},F_1)-(\boldsymbol{\beta^*},F^*)||\geq\epsilon\}$ must be greater than or equal to 0 by (5.9). Since the space of all distribution functions on the space $\mathcal{Y}$ is bounded by assumption and is thus compact with respect to the weak topology. This means that the set $\{(\boldsymbol{\beta},F):||(\boldsymbol{\beta},F)-(\boldsymbol{\beta^*},F^*)||\geq\epsilon\}$ is also compact and covered by finitely many neighbourhoods $U_{(\boldsymbol{\beta},F_1)}$. Here we index the finite number of neighbourhoods by $U_j = U_{(\boldsymbol{\beta}_j,F_{1,j})}$, $j=1,\ldots,J$ and denote the corresponding neighbourhoods $V_{(\boldsymbol{\beta}_j,F_{1,j})}$ of $F^*$ by $V_j$, where $V:=\cap_{j=1}^J V_j$.

As a result,

$$\begin{aligned}
P^*\left(\left\{||(\hat{\boldsymbol{\beta}},\hat{F})-(\boldsymbol{\beta^*},F^*)||\geq\epsilon\right\}\right) &\leq \sum_{j=1}^J P^*\left(\mathbb{P}_n \sup_{(\boldsymbol{\beta},F_1)\in U_j} m_{(\boldsymbol{\beta},F_1),\tilde{F}}\geq 0\right)\\
&= \sum_{j=1}^J P^*\left(\mathbb{P}_n \sup_{(\boldsymbol{\beta},F_1)\in U_j} m_{(\boldsymbol{\beta},F_1),\tilde{F}}\geq 0, \tilde{F}\in V_j\right)\\
&+ \sum_{j=1}^J P^*\left(\mathbb{P}_n \sup_{(\boldsymbol{\beta},F_1)\in U_j} m_{(\boldsymbol{\beta},F_1),\tilde{F}}\geq 0, \tilde{F}\notin V_j\right)\\
&\leq \sum_{j=1}^J P^*\left(\mathbb{P}_n \sup_{(\boldsymbol{\beta},F_1)\in U_j,F_2\in V_j} m_{(\boldsymbol{\beta},F_1),F_2}\geq 0\right)\\
&+ JP^*(\tilde{F}\notin V)
\end{aligned}$$

Here each of the probabilities in the summation on the right converges to 0 as the variables inside of the probabilities converge to negative constants by the law of large numbers. Furthermore, the last term on the right-hand side converges to 0 by the consistency of the estimator $\tilde{F}$.

The complete the proof, we relax the assumption that $\boldsymbol{\beta}$ lies in a compact set $\mathcal{B}$. To do this, we compactify $\mathbb{R}^q$ to it's one-point compactification $\bar{\mathbb{R}}^q = \mathbb{R}^q\cup\{\infty\}$ and extend the functions $m_{(\boldsymbol{\beta},F_1),F_2}$ onto $\bar{\mathbb{R}}^q$ by setting it to $-\infty$ when at least one component of $\boldsymbol{\beta}$ is infinite. This suffices if the true value $\boldsymbol{\beta^*}$ is bounded away from $\infty$.

## 5.3 Technical Details of Proposition 2.1: Joint Asymptotic Normality

To establish the joint asymptotic distribution of $(\hat{\boldsymbol{\beta}},\hat{F})$, we will use the following theorem from Van Der Vaart and Wellner (1996). For notation, let $\boldsymbol{\psi}=(\boldsymbol{\beta},F)$, $\boldsymbol{\psi}^*=(\boldsymbol{\beta^*},F^*)$ and $\hat{\boldsymbol{\psi}}_n=(\hat{\boldsymbol{\beta}},\hat{F})$

**Theorem 5.2** *Suppose that $\Psi_n$ and $\Psi$ are random maps and deterministic maps respectively. Suppose that $\boldsymbol{\psi}^*$ and $\hat{\boldsymbol{\psi}}_n$ satisfy $\Psi(\boldsymbol{\psi}^*) = 0$ and $\Psi_n(\hat{\boldsymbol{\psi}}_n) = 0$. Furthermore, suppose that*

$$\sqrt{n}\,(\Psi_n - \Psi)(\boldsymbol{\psi}^*) \rightsquigarrow \boldsymbol{Z} \tag{5.13}$$

$$||\sqrt{n}\,(\Psi_n - \Psi)(\hat{\boldsymbol{\psi}}_n) - \sqrt{n}\,(\Psi_n - \Psi)(\boldsymbol{\psi}^*)|| = o_p\left(1 + \sqrt{n}\,||\hat{\boldsymbol{\psi}}_n - \boldsymbol{\psi}^*||\right) \tag{5.14}$$

$$||\Psi(\boldsymbol{\psi}) - \Psi(\boldsymbol{\psi}^*) - \dot{\Psi}_{\boldsymbol{\psi}^*}(\boldsymbol{\psi} - \boldsymbol{\psi}^*)|| = o\left(||\boldsymbol{\psi} - \boldsymbol{\psi}^*||\right), \text{ as } \boldsymbol{\psi} \to \boldsymbol{\psi}^* \tag{5.15}$$

$$\hat{\boldsymbol{\psi}}_n \xrightarrow{\mathbb{P}} \boldsymbol{\psi}^* \text{ and } \Psi_n(\hat{\boldsymbol{\psi}}_n) = \Psi(\boldsymbol{\psi}^*) + o_p(n^{-\frac{1}{2}}) \tag{5.16}$$

*for a Gaussian random process $\boldsymbol{Z}$, and a linear, one-to-one map $\dot{\Psi}$ that is continuously-invertible and onto its range. Then,*

$$-\sqrt{n}\,\dot{\Psi}_{\psi^*}(\hat{\psi}_n - \psi^*) = \sqrt{n}\,(\Psi_n - \Psi)(\psi^*) + o_p(1)\,. \tag{5.17}$$

*Consequently, the sequence*

$$\sqrt{n}\,\left(\hat{\boldsymbol{\psi}}_n - \boldsymbol{\psi}^*\right) \rightsquigarrow -\dot{\Psi}^{-1}_{\psi^*}\boldsymbol{Z}\,, \tag{5.18}$$

*where $\dot{\Psi}^{-1}_{\boldsymbol{\psi}^*}$ is the inverse of $\dot{\Psi}$ evaluated at $\boldsymbol{\psi}^*$.*

We have that $(\hat{\boldsymbol{\beta}}, \hat{F})$ is consistent, so it remains to show the first three conditions. Condition (5.13) is regarding showing the score functions are Donsker Classes. Condition (5.14) is a stochastic approximation to the score equations, where the score function at the estimated parameters should be sufficiently close to the score functions evaluated at the true parameters. Condition (5.15) states that the map $\Psi$ is *Fréchet differentiable* at the true parameter.

We can verify conditions (5.13) and (5.14) with the following sufficient conditions, where the sufficiency is established in Lemma 3.3.5 of Van Der Vaart and Wellner (1996).

$$\{S_{\boldsymbol{\beta},F}, A_{\boldsymbol{\beta},F}h : h \in \mathcal{H}_L, ||\boldsymbol{\beta} - \boldsymbol{\beta}^*, F - F^*|| < \delta\} \tag{5.19}$$

$$\text{is } P^*\text{-Donsker for some } \delta > 0,$$

$$\sup_{h \in \mathcal{H}_L} P^*(A_{\boldsymbol{\beta},F} - A_{\boldsymbol{\beta^*},F^*}h)^2 \to 0, \text{ as } \boldsymbol{\beta} \to \boldsymbol{\beta^*}, F \to F^*\,, \tag{5.20}$$

$$P^*(S_{\boldsymbol{\beta},F} - S_{\boldsymbol{\beta^*},F^*})^2 \to 0, \text{ as } \boldsymbol{\beta} \to \boldsymbol{\beta^*}, F \to F^*\,, \tag{5.21}$$

and $(\hat{\boldsymbol{\beta}}, \hat{F})$ converges in probability to $(\boldsymbol{\beta^*}, F^*)$.

### 5.3.1 Showing Sufficient Conditions

#### Showing Score Equation for $\boldsymbol{\beta}$ is Donsker

Firstly, let us show condition (5.19), where we let $\delta = \min(\delta_1, \delta_2)$ from Assumption (A.2.2.a) and (A.2.2.b). Note that Donsker classes for vector-valued functions are defined by Van Der Vaart (1998) as being Donsker in each of its co-ordinate classes. The way in which we will show a class of score functions is Donsker is by noting a class of vector-valued Lipschitz functions is a Donsker class, which is used as a part of Theorem 5.21, and seen in Example 19.7 of Van Der Vaart (1998). For matrix norms, we use the Frobenius norm $||\cdot||_F$ which is submultiplicative and consistent with the Euclidean norm. This is used as $||.||_2$ induces a Euclidean topology, and the product topology on $K^2$ copies of $\mathbb{R}$ is equal to the Euclidean topology on $\mathbb{R}^{K^2} \cong \mathbb{R}^{K \times K}$.

Our score equation for $\boldsymbol{\beta}$ given below is a product of four terms, which we will show are Donsker classes.

$$S_{\boldsymbol{\beta},F}(\boldsymbol{X},\boldsymbol{Y}) = \text{Diag}(\boldsymbol{X}')\text{Diag}\left(\boldsymbol{\mu}'(\boldsymbol{X}^T\boldsymbol{\beta})\right)\Sigma_{\boldsymbol{Y}}^{-1}\left(\boldsymbol{Y}-\boldsymbol{\mu}\right)$$

The first term $\text{Diag}(\boldsymbol{X}')$ is trivially Donsker as it is a class of only one function. This can be seen further as it is trivially Donsker in its components.

The second term $(\boldsymbol{Y}-\boldsymbol{\mu})$ is continuous and continuously differentiable in $\boldsymbol{\beta}$ over $\{\boldsymbol{\beta}\in\mathbb{R}^q : ||\boldsymbol{\beta}-\boldsymbol{\beta}^*||\leq\delta\}$ by (A.2.2). Let us fix some $\boldsymbol{x}\in\mathcal{X}$ and suppose $\boldsymbol{\beta},\boldsymbol{\gamma}\in\{\boldsymbol{\beta}\in\mathbb{R}^q : ||\boldsymbol{\beta}-\boldsymbol{\beta}^*||\leq\delta\}$.
Therefore,

$$(\boldsymbol{Y}-\boldsymbol{\mu}(\boldsymbol{x}^T\boldsymbol{\beta}))-(\boldsymbol{Y}-\boldsymbol{\mu}(\boldsymbol{x}^T\boldsymbol{\gamma})) = \boldsymbol{\mu}(\boldsymbol{x}^T\boldsymbol{\beta})-\boldsymbol{\mu}(\boldsymbol{x}^T\boldsymbol{\gamma})\ .$$

It's equivalent to consider applying Taylor's Theorem either to the components or using multivariate Taylor's Theorem. By Multivariate Taylor's Theorem, there exists some $\tilde{\boldsymbol{\beta}}$ on the line segment from $\boldsymbol{\beta}$ to $\boldsymbol{\gamma}$, where $\tilde{\boldsymbol{\beta}}\in\{\boldsymbol{\beta}\in\mathbb{R}^q : ||\boldsymbol{\beta}-\boldsymbol{\beta}^*||\leq\delta\}$.

$$\begin{aligned}\boldsymbol{\mu}(\boldsymbol{x}^T\boldsymbol{\beta})-\boldsymbol{\mu}(\boldsymbol{x}^T\boldsymbol{\gamma}) &= \left(\frac{\partial\boldsymbol{\mu}(\boldsymbol{x}^T\tilde{\boldsymbol{\beta}})}{\partial\boldsymbol{\beta}}\right)^T(\boldsymbol{\beta}-\boldsymbol{\gamma})\\ &= D(\boldsymbol{X};\tilde{\boldsymbol{\beta}})^T(\boldsymbol{\beta}-\boldsymbol{\gamma})\ .\end{aligned}$$

Thus, noting that for some vector $\boldsymbol{a}$ that $||\text{Diag}\,(\boldsymbol{a})\,||_F = ||\boldsymbol{a}||_2$,

$$\begin{aligned}||\boldsymbol{\mu}(\boldsymbol{x}^T\boldsymbol{\beta})-\boldsymbol{\mu}(\boldsymbol{x}^T\boldsymbol{\gamma})||_2 &\leq \left|\left|\text{Diag}\left(\boldsymbol{\mu}'(\boldsymbol{x}^T\tilde{\boldsymbol{\beta}})\right)\right|\right|_F ||\text{Diag}\left((\boldsymbol{x}')^T\right)||_F||\boldsymbol{\beta}-\boldsymbol{\gamma}||_2\\ &= ||\boldsymbol{\mu}'(\boldsymbol{x}^T\tilde{\boldsymbol{\beta}})||_2\ ||\boldsymbol{x}'||_2\ ||(\boldsymbol{\beta}-\boldsymbol{\gamma})||_2\ .\end{aligned}$$

Assuming $k'$ is the associated index of $\boldsymbol{\beta}$ for component $k$, by Assumption A.2.2.a, as $||\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^*||\leq\delta$,

$$||\boldsymbol{\mu}'(\boldsymbol{x}^T\tilde{\boldsymbol{\beta}})||_2 = \sqrt{\sum_{k=1}^K\left(\mu'_{(k)}(\boldsymbol{x}_{(k)}^T\tilde{\boldsymbol{\beta}}_{(k')})\right)^2} \leq \sum_{k=1}^K\sqrt{\left(\mu'_{(k)}(\boldsymbol{x}_{(k)}^T\tilde{\boldsymbol{\beta}}_{(k')})\right)^2} \leq \sum_{k=1}^K M_{1k} \leq KM_1\ .$$

Furthermore, $||\boldsymbol{x}'||_2$ is bounded because $\boldsymbol{x}'\in\mathbb{R}^Q$ is a finite-dimensional vector where each component is bounded by Assumption A.2.1,

$$||\boldsymbol{\mu}(\boldsymbol{x}^T\boldsymbol{\beta})-\boldsymbol{\mu}(\boldsymbol{x}^T\boldsymbol{\gamma})||_2 \leq KM_1\ ||\boldsymbol{x}'||_2\ ||(\boldsymbol{\beta}-\boldsymbol{\gamma})||_2\ , \tag{5.22}$$

which implies that the class of functions $\{\boldsymbol{y}-\boldsymbol{\mu}(\boldsymbol{x}^T\boldsymbol{\beta}) : \boldsymbol{y}\in\mathcal{Y},\boldsymbol{x}\in\mathcal{X},||\boldsymbol{\beta}-\boldsymbol{\beta}^*||\leq\delta\}$ is Lipschitz in $\boldsymbol{\beta}$ and hence is a Donsker class.

The third term $\text{Diag}\left(\boldsymbol{\mu}'(\boldsymbol{X}^T\boldsymbol{\beta})\right)$ has each of its components continuous and continuously differentiable in $\boldsymbol{\beta}$ on $\{\boldsymbol{\beta}\in\mathbb{R}^Q : ||\boldsymbol{\beta}-\boldsymbol{\beta}^*||\leq\delta\}$. Fixing some $\boldsymbol{x}\in\mathcal{X}$ and suppose $\boldsymbol{\beta},\boldsymbol{\gamma}\in\{\boldsymbol{\beta}\in\mathbb{R}^q : ||\boldsymbol{\beta}-\boldsymbol{\beta}^*||,\leq\delta\}$. By Taylor's Theorem, there exists some $\tilde{\boldsymbol{\beta}}$ on the line segment from $\boldsymbol{\beta}$ to $\boldsymbol{\gamma}$, where $\tilde{\boldsymbol{\beta}}\in\{\boldsymbol{\beta}\in\mathbb{R}^q : ||\boldsymbol{\beta}-\boldsymbol{\beta}^*||\leq\delta\}$ and

$$\boldsymbol{\mu}'(\boldsymbol{x}^T\boldsymbol{\beta})-\boldsymbol{\mu}'(\boldsymbol{x}^T\boldsymbol{\gamma}) = \left(\frac{\partial\boldsymbol{\mu}''(\boldsymbol{x}^T\tilde{\boldsymbol{\beta}})}{\partial\boldsymbol{\beta}}\right)^T(\boldsymbol{\beta}-\boldsymbol{\gamma})\ .$$

Noting by a similar calculation to find $D(\boldsymbol{X};\boldsymbol{\beta})$, it can be shown that

$$\frac{\partial\boldsymbol{\mu}''}{\partial\boldsymbol{\beta}} = \text{Diag}\left(\boldsymbol{X}'\right)\text{Diag}\left(\boldsymbol{\mu}''(\boldsymbol{X}^T\boldsymbol{\beta})\right)\ .$$

Subbing this in, we have that

$$\begin{aligned}\left|\left|\text{Diag}\left(\boldsymbol{\mu}'(\boldsymbol{x}^T\boldsymbol{\beta})\right) - \text{Diag}\left(\boldsymbol{\mu}'(\boldsymbol{x}^T\boldsymbol{\gamma})\right)\right|\right|_F &= ||\boldsymbol{\mu}'(\boldsymbol{x}^T\boldsymbol{\beta}) - \boldsymbol{\mu}'(\boldsymbol{x}^T\boldsymbol{\gamma})||_2 \\ &\leq \left|\left|\text{Diag}\left(\boldsymbol{\mu}''(\boldsymbol{x}^T\tilde{\boldsymbol{\beta}})\right)\right|\right|_F ||\text{Diag}\left((x')^T\right)||_F ||\boldsymbol{\beta}-\boldsymbol{\gamma}||_2 \\ &= ||\boldsymbol{\mu}''(\boldsymbol{x}^T\boldsymbol{\beta})||_2 \, ||\boldsymbol{x}'||_2 \, ||(\boldsymbol{\beta}-\boldsymbol{\gamma})||_2 \, .\end{aligned}$$

Assuming that the index of $\mu_{(k)}$'s corresponding mean model is $k'$, by Assumption A.2.2.b, as $||\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*|| \leq \delta$,

$$||\boldsymbol{\mu}''(\boldsymbol{x}^T\tilde{\boldsymbol{\beta}})||_2 = \sqrt{\sum_{k=1}^{K}\left(\mu''_{(k)}(\boldsymbol{x}_{(k)}^T\tilde{\boldsymbol{\beta}}_{(k')})\right)^2} \leq \sum_{k=1}^{K}\sqrt{\left(\mu''_{(i)}(\boldsymbol{x}_{(k)}^T\tilde{\boldsymbol{\beta}}_{(k')})\right)^2} \leq \sum_{k=1}^{K} M_{2k} \leq KM_2$$

Thus,

$$\left|\left|\text{Diag}\left(\boldsymbol{\mu}'(\boldsymbol{x}^T\boldsymbol{\beta})\right) - \text{Diag}\left(\boldsymbol{\mu}'(\boldsymbol{x}^T\boldsymbol{\gamma})\right)\right|\right|_F \leq KM_2||\boldsymbol{x}'||_2||(\boldsymbol{\beta}-\boldsymbol{\gamma})||_2 \, . \tag{5.23}$$

Which implies that $\{\text{Diag}\left(\boldsymbol{\mu}'(\boldsymbol{x}^T\boldsymbol{\beta})\right) : \boldsymbol{x} \in \mathcal{X}, ||\boldsymbol{\beta}-\boldsymbol{\beta}^*|| \leq \delta\}$ is also Lipschitz in $\boldsymbol{\beta}$ and therefore a Donsker class.

Let us consider the last term $\Sigma_Y^{-1}(\boldsymbol{x};\boldsymbol{\beta},F)$. Let us fix $\boldsymbol{\beta}$, $F$ and consider $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{X}$. To show that this class of functions is Lipschitz in $\boldsymbol{x}$, note that for matrices $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{K\times K}$,

$$\boldsymbol{A}^{-1} - \boldsymbol{B}^{-1} = \boldsymbol{A}^{-1}(\boldsymbol{B}-\boldsymbol{A})\boldsymbol{B}^{-1} \, .$$

Plugging in $\Sigma_{\boldsymbol{Y}}^{-1}(\boldsymbol{x}_1)$ and $\Sigma_{\boldsymbol{Y}}^{-1}(\boldsymbol{x}_2)$,

$$\begin{aligned}\Sigma_{\boldsymbol{Y}}^{-1}(\boldsymbol{x}_1) - \Sigma_{\boldsymbol{Y}}^{-1}(\boldsymbol{x}_2) &= \Sigma_{\boldsymbol{Y}}^{-1}(\boldsymbol{x}_1)\left(\Sigma_Y(\boldsymbol{x}_2) - \Sigma_Y(\boldsymbol{x}_1)\right)\Sigma_{\boldsymbol{Y}}^{-1}(\boldsymbol{x}_2) \\ \left|\left|\Sigma_{\boldsymbol{Y}}^{-1}(\boldsymbol{x}_1) - \Sigma_{\boldsymbol{Y}}^{-1}(\boldsymbol{x}_2)\right|\right|_F &\leq \left|\left|\Sigma_{\boldsymbol{Y}}^{-1}(\boldsymbol{x}_1)\right|\right|_F \, ||\Sigma_Y(\boldsymbol{x}_1) - \Sigma_Y(\boldsymbol{x}_2)||_F \, \left|\left|\Sigma_{\boldsymbol{Y}}^{-1}(\boldsymbol{x}_2)\right|\right|_F \, .\end{aligned} \tag{5.24}$$

Firstly, let us show that $\left|\left|\Sigma_{\boldsymbol{Y}}^{-1}(\boldsymbol{x})\right|\right|_F$ is uniformly bounded for all $\boldsymbol{x} \in \mathcal{X}$.

Note that in the case where $K = 1$ we trivially have by Assumption A.2.3. that its bounded uniformly by $V_1^{-1}$. Suppose the case where $K > 1$ and let $r$ denote the rank of $\Sigma_{\boldsymbol{Y}}^{-1}$. We don't have much information about the explicit expression of the inverse covariance matrix, so let's try and express the upper bound in terms of the covariance matrix. We have the following inequality using the spectral norm of an inverse matrix

$$\left|\left|\Sigma_{\boldsymbol{Y}}^{-1}(\boldsymbol{x})\right|\right|_F \leq \sqrt{r} \, \left|\left|\Sigma_{\boldsymbol{Y}}^{-1}(\boldsymbol{x})\right|\right|_2 = \frac{\sqrt{r}}{\min\sigma(\Sigma_{\boldsymbol{Y}})} \, .$$

As $\Sigma_{\boldsymbol{Y}}(\boldsymbol{x})$ is semi-positive definite, symmetric and non-singular by assumption, it is positive-definite implying $\min\sigma(\Sigma_Y(\boldsymbol{x})) > 0$. We now wish to find a lower bound on the smallest singular value, of which there are many bounds to choose from. We consider one in terms of the Frobenius norm and determinant shown by Gungor (2010) where for a non-singular matrix $A \in \mathbb{R}^{n\times n}$, that

$$\min\sigma(A) > |\det(A)|\left(\frac{n-1}{||A||_F^2}\right)^{\frac{n-1}{2}} > 0 \, . \tag{5.25}$$

Applying this bound, we have that

$$\left|\left|\Sigma_{\boldsymbol{Y}}^{-1}(\boldsymbol{x})\right|\right|_F \leq \frac{\sqrt{r}}{\min\sigma(\Sigma_Y)} \leq \sqrt{r}\,(|\det(\Sigma_Y(\boldsymbol{x}))|)^{-1}\left(\frac{||\Sigma_Y||_F^2}{K-1}\right)^{\frac{K-1}{2}} \, ,$$

where by Assumption A.2.3., $|\det(\Sigma_Y(\boldsymbol{x}))| \geq V_1 > 0$. Furthermore, applying the Cauchy Schwarz inequality on covariances,

$$||\Sigma_{\boldsymbol{Y}}||_F^2 = \sum_{p=1}^{K}\sum_{q=1}^{K} |\text{Cov}(Y_{(p)}, Y_{(q)})|^2 \leq \sum_{p=1}^{K}\sum_{q=1}^{K} \text{Var}(Y_{(p)})\text{Var}(Y_{(q)}) \ .$$

We have a bounded probability distribution by Assumption A.2.1 which states that $\mathcal{Y}$ is contained in a closed finite hyperrectangle. Thus, by Popoviciu's Inequality on variances, for $Y_k \in [L_k, U_k]$ for $k = 1, 2, \ldots, K$,

$$\text{Var}(Y_k) \leq \frac{1}{4}(U_k - L_k)^2 \leq \frac{1}{4}(U - L)^2 \ ,$$

where $U = \max_{1\leq k\leq K} U_i, L = \min_{1\leq k\leq K} L_i$. Subbing this in we find that

$$||\Sigma_{\boldsymbol{Y}}||_F^2 \leq \frac{K^2}{16}(U - L)^4 \tag{5.26}$$

which gives us our upper bound on $\Sigma_{\boldsymbol{Y}}^{-1}$ for all $x \in \mathcal{X}$ to be

$$\left|\left|\Sigma_{\boldsymbol{Y}}^{-1}(\boldsymbol{x})\right|\right|_F \leq S_1 := \sqrt{r}\, V_2^{-1} \left(\frac{K^2(U-L)^4}{16(K-1)}\right)^{\frac{K-1}{2}} \ . \tag{5.27}$$

To continue with bounding (5.24), let us show that $\{\Sigma_{\boldsymbol{Y}}(\boldsymbol{x}; \boldsymbol{\beta}, F) : \boldsymbol{x} \in \mathcal{X}, ||\boldsymbol{\beta} - \boldsymbol{\beta}^*|| + ||F - F^*|| \leq \delta\}$ is Lipschitz in $\boldsymbol{x}$. Note that applying Taylor's theorem leads to 3-dimensional derivatives. Although it is possible to do it this way, we will remain within our denominator convention for vector-valued function derivatives and apply Taylor's theorem component-wise.

Denoting $\Sigma_Y(\boldsymbol{x})_{pq}$ as the $p$-th row and $q$-th column of the matrix,

$$||\Sigma_{\boldsymbol{Y}}(\boldsymbol{x}_1) - \Sigma_{\boldsymbol{Y}}(\boldsymbol{x}_2)||_F = \sqrt{\sum_{p=1}^{K}\sum_{q=1}^{K} \left(||\Sigma_{\boldsymbol{Y}}(\boldsymbol{x}_1)_{pq} - \Sigma_{\boldsymbol{Y}}(\boldsymbol{x}_2)_{pq}||_2\right)^2} \ .$$

Considering a single term, fix $\boldsymbol{\beta}$ and $F$ and consider $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{X}$. By Taylor's Theorem there exists some $\tilde{\boldsymbol{x}}$ on the line segment from $\boldsymbol{x}_1$ to $\boldsymbol{x}_2$ such that

$$\begin{aligned}
\Sigma_{\boldsymbol{Y}}(\boldsymbol{x}_1)_{pq} - \Sigma_{\boldsymbol{Y}}(\boldsymbol{x}_2)_{pq} &= \left(\frac{\partial}{\partial \boldsymbol{x}}\Sigma_{\boldsymbol{Y}}(\tilde{\boldsymbol{x}})_{pq}\right)^T (\boldsymbol{x}_1 - \boldsymbol{x}_2) \\
||\Sigma_{\boldsymbol{Y}}(\boldsymbol{x}_1)_{pq} - \Sigma_{\boldsymbol{Y}}(\boldsymbol{x}_2)_{pq}||_2 &\leq \left|\left|\frac{\partial}{\partial \boldsymbol{x}}\Sigma_{\boldsymbol{Y}}(\tilde{\boldsymbol{x}})_{pq}\right|\right|_2 ||(\boldsymbol{x}_1 - \boldsymbol{x}_2)||_2 \ .
\end{aligned}$$

Looking at the derivative for some $\boldsymbol{x} \in \mathcal{X}$,

$$\frac{\partial}{\partial \boldsymbol{x}}\Sigma_Y(\boldsymbol{x})_{pq} = \int_{\mathcal{Y}} \frac{\partial}{\partial \boldsymbol{x}}\left[(y_p - \mu_p)(y_q - \mu_q)\right] \exp\left(b + \boldsymbol{\theta}^T\boldsymbol{y}\right) \mathrm{d}F(\boldsymbol{y}) \tag{5.28}$$

$$+ \int_{\mathcal{Y}} (y_p - \mu_p)(y_q - \mu_q) \frac{\partial}{\partial \boldsymbol{x}}\left[\exp\left(b + \boldsymbol{\theta}^T\boldsymbol{y}\right)\right] \mathrm{d}F(\boldsymbol{y}) \tag{5.29}$$

Applying product rule and then noting the mean constraint 2.7, (5.28) becomes

$$\int_{\mathcal{Y}} \left(-\frac{\partial \mu_p}{\partial \boldsymbol{x}}(y_q - \mu_q) - \frac{\partial \mu_q}{\partial \boldsymbol{x}}(y_p - \mu_p)\right) \exp\left(b + \boldsymbol{\theta}^T\boldsymbol{y}\right) \mathrm{d}F(\boldsymbol{y}) = \boldsymbol{0} \ ,$$

and (5.29) becomes

$$\int_{\mathcal{Y}} (y_p - \mu_p)(y_q - \mu_q) \left(\frac{\partial b}{\partial \boldsymbol{x}} + \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{x}}\boldsymbol{y}\right) \left[\exp\left(b + \boldsymbol{\theta}^T\boldsymbol{y}\right)\right] \mathrm{d}F(\boldsymbol{y}) \ .$$

To simplify this, we have the following derivative identities

$$\begin{aligned}\frac{\partial b}{\partial \boldsymbol{x}} &= \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{x}}\frac{\partial \boldsymbol{b}}{\partial \boldsymbol{\theta}} = -\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{x}}\boldsymbol{\mu},\\ \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{x}} &= \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{x}}\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\mu}} = D_2(\boldsymbol{X};\boldsymbol{\beta})\Sigma_{\boldsymbol{Y}}^{-1}(\boldsymbol{X};\boldsymbol{\beta},F)\end{aligned}$$

$$D_2(\boldsymbol{X};\boldsymbol{\beta}) := \frac{\partial \boldsymbol{\mu}(\boldsymbol{x}^T\boldsymbol{\beta})}{\partial \boldsymbol{x}} = \begin{bmatrix} \frac{\partial \mu_1}{\partial \boldsymbol{x}_1} & \frac{\partial \mu_2}{\partial \boldsymbol{x}_1} & \cdots & \frac{\partial \mu_K}{\partial \boldsymbol{x}_1} \\ \frac{\partial \mu_1}{\partial \boldsymbol{x}_2} & \frac{\partial \mu_2}{\partial \boldsymbol{x}_2} & \cdots & \frac{\partial \mu_K}{\partial \boldsymbol{x}_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \mu_1}{\partial \boldsymbol{x}_q} & \frac{\partial \mu_2}{\partial \boldsymbol{x}_q} & \cdots & \frac{\partial \mu_K}{\partial \boldsymbol{x}_q} \end{bmatrix}$$

For component $(j,k)$ in $D_2$,

$$\begin{aligned}\frac{\partial \mu_{(k)}(\boldsymbol{X}_{(k)}^T\boldsymbol{\beta}_{(k')})}{\partial x_j} &= \frac{\partial(\boldsymbol{X}_{(k)}^T\boldsymbol{\beta}_{(k')})}{\partial x_j}\mu'_{(k)}(\boldsymbol{X}_{(k)}^T\boldsymbol{\beta}_{(k')})\\ &= \begin{cases}\boldsymbol{\beta}_j\mu'_{(k)}(\boldsymbol{X}_{(k)}^T\boldsymbol{\beta}_{(k')}) & , x_j \in X_{(k)}\\ 0 & , x_j \notin X_{(k)}\end{cases}\end{aligned}$$

Therefore, we can express $D_2(\boldsymbol{X};\boldsymbol{\beta})$ as

$$D_2(\boldsymbol{X};\boldsymbol{\beta}) = \frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{x}}\begin{bmatrix} \mu'_{(1)} & 0 & \dots & 0 \\ 0 & \mu'_{(2)} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \mu'_{(K)} \end{bmatrix}$$

where $\boldsymbol{\eta} \in \mathbb{R}^K$ is a vector of linear predictors for the $K$ components. Plugging these into (5.29),

$$\frac{\partial}{\partial \boldsymbol{x}}\Sigma_Y(\boldsymbol{x})_{pq} = D_2(\boldsymbol{X};\boldsymbol{\beta})\Sigma_{\boldsymbol{Y}}^{-1}(\boldsymbol{X};\boldsymbol{\beta},F)\int_{\mathcal{Y}}(y_p-\mu_p)(y_q-\mu_q)\,(\boldsymbol{y}-\boldsymbol{\mu})\exp\big(b+\boldsymbol{\theta}^T\boldsymbol{y}\big)\,\mathrm{d}F(\boldsymbol{y}) \tag{5.30}$$

Let us consider the upper bounds of the norms of each term. Firstly,

$$(y_p-\mu_p)(y_q-\mu_q)\,(\boldsymbol{y}-\boldsymbol{\mu}) = \begin{bmatrix}(y_p-\mu_p)(y_q-\mu_q)(y_1-\mu_1)\\(y_p-\mu_p)(y_q-\mu_q)(y_2-\mu_2)\\ \vdots \\ (y_p-\mu_p)(y_q-\mu_q)(y_K-\mu_K)\end{bmatrix} \le \begin{bmatrix}(U-L)^3\\(U-L)^3\\ \vdots \\ (U-L)^3\end{bmatrix}$$

implying

$$\left|\left|\int_{\mathcal{Y}}(y_p-\mu_p)(y_q-\mu_q)\,(\boldsymbol{y}-\boldsymbol{\mu})\exp\big(b+\boldsymbol{\theta}^T\boldsymbol{y}\big)\,\mathrm{d}F(\boldsymbol{y})\right|\right|_2 \le ||(\boldsymbol{U}-\boldsymbol{L})^3||_2 = K(U-L)^3\,.$$

Next, note that $||\Sigma_{\boldsymbol{Y}}^{-1}||_F \le S_1$ and

$$||D_2(\boldsymbol{X};\boldsymbol{\beta})||_F \le \left|\left|\frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{x}}\right|\right|_F \left|\left|\mathrm{Diag}\left(\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\eta}}\right)\right|\right|_F = \left|\left|\frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{x}}\right|\right|_F ||\boldsymbol{\mu}'||_2$$

We already have that $||\boldsymbol{\mu}'||_2 \leq KM_1$, and the element $(j,k)$ of $\frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{x}}$ are either $\boldsymbol{\beta}_j$ or 0. So we have that

$$\begin{aligned}
\left|\left|\frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{x}}\right|\right|_F &\leq \left|\left|\begin{bmatrix} \beta_1 & \beta_1 & \dots & \beta_1 \\ \beta_2 & \beta_2 & \dots & \beta_2 \\ \vdots & \vdots & \ddots & \vdots \\ \beta_q & \beta_q & \dots & \beta_q \end{bmatrix}\right|\right|_F \\
&= \sqrt{\sum_{k=1}^{K}\sum_{j=1}^{q}|\beta_j|^2} \\
&\leq \sum_{k=1}^{K}\sqrt{\sum_{j=1}^{q}|\beta_j|^2} \\
&= \sum_{k=1}^{K}||\boldsymbol{\beta}||_2 \\
&= K||\boldsymbol{\beta}||_2
\end{aligned}$$

Therefore,

$$||D_2(\boldsymbol{X};\boldsymbol{\beta})||_F \leq \left|\left|\frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{x}}\right|\right|_F ||\boldsymbol{\mu}'||_2 \leq K^2 M_1 ||\boldsymbol{\beta}||_2 \leq K^2 M_1 \left(||\boldsymbol{\beta}^*||_2 + \delta\right) \tag{5.31}$$

Bringing this all together we have that

$$||\Sigma_{\boldsymbol{Y}}(\boldsymbol{x}_1)_{pq} - \Sigma_{\boldsymbol{Y}}(\boldsymbol{x}_2)_{pq}||_2 \leq S_1 K^3 M_1 (||\boldsymbol{\beta}^*||_2 + \delta)(U-L)^3 ||\boldsymbol{x}_1 - \boldsymbol{x}_2||_2 \,. \tag{5.32}$$

Defining

$$S_2 := S_1 K^3 M_1 (||\boldsymbol{\beta}^*||_2 + \delta)(U-L)^3,$$

we finally find that

$$||\Sigma_{\boldsymbol{Y}}(\boldsymbol{x}_1) - \Sigma_{\boldsymbol{Y}}(\boldsymbol{x}_2)||_F \leq \sqrt{\sum_{p=1}^{K}\sum_{q=1}^{K}\left(S_2||\boldsymbol{x}_1 - \boldsymbol{x}_2||_2\right)^2} = KS_2||\boldsymbol{x}_1 - \boldsymbol{x}_2||_2 \tag{5.33}$$

As a result the class of functions $\{\Sigma_Y(\boldsymbol{x};\boldsymbol{\beta},F) : \boldsymbol{x} \in \mathcal{X}, ||\boldsymbol{\beta} - \boldsymbol{\beta}^*|| + ||F - F^*|| \leq \delta\}$ is Lipschitz in $\boldsymbol{x}$ and is a Donsker class. Now, subbing (5.27) and (5.33) into (5.24)

$$||\Sigma_{\boldsymbol{Y}}^{-1}(\boldsymbol{x}_1) - \Sigma_{\boldsymbol{Y}}^{-1}(\boldsymbol{x}_2)||_F \leq S_1^3 K^4 M_1 (||\boldsymbol{\beta}^*|| + \delta)(U-L)^3 ||\boldsymbol{x}_1 - \boldsymbol{x}_2||_2 \tag{5.34}$$

As a result the class of functions $\{\Sigma_{\boldsymbol{Y}}^{-1}(\boldsymbol{x};\boldsymbol{\beta},F) : \boldsymbol{x} \in \mathcal{X}, ||\boldsymbol{\beta} - \boldsymbol{\beta}^*|| + ||F - F^*|| \leq \delta\}$ is Lipschitz in $\boldsymbol{x}$ and is a Donsker class.

From Van Der Vaart and Wellner (1996) we have that the product of uniformly bounded Donsker classes forms a Donsker class. We have that $||\text{Diag}(\boldsymbol{X}')||_F$ is uniformly bounded by the hyper-rectangle, $||(\boldsymbol{Y} - \boldsymbol{\mu})||_2$ is uniformly bounded by $K(U-L)$, $||\boldsymbol{\mu}'||_2$ is uniformly bounded by $KM_1$ and $||\Sigma_{\boldsymbol{Y}}^{-1}||_F$ is uniformly bounded by $S_1$.

Therefore $\{S_{\boldsymbol{\beta},F} : ||\boldsymbol{\beta} - \boldsymbol{\beta}^*|| + ||F - F^*|| \leq \delta\}$ is $P^*$ Donsker as it's the product of four uniformly bounded Donsker classes.

**Showing Score Equation for $F$ is Donsker**

We have that the Score equation for $F$ given below is the sum of three terms which we will show are Donsker classes.

$$A_{\boldsymbol{\beta},F}h(\boldsymbol{x},\boldsymbol{y}) = h(\boldsymbol{y}) - B_{\boldsymbol{\beta},F}h(\boldsymbol{x}) - C_{\boldsymbol{\beta},F}h(\boldsymbol{x})\Sigma_{\boldsymbol{Y}}^{-\frac{1}{2}}(\boldsymbol{Y}-\boldsymbol{\mu})$$

The first term $h(\boldsymbol{y})$ belongs to the class $\mathcal{H}_L$ of left-half plane indicator functions which is a Vapink-Cervonenkis (VC) class and hence a Donsker class. Note that it is also bounded by 1 so it's a bounded Donsker class.

The second term $B_{\boldsymbol{\beta},F}h(\boldsymbol{X})$ is bounded in absolute value by 1, continuous and continuously differentiable in $\boldsymbol{x}$. Let us fix $\boldsymbol{\beta}$, $F$ and consider $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{X}$. By Taylor's theorem, there exists some $\tilde{\boldsymbol{x}}$ on the line segment between $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ where

$$\begin{aligned} B_{\boldsymbol{\beta},F}h(\boldsymbol{x}_1) - B_{\boldsymbol{\beta},F}h(\boldsymbol{x}_2) &= \left(\frac{\partial}{\partial \boldsymbol{x}}B_{\boldsymbol{\beta},F}h(\tilde{\boldsymbol{x}})\right)^T(\boldsymbol{x}_1-\boldsymbol{x}_2) \\ ||B_{\boldsymbol{\beta},F}h(\boldsymbol{x}_1) - B_{\boldsymbol{\beta},F}h(\boldsymbol{x}_2)||_2 &\leq \left|\left|\frac{\partial}{\partial \boldsymbol{x}}B_{\boldsymbol{\beta},F}h(\tilde{\boldsymbol{x}})\right|\right|_2 ||\boldsymbol{x}_1-\boldsymbol{x}_2||_2 \end{aligned} \tag{5.35}$$

Applying the Leibniz rule and derivative identities found previously.

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{x}}B_{\boldsymbol{\beta},F}h(\tilde{\boldsymbol{x}}) &= \int_{\mathcal{Y}} \frac{\partial}{\partial \boldsymbol{x}}h(\boldsymbol{y})\exp\left(b+\boldsymbol{\theta}^T\boldsymbol{y}\right)\mathrm{d}F(\boldsymbol{y}) \\ &= \int_{\mathcal{Y}} h(\boldsymbol{y})\left(\frac{\partial b}{\partial \boldsymbol{x}} + \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{x}}\boldsymbol{y}\right)\exp\left(b+\boldsymbol{\theta}^T\boldsymbol{y}\right)\mathrm{d}F(\boldsymbol{y}) \\ &= D_2\Sigma_{\boldsymbol{Y}}^{-1}\int_{\mathcal{Y}} h(\boldsymbol{y})\left(\boldsymbol{y}-\boldsymbol{\mu}\right)\exp\left(b+\boldsymbol{\theta}^T\boldsymbol{y}\right)\mathrm{d}F(\boldsymbol{y}) \end{aligned}$$

Therefore, noting that

$$\left|\left|\int_{\mathcal{Y}} h(\boldsymbol{y})\left(\boldsymbol{y}-\boldsymbol{\mu}\right)\exp\left(b+\boldsymbol{\theta}^T\boldsymbol{y}\right)\mathrm{d}F(\boldsymbol{y})\right|\right|_2 \leq \left|\left|\int_{\mathcal{Y}} \left(\boldsymbol{y}-\boldsymbol{\mu}\right)\exp\left(b+\boldsymbol{\theta}^T\boldsymbol{y}\right)\mathrm{d}F(\boldsymbol{y})\right|\right|_2 \leq K(U-L)$$

then we can establish the following upper bound on the derivative

$$\begin{aligned} \left|\left|\frac{\partial}{\partial \boldsymbol{x}}B_{\boldsymbol{\beta},F}h(\tilde{\boldsymbol{x}})\right|\right|_2 &\leq ||D_2||_F||\Sigma_{\boldsymbol{Y}}^{-1}||_F\left|\left|\int_{\mathcal{Y}} h(\boldsymbol{y})\left(\boldsymbol{y}-\boldsymbol{\mu}\right)\exp\left(b+\boldsymbol{\theta}^T\boldsymbol{y}\right)\mathrm{d}F(\boldsymbol{y})\right|\right|_2 \\ \left|\left|\frac{\partial}{\partial \boldsymbol{x}}B_{\boldsymbol{\beta},F}h(\tilde{\boldsymbol{x}})\right|\right|_2 &\leq S_1K^3M_1(||\boldsymbol{\beta}^*||_2+\delta)(U-L) \end{aligned} \tag{5.36}$$

Plugging (5.36) into (5.35)

$$||B_{\boldsymbol{\beta},F}h(\boldsymbol{x}_1) - B_{\boldsymbol{\beta},F}h(\boldsymbol{x}_2)||_2 \leq S_1K^3M_1(||\boldsymbol{\beta}^*||_2+\delta)(U-L)\,||\boldsymbol{x}_1-\boldsymbol{x}_2||_2 \tag{5.37}$$

implying $\{B_{\boldsymbol{\beta},F}h : h \in \mathcal{H}_L, ||\boldsymbol{\beta}-\boldsymbol{\beta}^*|| + ||F-F^*|| \leq \delta\}$ is uniformly bounded, continuous and Lipschitz in $\boldsymbol{x}$, and hence a Donsker class.

Considering the third term $C_{\boldsymbol{\beta},F}h(\boldsymbol{x})\Sigma_Y^{-\frac{1}{2}}(\boldsymbol{Y}-\boldsymbol{\mu}) = \mathbb{E}_{\boldsymbol{\beta},F}\left[h(\boldsymbol{Y})(\boldsymbol{Y}-\boldsymbol{\mu})^T|\boldsymbol{X}=\boldsymbol{x}\right]\Sigma_{\boldsymbol{Y}}^{-1}(\boldsymbol{X};\boldsymbol{\beta},F)(\boldsymbol{Y}-\boldsymbol{\mu})$, we have shown that $\{\boldsymbol{y}-\boldsymbol{\mu}(\boldsymbol{x}^T\boldsymbol{\beta}) : ||\boldsymbol{\beta}-\boldsymbol{\beta}^*|| \leq \delta\}$ and $\{\Sigma_{\boldsymbol{Y}}^{-1}(\boldsymbol{x};\boldsymbol{\beta},F) : \boldsymbol{x}\in\mathcal{X}, ||\boldsymbol{\beta}-\boldsymbol{\beta}^*||+||F-F^*|| \leq \delta\}$ are bounded Donsker classes, so it remains to consider the term $\mathbb{E}_{\boldsymbol{\beta},F}\left[h(\boldsymbol{Y})(\boldsymbol{Y}-\boldsymbol{\mu})^T|\boldsymbol{X}=\boldsymbol{x}\right]$. Fixing $\boldsymbol{\beta}$, $F$ and consider $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{X}$. By Taylor's theorem, there exists some $\tilde{\boldsymbol{x}}$ on the line segment between $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ where

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\beta},F}\left[h(\boldsymbol{Y})(\boldsymbol{Y}-\boldsymbol{\mu})|\boldsymbol{x}_1\right] - \mathbb{E}_{\boldsymbol{\beta},F}\left[h(\boldsymbol{Y})(\boldsymbol{Y}-\boldsymbol{\mu})|\boldsymbol{x}_2\right] &= \left(\frac{\partial}{\partial \boldsymbol{x}}\mathbb{E}_{\boldsymbol{\beta},F}\left[h(\boldsymbol{Y})(\boldsymbol{Y}-\boldsymbol{\mu})|\tilde{\boldsymbol{x}}\right]\right)^T(\boldsymbol{x}_1-\boldsymbol{x}_2) \\ ||\mathbb{E}_{\boldsymbol{\beta},F}\left[h(\boldsymbol{Y})(\boldsymbol{Y}-\boldsymbol{\mu})|\boldsymbol{x}_1\right] - \mathbb{E}_{\boldsymbol{\beta},F}\left[h(\boldsymbol{Y})(\boldsymbol{Y}-\boldsymbol{\mu})|\boldsymbol{x}_2\right]||_2 &\leq \left|\left|\frac{\partial}{\partial \boldsymbol{x}}\mathbb{E}_{\boldsymbol{\beta},F}\left[h(\boldsymbol{Y})(\boldsymbol{Y}-\boldsymbol{\mu})|\tilde{\boldsymbol{x}}\right]\right|\right|_F ||\boldsymbol{x}_1-\boldsymbol{x}_2||_2 \end{aligned}$$

Considering the derivative,

$$\frac{\partial}{\partial \boldsymbol{x}}\mathbb{E}_{\boldsymbol{\beta},F}\left[h(\boldsymbol{Y})(\boldsymbol{Y}-\boldsymbol{\mu})|\boldsymbol{x}\right] = \int_{\mathcal{Y}} \frac{\partial}{\partial \boldsymbol{x}} h(\boldsymbol{y})(\boldsymbol{y}-\boldsymbol{\mu}) \exp\left(b+\boldsymbol{\theta}^T\boldsymbol{y}\right) \mathrm{d}F(\boldsymbol{y})$$

Using the matrix identity, $\frac{\partial (v(\boldsymbol{x})\boldsymbol{u}(\boldsymbol{x}))}{\partial \boldsymbol{x}} = v(\boldsymbol{x})\frac{\partial \boldsymbol{u}(\boldsymbol{x})}{\partial \boldsymbol{x}} + \frac{\partial v(\boldsymbol{x})}{\partial \boldsymbol{x}}\left(\boldsymbol{u}(\boldsymbol{x})\right)^T$

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{x}}\mathbb{E}_{\boldsymbol{\beta},F}\left[h(\boldsymbol{Y})(\boldsymbol{Y}-\boldsymbol{\mu})|\boldsymbol{x}\right] &= \int_{\mathcal{Y}} \left(-\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{x}}\right) h(\boldsymbol{y}) \exp\left(b+\boldsymbol{\theta}^T\boldsymbol{y}\right) \mathrm{d}F(\boldsymbol{y}) \\
&\quad + \int_{\mathcal{Y}} \left(\frac{\partial b}{\partial \boldsymbol{x}} + \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{x}}\boldsymbol{y}\right) h(\boldsymbol{y}) \exp\left(b+\boldsymbol{\theta}^T\boldsymbol{y}\right)(\boldsymbol{y}-\boldsymbol{\mu})^T \mathrm{d}F(\boldsymbol{y}) \\
&= -D_2 \int_{\mathcal{Y}} h(\boldsymbol{y}) \exp\left(b+\boldsymbol{\theta}^T\boldsymbol{y}\right) \mathrm{d}F(\boldsymbol{y}) \\
&\quad + D_2\Sigma_{\boldsymbol{Y}}^{-1} \int_{\mathcal{Y}} h(\boldsymbol{y})(\boldsymbol{y}-\boldsymbol{\mu})(\boldsymbol{y}-\boldsymbol{\mu})^T \exp\left(b+\boldsymbol{\theta}^T\boldsymbol{y}\right) \mathrm{d}F(\boldsymbol{y})
\end{aligned}$$

Therefore,

$$\begin{aligned}
\left|\left|\frac{\partial}{\partial \boldsymbol{x}}\mathbb{E}_{\boldsymbol{\beta},F}\left[h(\boldsymbol{Y})(\boldsymbol{Y}-\boldsymbol{\mu})|\tilde{\boldsymbol{x}}\right]\right|\right|_F &\leq ||D_2||_F + ||D_2||_F||\Sigma_{\boldsymbol{Y}}^{-1}||_F|| \int_{\mathcal{Y}} (\boldsymbol{y}-\boldsymbol{\mu})(\boldsymbol{y}-\boldsymbol{\mu})^T \exp\left(b+\boldsymbol{\theta}^T\boldsymbol{y}\right) \mathrm{d}F(\boldsymbol{y})||_F \\
&= ||D_2||_F + ||D_2||_F||\Sigma_{\boldsymbol{Y}}^{-1}||_F||\Sigma_{\boldsymbol{Y}}||_F \\
&\leq K^2 M_1(||\boldsymbol{\beta}^*||_2+\delta) + \frac{K^3}{4} M_1(||\boldsymbol{\beta}^*||_2+\delta) S_1 (U-L)^2 \qquad (5.38)
\end{aligned}$$

Let us define $S_3 := K^2 M_1(||\boldsymbol{\beta}^*||_2+\delta) + \frac{K^3}{4} M_1(||\boldsymbol{\beta}^*||_2+\delta) S_1 (U-L)^2$. Then, noting the norm of a vector is equal to the norm of its transpose,

$$||\mathbb{E}_{\boldsymbol{\beta},F}\left[h(\boldsymbol{Y})(\boldsymbol{Y}-\boldsymbol{\mu})|\boldsymbol{x}_1\right] - \mathbb{E}_{\boldsymbol{\beta},F}\left[h(\boldsymbol{Y})(\boldsymbol{Y}-\boldsymbol{\mu})|\boldsymbol{x}_2\right]||_2 \leq S_3||\boldsymbol{x}_1-\boldsymbol{x}_2||_2 \qquad (5.39)$$

$$\implies ||\mathbb{E}_{\boldsymbol{\beta},F}\left[h(\boldsymbol{Y})(\boldsymbol{Y}-\boldsymbol{\mu})^T|\boldsymbol{x}_1\right] - \mathbb{E}_{\boldsymbol{\beta},F}\left[h(\boldsymbol{Y})(\boldsymbol{Y}-\boldsymbol{\mu})^T|\boldsymbol{x}_2\right]||_2 \leq S_3||\boldsymbol{x}_1-\boldsymbol{x}_2||_2 \qquad (5.40)$$

Thus, we have shown that $\{\mathbb{E}_{\boldsymbol{\beta},F}\left[h(\boldsymbol{Y})(\boldsymbol{Y}-\boldsymbol{\mu})^T|\boldsymbol{x}\right] : ||\boldsymbol{\beta}-\boldsymbol{\beta}^*|| + ||F-F^*|| \leq \delta\}$ is a Lipschitz in $\boldsymbol{x}$ and hence a Donsker class. Note that it is also uniformly bounded by $U-L$, which gives that $\{C_{\boldsymbol{\beta},F}h(\boldsymbol{x})\Sigma_{\boldsymbol{Y}}^{-\frac{1}{2}}(\boldsymbol{Y}-\boldsymbol{\mu}) : \boldsymbol{x}\in\mathcal{X}, ||\boldsymbol{\beta}-\boldsymbol{\beta}^*|| + ||F-F^*|| \leq \delta\}$ is uniformly bounded and hence a Donsker Class.

Finally, from Van Der Vaart and Wellner (1996) we have that the sum of three Donsker classes is a Donsker class, implying that $\{A_{\boldsymbol{\beta},F}h : h\in\mathcal{H}_L, ||\boldsymbol{\beta}-\boldsymbol{\beta}^*|| + ||F-F^*|| \leq \delta\}$ is $P^*$-Donsker.

**Convergence of Score Equations**

Firstly, note that $S_{\boldsymbol{\beta},F} \to S_{\boldsymbol{\beta}^*,F^*}$ point-wise by the continuous mapping theorem as $\boldsymbol{\beta}\to\boldsymbol{\beta}^*$ and $F\to F^*$, and $|S_{\boldsymbol{\beta},F}|$ is uniformly bounded by a constant as each of its terms are. Therefore, by the dominated convergence theorem, we have that

$$P^*(S_{\boldsymbol{\beta},F} - S_{\boldsymbol{\beta}^*,F^*})^2 \to 0, \text{ as } \boldsymbol{\beta}\to\boldsymbol{\beta}^*, F\to F^*$$

Similarly, $A_{\boldsymbol{\beta},F}h \to A_{\boldsymbol{\beta}^*,F^*}h$ pointwise, uniformly in $h$ as $\boldsymbol{\beta}\to\boldsymbol{\beta}^*$ and $F\to F^*$, and $|A_{\boldsymbol{\beta},F}h|$ is uniformly bounded by a constant as each of its terms are. Therefore, by the dominated convergence theorem, we have that

$$\sup_{h\in\mathcal{H}_L} P^*(A_{\boldsymbol{\beta},F}h - A_{\boldsymbol{\beta}^*,F^*}h)^2 \to 0, \text{ as } \boldsymbol{\beta}\to\boldsymbol{\beta}^*, F\to F^*$$

Thus, all sufficient conditions (5.19), (5.20) (5.21) hold, establishing conditions (5.13) and (5.14) from Theorem 5.2.

### 5.3.2 Asymptotic Normality of Score Functions

By establishing condition (5.19), by the definition of Donsker classes we can establish asymptotic normality of the score functions in Lemma 5.1.

**Lemma 5.1** *As $n \to \infty$, $\sqrt{n}\,(\Psi_n - \Psi)(\boldsymbol{\beta}^*, F^*) \rightsquigarrow \boldsymbol{Z} = (Z_1, Z_2)^T$ in distribution in $\mathbb{R}^q \times \ell^\infty(\mathcal{H}_L)$ where $Z_1$ is a mean zero Gaussian random vector with covariance matrix*

$$W_{\boldsymbol{\beta}} = \mathbb{E}^{\mathcal{X}}\left[D(\boldsymbol{X};\boldsymbol{\beta}^*)\Sigma_Y^{-1}(\boldsymbol{X};\boldsymbol{\beta}^*,F^*)D(\boldsymbol{X};\boldsymbol{\beta}^*)^T\right] \tag{5.41}$$

*$Z_2$ is a mean zero Gaussian random process indexed by $h \in \mathcal{H}_L$ with covariance function*

$$W_F(h_1,h_2) = \mathbb{E}^{\mathcal{X}}\left[Cov(h_1(\boldsymbol{Y}),h_2(\boldsymbol{Y})|\boldsymbol{X}) - C_{\boldsymbol{\beta}^*,F^*}h_1(\boldsymbol{X})C_{\boldsymbol{\beta}^*,F^*}h_2(\boldsymbol{X})^T\right] \tag{5.42}$$

*and $Z_1, Z_2$ are independent*

This is a vectorized version of Lemma 1.1 in the supplementary material of Huang (2014), and we will verify the covariance structure of the limiting variable $Z$ directly. The first component of $\sqrt{n}\,(\Psi_n - \Psi)(\boldsymbol{\beta}^*, F^*)$, we have shown that the expectation of the score function $S_{\boldsymbol{\beta}^*,F^*}$ is 0.

We also have that $\mathrm{Cov}\left(\sqrt{n}\,(\Psi_{1n} - \Psi_1)(\boldsymbol{\beta}^*, F^*), \sqrt{n}\,(\Psi_{1n} - \Psi_1)(\boldsymbol{\beta}^*, F^*)\right)$ is equal to

$$\begin{aligned}\mathrm{Cov}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n S_{\boldsymbol{\beta}^*,F^*}(\boldsymbol{X}_i,\boldsymbol{Y}_i), \frac{1}{\sqrt{n}}\sum_{i=1}^n S_{\boldsymbol{\beta}^*,F^*}(\boldsymbol{X}_i,\boldsymbol{Y}_i)\right) &= \frac{1}{n}\sum_{i=1}^n \mathrm{Cov}\left(S_{\boldsymbol{\beta}^*,F^*}(\boldsymbol{X}_i,\boldsymbol{Y}_i), S_{\boldsymbol{\beta}^*,F^*}(\boldsymbol{X}_i,\boldsymbol{Y}_i)\right)\\ &= \mathrm{Cov}\left(S_{\boldsymbol{\beta}^*,F^*}(\boldsymbol{X},\boldsymbol{Y}), S_{\boldsymbol{\beta}^*,F^*}(\boldsymbol{X},\boldsymbol{Y})\right)\end{aligned}$$

as $(\boldsymbol{X}_i, \boldsymbol{Y}_i)$ is an i.i.d copy of a generic pair $(\boldsymbol{X}, \boldsymbol{Y})$.

Thus, evaluating the expression $W_{\boldsymbol{\beta}} := \mathrm{Cov}\left(S_{\boldsymbol{\beta}^*,F^*}(\boldsymbol{X},\boldsymbol{Y}), S_{\boldsymbol{\beta}^*,F^*}(\boldsymbol{X},\boldsymbol{Y})\right)$, we find

$$\begin{aligned}W_{\boldsymbol{\beta}} &= \mathbb{E}^{\mathcal{X}}\left[\mathbb{E}_{\boldsymbol{\beta}^*}\left[D(\boldsymbol{X};\boldsymbol{\beta})\Sigma_{\boldsymbol{Y}}^{-1}(\boldsymbol{X};\boldsymbol{\beta},F)(\boldsymbol{Y}-\boldsymbol{\mu})(\boldsymbol{Y}-\boldsymbol{\mu})^T\Sigma_{\boldsymbol{Y}}^{-1}(\boldsymbol{X};\boldsymbol{\beta},F)D(\boldsymbol{X};\boldsymbol{\beta})^T|\boldsymbol{X}\right]\right]\\ &= \mathbb{E}^{\mathcal{X}}\left[D(\boldsymbol{X};\boldsymbol{\beta}^*)\Sigma_{\boldsymbol{Y}}^{-1}(\boldsymbol{X};\boldsymbol{\beta}^*,F^*)\mathbb{E}_{\boldsymbol{\beta}^*}\left[(\boldsymbol{Y}-\boldsymbol{\mu})(\boldsymbol{Y}-\boldsymbol{\mu})^T|\boldsymbol{X}\right]\Sigma_{\boldsymbol{Y}}^{-1}(\boldsymbol{X};\boldsymbol{\beta}^*,F^*)D(\boldsymbol{X};\boldsymbol{\beta}^*)^T\right]\end{aligned}$$

Here the operator $\mathbb{E}_{\boldsymbol{\beta}^*}(\cdot)$ is with respect to the true distribution $F^*(\boldsymbol{y}|\boldsymbol{x})$ where $F^*(\boldsymbol{y}|\boldsymbol{x})$ is some multivariate distribution. Given the assumption that $F^*(\boldsymbol{y}|\boldsymbol{x})$ is within the multivariate exponential family, we have that

$$\mathbb{E}_{\boldsymbol{\beta}^*}\left[(\boldsymbol{Y}-\boldsymbol{\mu})(\boldsymbol{Y}-\boldsymbol{\mu})^T|\boldsymbol{X}\right] = \mathbb{E}_{\boldsymbol{\beta}^*,F^*}\left[(\boldsymbol{Y}-\boldsymbol{\mu})(\boldsymbol{Y}-\boldsymbol{\mu})^T|\boldsymbol{X}\right] = \Sigma_{\boldsymbol{Y}}(\boldsymbol{X};\boldsymbol{\beta}^*,F^*), \tag{5.43}$$

which gives

$$W_{\boldsymbol{\beta}} = \mathbb{E}^{\mathcal{X}}\left[D(\boldsymbol{X};\boldsymbol{\beta}^*)\Sigma_Y^{-1}(\boldsymbol{X};\boldsymbol{\beta}^*,F^*)D(\boldsymbol{X};\boldsymbol{\beta}^*)^T\right]. \tag{5.44}$$

The second component of $\sqrt{n}\,(\Psi_n - \Psi)(\boldsymbol{\beta}^*, F^*)$ has mean 0 as the expectation of the score function $A_{\boldsymbol{\beta}^*,F^*}h$ has zero mean for any $h \in \mathcal{H}_L$. For any $h_1, h_2 \in \mathcal{H}_L$ we similarly have that

$$\mathrm{Cov}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n A_{\boldsymbol{\beta}^*,F^*}h_1(\boldsymbol{X},\boldsymbol{Y}), \frac{1}{\sqrt{n}}\sum_{i=1}^n A_{\boldsymbol{\beta}^*,F^*}h_2(\boldsymbol{X},\boldsymbol{Y})\right) = \mathrm{Cov}\left(A_{\boldsymbol{\beta}^*,F^*}h_1(\boldsymbol{X},\boldsymbol{Y}), A_{\boldsymbol{\beta}^*,F^*}h_2(\boldsymbol{X},\boldsymbol{Y})\right)$$

By the law of total covariance, noting that the conditional expectation of the score is 0 for all $h \in \mathcal{H}_L$, we can express this as

$$\begin{aligned}\mathrm{Cov}\left(A_{\boldsymbol{\beta}^*,F^*}h_1(\boldsymbol{X}_i,\boldsymbol{Y}_i), A_{\boldsymbol{\beta}^*,F^*}h_2(\boldsymbol{X}_i,\boldsymbol{Y}_i)\right) &= \mathbb{E}^{\mathcal{X}}\left[\mathrm{Cov}\left(A_{\boldsymbol{\beta}^*,F^*}h_1(\boldsymbol{X},\boldsymbol{Y}), A_{\boldsymbol{\beta}^*,F^*}h_2(\boldsymbol{X},\boldsymbol{Y})|\boldsymbol{X}\right)\right]\\ &\quad + \mathrm{Cov}\left[\mathbb{E}\left(A_{\boldsymbol{\beta}^*,F^*}h_1(\boldsymbol{X},\boldsymbol{Y})|\boldsymbol{X}\right)\mathbb{E}\left(A_{\boldsymbol{\beta}^*,F^*}h_2(\boldsymbol{X},\boldsymbol{Y})|\boldsymbol{X}\right)\right]\\ &= \mathbb{E}^{\mathcal{X}}\left[\mathrm{Cov}\left(A_{\boldsymbol{\beta}^*,F^*}h_1(\boldsymbol{X},\boldsymbol{Y}), A_{\boldsymbol{\beta}^*,F^*}h_2(\boldsymbol{X},\boldsymbol{Y})|\boldsymbol{X}\right)\right]\end{aligned}$$

To simplify the working, for $j = 1, 2$ let

$$g_j(\boldsymbol{X}, \boldsymbol{Y}) = B_{\boldsymbol{\beta}^*, F^*} h_j(\boldsymbol{X}) - C_{\boldsymbol{\beta}^*, F^*} h_j(\boldsymbol{X}) \Sigma_{\boldsymbol{Y}}^{-1} (\boldsymbol{Y} - \boldsymbol{\mu}) \tag{5.45}$$

Therefore the conditional covariance becomes,

$$\begin{aligned}\operatorname{Cov}\left(A_{\boldsymbol{\beta}^*, F^*} h_1(\boldsymbol{X}, \boldsymbol{Y}), A_{\boldsymbol{\beta}^*, F^*} h_2(\boldsymbol{X}, \boldsymbol{Y}) | \boldsymbol{X}\right) &= \operatorname{Cov}(h_1(\boldsymbol{Y}), h_2(\boldsymbol{Y}) | \boldsymbol{X}) - \operatorname{Cov}(h_1(\boldsymbol{Y}), g_2(\boldsymbol{X}, \boldsymbol{Y}) | \boldsymbol{X}) \\ &\quad - \operatorname{Cov}(h_2(\boldsymbol{Y}), g_1(\boldsymbol{X}, \boldsymbol{Y}) | \boldsymbol{X}) + \operatorname{Cov}(g_1(\boldsymbol{X}, \boldsymbol{Y}), g_2(\boldsymbol{X}, \boldsymbol{Y}) | \boldsymbol{X})\end{aligned}$$

Considering the second term $\operatorname{Cov}(h_1(\boldsymbol{Y}), g_2(\boldsymbol{X}, \boldsymbol{Y}) | \boldsymbol{X})$, noting that

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\beta}^*, F^*}\left[g_2(\boldsymbol{X}, \boldsymbol{Y}) | \boldsymbol{X}\right] &= \mathbb{E}_{\boldsymbol{\beta}^*, F^*}\left[B_{\boldsymbol{\beta}^*, F^*} h_2(\boldsymbol{X}) + C_{\boldsymbol{\beta}^*, F^*} h_2(\boldsymbol{X}) \Sigma_{\boldsymbol{Y}}^{-\frac{1}{2}} (\boldsymbol{Y} - \boldsymbol{\mu}) | \boldsymbol{X}\right] \\ &= B_{\boldsymbol{\beta}^*, F^*} h_2(\boldsymbol{X}) + C_{\boldsymbol{\beta}^*, F^*} h_2(\boldsymbol{X}) \Sigma_{Y}^{-\frac{1}{2}} \mathbb{E}_{\boldsymbol{\beta}^*, F^*}\left[(\boldsymbol{Y} - \boldsymbol{\mu}) | \boldsymbol{X}\right] \\ &= B_{\boldsymbol{\beta}^*, F^*} h_2(\boldsymbol{X}) \\ \mathbb{E}_{\boldsymbol{\beta}^*, F^*}\left[h_1(\boldsymbol{Y}) g_2(\boldsymbol{X}, \boldsymbol{Y}) | \boldsymbol{X}\right] &= \mathbb{E}_{\boldsymbol{\beta}^*, F^*}\left[h_1(\boldsymbol{Y}) B_{\boldsymbol{\beta}^*, F^*} h_2(\boldsymbol{X}) | \boldsymbol{X}\right] \\ &\quad + \mathbb{E}_{\boldsymbol{\beta}^*, F^*}\left[h_1(\boldsymbol{Y}) C_{\boldsymbol{\beta}^*, F^*} h_2(\boldsymbol{X}) \Sigma_{\boldsymbol{Y}}^{-\frac{1}{2}} (\boldsymbol{Y} - \boldsymbol{\mu}) | \boldsymbol{X}\right] \\ &= \mathbb{E}_{\boldsymbol{\beta}^*, F^*} B_{\boldsymbol{\beta}^*, F^*} h_2(\boldsymbol{X}) \left[h_1(\boldsymbol{Y}) | \boldsymbol{X}\right] \\ &\quad + C_{\boldsymbol{\beta}^*, F^*} h_2(\boldsymbol{X}) \Sigma_{\boldsymbol{Y}}^{-\frac{1}{2}} \mathbb{E}_{\boldsymbol{\beta}^*, F^*}\left[h_1(\boldsymbol{Y}) (\boldsymbol{Y} - \boldsymbol{\mu}) | \boldsymbol{X}\right] \\ &= B_{\boldsymbol{\beta}^*, F^*} h_1(\boldsymbol{X}) B_{\boldsymbol{\beta}^*, F^*} h_2(\boldsymbol{X}) + C_{\boldsymbol{\beta}^*, F^*} h_2(\boldsymbol{X}) \left(C_{\boldsymbol{\beta}^*, F^*} h_1(\boldsymbol{X})\right)^T \\ C_{\boldsymbol{\beta}^*, F^*} h_2(\boldsymbol{X}) \left(C_{\boldsymbol{\beta}^*, F^*} h_1(\boldsymbol{X})\right)^T &= C_{\boldsymbol{\beta}^*, F^*} h_1(\boldsymbol{X}) \left(C_{\boldsymbol{\beta}^*, F^*} h_2(\boldsymbol{X})\right)^T\end{aligned}$$

Subbing the above into our covariance expression.

$$\begin{aligned}\operatorname{Cov}(h_1(\boldsymbol{Y}), g_2(\boldsymbol{X}, \boldsymbol{Y}) | \boldsymbol{X}) &= \mathbb{E}_{\boldsymbol{\beta}^*, F^*}\left[h_1(\boldsymbol{Y}) g_2(\boldsymbol{X}, \boldsymbol{Y}) | \boldsymbol{X}\right] - B_{\boldsymbol{\beta}^*, F^*} h_1(\boldsymbol{X}) \mathbb{E}_{\boldsymbol{\beta}^*, F^*}\left[g_2(\boldsymbol{X}, \boldsymbol{Y}) | \boldsymbol{X}\right] \\ &= C_{\boldsymbol{\beta}^*, F^*} h_1(\boldsymbol{X}) \left(C_{\boldsymbol{\beta}^*, F^*} h_2(\boldsymbol{X})\right)^T\end{aligned} \tag{5.46}$$

Similarly, we can evaluate the third term $\operatorname{Cov}(h_2(\boldsymbol{Y}), g_1(\boldsymbol{X}, \boldsymbol{Y}) | \boldsymbol{X})$ as

$$\operatorname{Cov}(h_2(\boldsymbol{Y}), g_1(\boldsymbol{X}, \boldsymbol{Y}) | \boldsymbol{X}) = C_{\boldsymbol{\beta}^*, F^*} h_1(\boldsymbol{X}) \left(C_{\boldsymbol{\beta}^*, F^*} h_2(\boldsymbol{X})\right)^T \tag{5.47}$$

The final term $\operatorname{Cov}(g_1(\boldsymbol{X}, \boldsymbol{Y}), g_2(\boldsymbol{X}, \boldsymbol{Y}) | \boldsymbol{X})$ can be found through similar calculations using the following identity for $j = 1, 2$

$$C_{\boldsymbol{\beta}^*, F^*} h_j(\boldsymbol{X}) \Sigma_{Y}^{-\frac{1}{2}} (\boldsymbol{Y} - \boldsymbol{\mu}) = (\boldsymbol{Y} - \boldsymbol{\mu})^T \Sigma_{\boldsymbol{Y}}^{-\frac{1}{2}} \left(C_{\boldsymbol{\beta}^*, F^*} h_j(\boldsymbol{X})\right)^T \tag{5.48}$$

Thus,

$$\operatorname{Cov}(g_1(\boldsymbol{X}, \boldsymbol{Y}), g_2(\boldsymbol{X}, \boldsymbol{Y}) | \boldsymbol{X}) = C_{\boldsymbol{\beta}^*, F^*} h_1(\boldsymbol{X}) \left(C_{\boldsymbol{\beta}^*, F^*} h_2(\boldsymbol{X})\right)^T \tag{5.49}$$

Bringing together (5.46), (5.47), (5.49), into the law of total covariance,

$$\begin{aligned}\operatorname{Cov}\left(A_{\boldsymbol{\beta}^*, F^*} h_1(\boldsymbol{X}, \boldsymbol{Y}), A_{\boldsymbol{\beta}^*, F^*} h_2(\boldsymbol{X}, \boldsymbol{Y})\right) &= \mathbb{E}^{\mathcal{X}}\left[\operatorname{Cov}(h_1(\boldsymbol{Y}), h_2(\boldsymbol{Y}) | \boldsymbol{X}) - C_{\boldsymbol{\beta}^*, F^*} h_1(\boldsymbol{X}) \left(C_{\boldsymbol{\beta}^*, F^*} h_2(\boldsymbol{X})\right)^T\right] \\ &=: W_F(h_1, h_2)\end{aligned} \tag{5.50}$$

Finally, considering the covariance between the two components, for any $h \in \mathcal{H}_L$

$$\operatorname{Cov}\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} S_{\boldsymbol{\beta}^*, F^*}(\boldsymbol{X}_i, \boldsymbol{Y}_i), \frac{1}{\sqrt{n}} \sum_{i=1}^{n} A_{\boldsymbol{\beta}^*, F^*} h_2(\boldsymbol{X}_i, \boldsymbol{Y}_i)\right) = \operatorname{Cov}\left(S_{\boldsymbol{\beta}^*, F^*}(\boldsymbol{X}, \boldsymbol{Y}), A_{\boldsymbol{\beta}^*, F^*} h_2(\boldsymbol{X}, \boldsymbol{Y})\right)$$

By the law of total covariance, and noting the conditional expectation of the score functions is 0.

$$\begin{aligned}\text{Cov}\left(S_{\boldsymbol{\beta}^*,F^*}(\boldsymbol{X},\boldsymbol{Y}), A_{\boldsymbol{\beta}^*,F^*}h(\boldsymbol{X},\boldsymbol{Y})\right) &= \mathbb{E}^{\mathcal{X}}\left[\text{Cov}\left(S_{\boldsymbol{\beta}^*,F^*}(\boldsymbol{X},\boldsymbol{Y}), A_{\boldsymbol{\beta}^*,F^*}h(\boldsymbol{X},\boldsymbol{Y})|\boldsymbol{X}\right)\right] \\ &\quad + \text{Cov}\left[\mathbb{E}\left(S_{\boldsymbol{\beta}^*,F^*}(\boldsymbol{X},\boldsymbol{Y})|\boldsymbol{X}\right)\left(A_{\boldsymbol{\beta}^*,F^*}h(\boldsymbol{X},\boldsymbol{Y})|\boldsymbol{X}\right)\right] \\ &= \mathbb{E}^{\mathcal{X}}\left[\text{Cov}\left(S_{\boldsymbol{\beta}^*,F^*}(\boldsymbol{X},\boldsymbol{Y}), A_{\boldsymbol{\beta}^*,F^*}h(\boldsymbol{X},\boldsymbol{Y})|\boldsymbol{X}\right)\right]\end{aligned}$$

Considering the term $\text{Cov}\left(S_{\boldsymbol{\beta}^*,F^*}(\boldsymbol{X},\boldsymbol{Y}), A_{\boldsymbol{\beta}^*,F^*}h(\boldsymbol{X},\boldsymbol{Y})|\boldsymbol{X}\right)$ and using (5.48),

$$\begin{aligned}&= \mathbb{E}_{\boldsymbol{\beta}^*,F^*}\left[S_{\boldsymbol{\beta}^*,F^*}(\boldsymbol{X},\boldsymbol{Y})A_{\boldsymbol{\beta}^*,F^*}h(\boldsymbol{X},\boldsymbol{Y})|\boldsymbol{X}\right] \\ &= D(\boldsymbol{X};\boldsymbol{\beta}^*)\Sigma_{\boldsymbol{Y}}^{-1}\mathbb{E}_{\boldsymbol{\beta}^*,F^*}\left[h(\boldsymbol{Y})(\boldsymbol{Y}-\boldsymbol{\mu})|\boldsymbol{X}\right] - D(\boldsymbol{X};\boldsymbol{\beta}^*)\Sigma_{\boldsymbol{Y}}^{-1}B_{\boldsymbol{\beta}^*,F^*}h(\boldsymbol{X})\mathbb{E}_{\boldsymbol{\beta}^*,F^*}\left[(\boldsymbol{Y}-\boldsymbol{\mu})|\boldsymbol{X}\right] \\ &- \mathbb{E}_{\boldsymbol{\beta}^*,F^*}\left[D(\boldsymbol{X};\boldsymbol{\beta}^*)\Sigma_{\boldsymbol{Y}}^{-1}(\boldsymbol{Y}-\boldsymbol{\mu})(\boldsymbol{Y}-\boldsymbol{\mu})^T\Sigma_Y^{-\frac{1}{2}}C_{\boldsymbol{\beta}^*,F^*}h(\boldsymbol{X})^T|\boldsymbol{X}\right] \\ &= D(\boldsymbol{X};\boldsymbol{\beta}^*)\Sigma_{\boldsymbol{Y}}^{-1}\mathbb{E}_{\boldsymbol{\beta}^*,F^*}\left[h(\boldsymbol{Y})(\boldsymbol{Y}-\boldsymbol{\mu})|\boldsymbol{X}\right] - 0 - D(\boldsymbol{X};\boldsymbol{\beta}^*)\Sigma_{\boldsymbol{Y}}^{-1}\mathbb{E}_{\boldsymbol{\beta}^*,F^*}\left[h(\boldsymbol{Y})(\boldsymbol{Y}-\boldsymbol{\mu})|\boldsymbol{X}\right]\end{aligned}$$

Thus,

$$\text{Cov}\left(S_{\boldsymbol{\beta}^*,F^*}(\boldsymbol{X},\boldsymbol{Y}), A_{\boldsymbol{\beta}^*,F^*}h(\boldsymbol{X},\boldsymbol{Y})|\boldsymbol{X}\right) = 0\,,$$

which implies that

$$\text{Cov}\left(S_{\boldsymbol{\beta}^*,F^*}(\boldsymbol{X},\boldsymbol{Y}), A_{\boldsymbol{\beta}^*,F^*}h(\boldsymbol{X},\boldsymbol{Y})\right) = 0\,. \tag{5.51}$$

The two score functions were shown to be orthogonal in Lemma 2.1, so we expect the covariance between the score functions to be 0. This shows that $\boldsymbol{Z}$ contains two independent components, $Z_1$ which is a mean zero multivariate Gaussian with covariance matrix $W_{\boldsymbol{\beta}}$, and $Z_2$ is a mean zero Gaussian random process with covariance function $W_F(h_1,h_2)$ where $h_1,h_2 \in \mathcal{H}_L$

### 5.3.3 Fréchet Derivative

Now that we have shown (5.13), (5.14), the next step is to explicitly find the Fréchet derivative $\dot{\Psi}$ of $\Psi$ at $(\boldsymbol{\beta}^*,F^*)$. A Fréchet derivative $\dot{\Psi}$ is a linear map defined by the requirement that as $\boldsymbol{\beta}\to\boldsymbol{\beta},F\to F^*$,

$$||\Psi(\boldsymbol{\beta},F) - \Psi(\boldsymbol{\beta}^*,F) - \dot{\Psi}(\boldsymbol{\beta}-\boldsymbol{\beta}^*,F-F^*)|| = o(||\boldsymbol{\beta}-\boldsymbol{\beta}^*,F-F^*||)\,,$$

where

$$\dot{\Psi}(\boldsymbol{\beta}-\boldsymbol{\beta}^*,F-F^*) = \begin{pmatrix}\dot{\Psi}_{11} & \dot{\Psi}_{12}\\ \dot{\Psi}_{21} & \dot{\Psi}_{22}\end{pmatrix}\begin{pmatrix}\boldsymbol{\beta}-\boldsymbol{\beta}^*\\ F-F^*\end{pmatrix}\,. \tag{5.52}$$

The Fréchet derivative can be thought of as an analogue to a first-order Taylor expansion in the finite-dimensional case. Thus, let us consider the first component $\Psi(\boldsymbol{\beta},F)-\Psi(\boldsymbol{\beta}^*,F^*)$. Note that under the true parameters $(\boldsymbol{\beta}^*,F^*)$,

$$\begin{aligned}P^*S_{\boldsymbol{\beta}^*,F^*} &= \mathbb{E}_{\boldsymbol{\beta}^*,F^*}\left[D(\boldsymbol{X};\boldsymbol{\beta}^*)\Sigma_{\boldsymbol{Y}}^{-1}(\boldsymbol{X};\boldsymbol{\beta}^*,F^*)\left(\boldsymbol{Y}-\boldsymbol{\mu}(\boldsymbol{X}^T\boldsymbol{\beta}^*)\right)\right] \\ &= \mathbb{E}^{\mathcal{X}}\left[D(\boldsymbol{X};\boldsymbol{\beta}^*)\Sigma_{\boldsymbol{Y}}^{-1}(\boldsymbol{X};\boldsymbol{\beta}^*,F^*)\mathbb{E}_{\boldsymbol{\beta}^*,F^*}\left[\left(\boldsymbol{Y}-\boldsymbol{\mu}(\boldsymbol{X}^T\boldsymbol{\beta}^*)\right)|\boldsymbol{X}\right]\right] \\ &= \mathbb{E}^{\mathcal{X}}\left[D(\boldsymbol{X};\boldsymbol{\beta}^*)\Sigma_{\boldsymbol{Y}}^{-1}(\boldsymbol{X};\boldsymbol{\beta}^*,F^*)\left(\boldsymbol{\mu}(\boldsymbol{X}^T\boldsymbol{\beta}^*)-\boldsymbol{\mu}(\boldsymbol{X}^T\boldsymbol{\beta}^*)\right)\right] \\ &= 0\,.\end{aligned} \tag{5.53}$$

Using (5.53), we have

$$\begin{aligned}\Psi_1(\boldsymbol{\beta},F)-\Psi_1(\boldsymbol{\beta}^*,F^*) &= P^*S_{\boldsymbol{\beta},F} - P^*S_{\boldsymbol{\beta}^*,F^*} \\ &= P^*S_{\boldsymbol{\beta},F} \\ &= \mathbb{E}_{\boldsymbol{\beta}^*,F^*}\left[D(\boldsymbol{X};\boldsymbol{\beta})\Sigma_{\boldsymbol{Y}}^{-1}(\boldsymbol{X};\boldsymbol{\beta},F)\left(\boldsymbol{Y}-\boldsymbol{\mu}(\boldsymbol{X}^T\boldsymbol{\beta})\right)\right] \\ &= \mathbb{E}^{\mathcal{X}}\left[D(\boldsymbol{X};\boldsymbol{\beta}^*)\Sigma_{\boldsymbol{Y}}^{-1}(\boldsymbol{X};\boldsymbol{\beta}^*,F^*)\mathbb{E}_{\boldsymbol{\beta}^*,F^*}\left[\left(\boldsymbol{Y}-\boldsymbol{\mu}(\boldsymbol{X}^T\boldsymbol{\beta})\right)|\boldsymbol{X}\right]\right] \\ &= \mathbb{E}^{\mathcal{X}}\left[D(\boldsymbol{X};\boldsymbol{\beta}^*)\Sigma_{\boldsymbol{Y}}^{-1}(\boldsymbol{X};\boldsymbol{\beta}^*,F^*)\left(\boldsymbol{\mu}(\boldsymbol{X}^T\boldsymbol{\beta}^*)-\boldsymbol{\mu}(\boldsymbol{X}^T\boldsymbol{\beta})\right)\right]\,.\end{aligned}$$

We can approximate the expression above by taking that the difference in the score functions tends to 0 as $\boldsymbol{\beta} \to \boldsymbol{\beta^*}, F \to F^*$. This should depend on the rate that $\boldsymbol{\mu}(\boldsymbol{X}^T\boldsymbol{\beta^*}) - \boldsymbol{\mu}(\boldsymbol{X}^T\boldsymbol{\beta})$ tends to 0. We can consider a first-order Taylor expansion similar to (5.22), giving

$$\mu(\boldsymbol{X}^T\boldsymbol{\beta^*}) - \mu(\boldsymbol{X}^T\boldsymbol{\beta}) \approx -D(\boldsymbol{X};\boldsymbol{\beta^*})^T(\boldsymbol{\beta} - \boldsymbol{\beta^*}) \,. \tag{5.54}$$

Therefore, approximately we find

$$\Psi_1(\boldsymbol{\beta}, F) - \Psi_1(\boldsymbol{\beta^*}, F^*) \approx -\mathbb{E}^{\mathcal{X}}\left[D(\boldsymbol{X};\boldsymbol{\beta^*})\Sigma_{\boldsymbol{Y}}^{-1}(\boldsymbol{X};\boldsymbol{\beta^*}, F^*)D(\boldsymbol{X};\boldsymbol{\beta^*})^T\right](\boldsymbol{\beta} - \boldsymbol{\beta^*}) \,. \tag{5.55}$$

The above can be verified by taking partial derivatives and bounding second-order terms using all the Assumptions in A.2. By the definition of the Fréchet derivative, we find that as $\boldsymbol{\beta} \to \boldsymbol{\beta^*}, F \to F^*$

$$\frac{||\Psi_1(\boldsymbol{\beta}, F) - \Psi_1(\boldsymbol{\beta^*}, F^*) - \dot{\Psi}_{11}(\boldsymbol{\beta} - \boldsymbol{\beta^*}) - \dot{\Psi}_{12}(F - F^*)||}{||\boldsymbol{\beta} - \boldsymbol{\beta^*}, F - F^*||} \to 0 \,, \tag{5.56}$$

where from (5.55)

$$\dot{\Psi}_{11} = -\mathbb{E}^{\mathcal{X}}\left[D(\boldsymbol{X};\boldsymbol{\beta^*})\Sigma_{\boldsymbol{Y}}^{-1}(\boldsymbol{X};\boldsymbol{\beta^*}, F^*)D(\boldsymbol{X};\boldsymbol{\beta^*})^T\right] \tag{5.57}$$

$$\dot{\Psi}_{12} \equiv 0 \tag{5.58}$$

Now, let us consider the second component $\Psi(\boldsymbol{\beta}, F) - \Psi(\boldsymbol{\beta^*}, F^*)$. By similar working to (5.53), we have that $P^*A_{\boldsymbol{\beta^*},F^*}h = 0$, for some $h \in \mathcal{H}_L$.

$$\begin{aligned}
\Psi_2(\boldsymbol{\beta}, F) - \Psi_2(\boldsymbol{\beta^*}, F^*) &= P^*A_{\boldsymbol{\beta},F}h - P^*A_{\boldsymbol{\beta^*},F^*}h \\
&= \mathbb{E}_{\boldsymbol{\beta^*},F^*}\left[h(\boldsymbol{Y}) - B_{\boldsymbol{\beta},F}h(\boldsymbol{X}) - C_{\boldsymbol{\beta},F}\Sigma_{\boldsymbol{Y}}^{-1}(\boldsymbol{X})\left(\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{X}^T\boldsymbol{\beta})\right)\right] \\
&= \mathbb{E}^{\mathcal{X}}\left[\mathbb{E}_{\boldsymbol{\beta^*},F^*}\left[h(\boldsymbol{Y})|\boldsymbol{X}\right] - B_{\boldsymbol{\beta},F}h(\boldsymbol{X})\right] \\
&- \mathbb{E}^{\mathcal{X}}\left[C_{\boldsymbol{\beta},F}\Sigma_{\boldsymbol{Y}}^{-1}(\boldsymbol{X})\mathbb{E}_{\boldsymbol{\beta^*},F^*}\left[\left(\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{X}^T\boldsymbol{\beta})\right)|\boldsymbol{X}\right]\right] \\
&= \mathbb{E}^{\mathcal{X}}\left[B_{\boldsymbol{\beta^*},F^*}h(\boldsymbol{X}) - B_{\boldsymbol{\beta},F}h(\boldsymbol{X})\right] - \mathbb{E}^{\mathcal{X}}\left[C_{\boldsymbol{\beta},F}\Sigma_{\boldsymbol{Y}}^{-1}(\boldsymbol{X})\left(\boldsymbol{\mu}(\boldsymbol{X}^T\boldsymbol{\beta^*}) - \boldsymbol{\mu}(\boldsymbol{X}^T\boldsymbol{\beta})\right)\right]
\end{aligned}$$

Analogous to notation in Huang (2014), let

$$B^*_{\boldsymbol{\beta^*},F}g(\boldsymbol{y}) = \int_{\mathcal{X}} g(\boldsymbol{x}) \exp\left(b(\boldsymbol{x};\boldsymbol{\beta^*}, F^*) + \boldsymbol{\theta}(\boldsymbol{x};\boldsymbol{\beta}^*, F^*)^T\boldsymbol{y}\right) \mathrm{d}G_{\boldsymbol{X}}(\boldsymbol{x}) \tag{5.59}$$

$$C^*_{\boldsymbol{\beta^*},F^*}g(\boldsymbol{y}) = \int_{\mathcal{X}} g(\boldsymbol{x})\Sigma_{\boldsymbol{Y}}^{-\frac{1}{2}}\left(\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{X}^T\boldsymbol{\beta^*})\right) \exp\left(b(\boldsymbol{x};\boldsymbol{\beta^*}, F^*) + \boldsymbol{\theta}(\boldsymbol{x};\boldsymbol{\beta}^*, F^*)^T\boldsymbol{y}\right) \mathrm{d}G_{\boldsymbol{X}}(\boldsymbol{x}) \tag{5.60}$$

which are Hilbert Adjoints of $B_{\boldsymbol{\beta^*},F^*}$, $C_{\boldsymbol{\beta^*},F^*}$.

**Definition 5.1 (Hilbert Adjoint)** *Suppose $\mathcal{X}, \mathcal{Y}$ are Hilbert spaces with an inner product. Suppose a continuous bounded linear operator $\mathcal{A} : \mathcal{X} \to \mathcal{Y}$, then the adjoint operator $\mathcal{A}^* : \mathcal{Y}^* \to \mathcal{X}^*$ satisfies*

$$\langle \mathcal{A}^*y, x\rangle_{\mathcal{X}} = \langle y, \mathcal{A}x\rangle_{\mathcal{Y}}$$

*Where $y \in \mathcal{Y}, x \in \mathcal{X}$.*

For the second term, by the same argument given in (5.54) and (5.55), we informally find

$$\begin{aligned}
&\mathbb{E}^{\mathcal{X}}\left[C_{\boldsymbol{\beta},F}h(\boldsymbol{X})\Sigma_{\boldsymbol{Y}}^{-\frac{1}{2}}(\boldsymbol{X};\boldsymbol{\beta}, F)\left(\boldsymbol{\mu}(\boldsymbol{X}^T\boldsymbol{\beta^*}) - \mu(\boldsymbol{X}^T\boldsymbol{\beta})\right)\right] \\
&\approx -\mathbb{E}^{\mathcal{X}}\left[C_{\boldsymbol{\beta}^*,F^*}h(\boldsymbol{X})\Sigma_{\boldsymbol{Y}}^{-\frac{1}{2}}(\boldsymbol{X};\boldsymbol{\beta^*}, F^*)D(\boldsymbol{X};\boldsymbol{\beta^*})^T\right](\boldsymbol{\beta} - \boldsymbol{\beta^*}) \,.
\end{aligned} \tag{5.61}$$

For the first term, we will use a relationship between directional derivatives and Fréchet derivatives. To make this clear, let us consider an equivalent definition of a Fréchet derivative from Abraham et al. (1988).

**Definition 5.2 (Fréchet Derivative)** *For $f : \mathcal{U} \subset \mathcal{E} \to \mathcal{F}$ and $u_0 \in \mathcal{U}$ there is at most one bounded linear operator $L \in L(\mathcal{E}, \mathcal{F})$ such that the map $g : \mathcal{U} \subset \mathcal{E} \to \mathcal{F}$ given by $g(u) = f(u_0) + L(u - u_0)$ is tangent to $f$ at $u_0$. If there exists $L$ which is unique, we say $f$ is Fréchet differentiable at $u_0$ and define the derivative of $f$ at $u_0$ to be $\boldsymbol{D}f(u_0) = L$. The evaluation of $\boldsymbol{D}f(u_0)$ on $e \in \mathcal{E}$ is denoted $\boldsymbol{D}f(u_0) \cdot e$ and the map*

$$\boldsymbol{D}f : \mathcal{U} \to L(\mathcal{E}, \mathcal{F}); u \mapsto \boldsymbol{D}f(u)$$

*is the derivative of $f$ if it is differentiable at each $u_0 \in \mathcal{U}$. The partial derivative with respect to some parameter $x$ is denoted by $\boldsymbol{D}_x f$.*

**Definition 5.3 (Directional Derivative)** *Let $f : \mathcal{U} \subset \mathcal{E} \to \mathcal{F}$ and let $u \in \mathcal{U}$. We say that $f$ has a derivative in the direction $e \in \mathcal{E}$ at $u$ if*

$$\left.\frac{d}{dt} f(u + te)\right|_{t=0}$$

*exists. We call this element of $\mathcal{F}$ the directional derivative of $f$ in the direction $e$ at $u$.*

Thus, by Proposition 2.4.6 from Abraham et al. (1988), if $f$ is differentiable at $u$, then the directional derivatives of $f$ exist at $u$ and is given by

$$\left.\frac{d}{dt} f(u + te)\right|_{t=0} = \boldsymbol{D}f(u) \cdot e$$

This is important as it shows to use the sub-model $u_t = u + te$ when evaluating the directional derivative.

Performing a first-order Taylor series expansion of the expression inside the expectation about $(\boldsymbol{\beta^*}, F^*)$,

$$\begin{aligned} B_{\boldsymbol{\beta},F} h(X) - B_{\boldsymbol{\beta}^*,F^*} h(X) &\approx \left(\boldsymbol{D} B_{\boldsymbol{\beta}^*,F^*} h(X)\right)^T \begin{pmatrix} \boldsymbol{\beta} - \boldsymbol{\beta^*} \\ F - F^* \end{pmatrix} \\ &= \left(\boldsymbol{D}_{\boldsymbol{\beta}} B_{\boldsymbol{\beta^*},F^*} h(X)\right)^T \cdot (\boldsymbol{\beta} - \boldsymbol{\beta}^*) + \boldsymbol{D}_F B_{\boldsymbol{\beta}^*,F^*} h(X) \cdot (F - F^*) \\ &= \left[\frac{\partial}{\partial \boldsymbol{\beta}} B_{\boldsymbol{\beta},F} h(X)\right]^T_{\boldsymbol{\beta}=\boldsymbol{\beta^*},F=F^*} (\boldsymbol{\beta} - \boldsymbol{\beta^*}) + \left.\frac{\partial}{\partial t} B_{\boldsymbol{\beta^*},F_t} h(X)\right|_{t=0}, \end{aligned} \tag{5.62}$$

whereby Proposition 2.4.6 we have the sub-model

$$F_t = F^* + t(F - F^*) \, .$$

Evaluating the partial derivative with respect to $\boldsymbol{\beta}$,

$$\begin{aligned} \left[\frac{\partial}{\partial \boldsymbol{\beta}} B_{\boldsymbol{\beta},F} h(X)\right]^T &= \left[\frac{\partial}{\partial \boldsymbol{\beta}} \int_{\mathcal{Y}} h(\boldsymbol{y}) \exp\left(b + \boldsymbol{\theta}^T \boldsymbol{y}\right) \mathrm{d}F(\boldsymbol{y})\right]^T \\ &= \left[\int_{\mathcal{Y}} h(\boldsymbol{y}) \left(\frac{\partial b}{\partial \boldsymbol{\beta}} + \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\beta}} \boldsymbol{y}\right) \exp\left(b + \boldsymbol{\theta}^T \boldsymbol{y}\right) \mathrm{d}F(\boldsymbol{y})\right]^T \\ &= \left[\int_{\mathcal{Y}} h(\boldsymbol{y}) \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\beta}} (\boldsymbol{y} - \boldsymbol{\mu}) \exp\left(b + \boldsymbol{\theta}^T \boldsymbol{y}\right) \mathrm{d}F(\boldsymbol{y})\right]^T \\ &= \left[\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\beta}} \int_{\mathcal{Y}} h(\boldsymbol{y}) (\boldsymbol{y} - \boldsymbol{\mu}) \exp\left(b + \boldsymbol{\theta}^T \boldsymbol{y}\right) \mathrm{d}F(\boldsymbol{y})\right]^T \\ &= \left[\int_{\mathcal{Y}} h(\boldsymbol{y}) (\boldsymbol{y} - \boldsymbol{\mu}) \exp\left(b + \boldsymbol{\theta}^T \boldsymbol{y}\right) \mathrm{d}F(\boldsymbol{y})\right]^T \left[\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\beta}}\right]^T \\ &= \mathbb{E}\left(h(\boldsymbol{Y})(\boldsymbol{Y} - \boldsymbol{\mu}) | \boldsymbol{X}\right)^T \Sigma_Y^{-1}(\boldsymbol{X}; \boldsymbol{\beta}, F) \boldsymbol{D}(\boldsymbol{X}; \boldsymbol{\beta})^T \\ &= C_{\boldsymbol{\beta},F} h(\boldsymbol{X}) \Sigma_{\boldsymbol{Y}}^{-\frac{1}{2}}(\boldsymbol{X}; \boldsymbol{\beta}, F) \boldsymbol{D}(\boldsymbol{X}; \boldsymbol{\beta})^T. \end{aligned}$$

As a result,

$$\left[\frac{\partial}{\partial \boldsymbol{\beta}} B_{\boldsymbol{\beta},F} h(\boldsymbol{X})\right]_{\boldsymbol{\beta}=\boldsymbol{\beta}^*,F=F^*}^{T} (\boldsymbol{\beta}-\boldsymbol{\beta}^*) = C_{\boldsymbol{\beta}^*,F^*} h(\boldsymbol{X}) \Sigma_{\boldsymbol{Y}}^{-\frac{1}{2}}(\boldsymbol{X};\boldsymbol{\beta}^*,F^*) \boldsymbol{D}(\boldsymbol{X};\boldsymbol{\beta}^*)^T (\boldsymbol{\beta}-\boldsymbol{\beta}^*) \tag{5.63}$$

$$\mathbb{E}^{\mathcal{X}}\left[\left[\frac{\partial}{\partial \boldsymbol{\beta}} B_{\boldsymbol{\beta},F} h(\boldsymbol{X})\right]_{\boldsymbol{\beta}=\boldsymbol{\beta}^*,F=F^*}^{T} (\boldsymbol{\beta}-\boldsymbol{\beta}^*)\right] = \mathbb{E}^{\mathcal{X}}\left[C_{\boldsymbol{\beta}^*,F^*} h(\boldsymbol{X}) \Sigma_{\boldsymbol{Y}}^{-\frac{1}{2}}(\boldsymbol{X};\boldsymbol{\beta}^*,F^*) \boldsymbol{D}(\boldsymbol{X};\boldsymbol{\beta}^*)^T\right] (\boldsymbol{\beta}-\boldsymbol{\beta}^*)\,, \tag{5.64}$$

where (5.64) cancels with (5.61). This leaves us with

$$\Psi_2(\boldsymbol{\beta},F) - \Psi_2(\boldsymbol{\beta}^*,F) \approx -\mathbb{E}^{\mathcal{X}}\left(\frac{\partial}{\partial t} B_{\boldsymbol{\beta}^*,F_t} h(\boldsymbol{X})\Big|_{t=0}\right) \tag{5.65}$$

For convenience let $b_t = b(\boldsymbol{X};\boldsymbol{\beta}^*,F_t), \boldsymbol{\theta}_t = \boldsymbol{\theta}(\boldsymbol{X};\boldsymbol{\beta}^*,F_t)$ and $b^* = b(\boldsymbol{X};\boldsymbol{\beta}^*,F^*), \boldsymbol{\theta}^* = \boldsymbol{\theta}(\boldsymbol{X};\boldsymbol{\beta}^*,F^*), \boldsymbol{\mu}^* = \boldsymbol{\mu}(\boldsymbol{X}^T\boldsymbol{\beta}^*), \Sigma_{\boldsymbol{Y}}^{-1^*} = \Sigma_{\boldsymbol{Y}}^{-1}(\boldsymbol{X};\boldsymbol{\beta}^*,F^*)$, and $\boldsymbol{y}_1$ be a dummy variable.

To evaluate the partial derivative with respect to $t$ along our submodel, fix some $\boldsymbol{X} \in \mathcal{X}$. Using submodel $F_t$ suppose $F_t, F^*, F-F^*$ have some common dominating measure $\lambda$, and densities $dF_t, dF^*, d(F-F^*)$ respectively. Then for any measurable function $g$,

$$\begin{aligned}
\int_{\mathcal{Y}} g(\boldsymbol{y}) \mathrm{d}F_t(\boldsymbol{y}) &= \int_{\mathcal{Y}} g(\boldsymbol{y})\, \mathrm{d}F^*(\boldsymbol{y}) + \int_{\mathcal{Y}} g(\boldsymbol{y}) t\, \mathrm{d}(F-F^*)(\boldsymbol{y}) && (5.66)\\
&= \int_{\mathcal{Y}} g(\boldsymbol{y}) dF^*(\boldsymbol{y})\, \mathrm{d}\lambda(\boldsymbol{y}) + \int_{\mathcal{Y}} g(\boldsymbol{y}) t\, d(F-F^*)(\boldsymbol{y})\, \mathrm{d}\lambda(\boldsymbol{y}) && (5.67)\\
&= \int_{\mathcal{Y}} g(\boldsymbol{y}) \left(dF^*(\boldsymbol{y}) + t d(F-F^*)(\boldsymbol{y})\right) \mathrm{d}\lambda(\boldsymbol{y}) && (5.68)\\
&= \int_{\mathcal{Y}} g(\boldsymbol{y}) \left(1 + t\frac{d(F-F^*)(\boldsymbol{y})}{dF^*(\boldsymbol{y})}\right) dF^*(\boldsymbol{y})\, \mathrm{d}\lambda(\boldsymbol{y}) && (5.69)\\
&= \int_{\mathcal{Y}} g(\boldsymbol{y}) \left(1 + t\frac{d(F-F^*)(\boldsymbol{y})}{dF^*(\boldsymbol{y})}\right) \mathrm{d}F^*(\boldsymbol{y}) && (5.70)
\end{aligned}$$

Letting $\psi(\boldsymbol{y}) = \dfrac{d(F-F^*)(\boldsymbol{y})}{\mathrm{d}F^*(\boldsymbol{y})}$, we have that

$$\int_{\mathcal{Y}} g(\boldsymbol{y}) \mathrm{d}F_t(\boldsymbol{y}) = \int_{\mathcal{Y}} g(\boldsymbol{y}) \left(1 + t\psi(\boldsymbol{y})\right) \mathrm{d}F^*(\boldsymbol{y}) \tag{5.71}$$

It's useful to factorize the densities as it's in the same form as in Section 2.3.2, except the integral is with respect to $F^*$ and $\psi(\boldsymbol{y}) = h(\boldsymbol{y})$. Thus,

$$\begin{aligned}
\frac{\partial}{\partial t} B_{\boldsymbol{\beta}^*,F_t} h(\boldsymbol{X}) &= \frac{\partial}{\partial t} \int_{\mathcal{Y}} h(\boldsymbol{y}) \exp\left(b_t + \boldsymbol{\theta}_t^T \boldsymbol{y}\right)\left(1 + t\psi(\boldsymbol{y})\right) \mathrm{d}F^*(\boldsymbol{y})\\
&= \int_{\mathcal{Y}} h(\boldsymbol{y}) \exp\left(b_t + \boldsymbol{\theta}_t^T \boldsymbol{y}\right) \psi(\boldsymbol{y})\, \mathrm{d}F^*(\boldsymbol{y}) && (5.72)\\
&+ \int_{\mathcal{Y}} h(\boldsymbol{y}) \left(\frac{\partial b_t}{\partial t} + \frac{\partial \boldsymbol{\theta}_t}{\partial t}\boldsymbol{y}\right) \exp\left(b_t + \boldsymbol{\theta}_t^T \boldsymbol{y}\right) \left(1 + t\psi(\boldsymbol{y})\right) \mathrm{d}F^*(\boldsymbol{y}) && (5.73)
\end{aligned}$$

We can find the partial derivatives with respect to $t$ by implicit differentiation as done in Section 2.3.2, subbing in $\psi$ for $h$ and noting that

$$\psi(\boldsymbol{y})\mathrm{d}F^*(\boldsymbol{y}) = \psi(\boldsymbol{y}) dF^*(\boldsymbol{y})\mathrm{d}\lambda(\boldsymbol{y}) = d(F-F^*)(\boldsymbol{y})\mathrm{d}\lambda(\boldsymbol{y}) = \mathrm{d}(F-F^*)(\boldsymbol{y}), \tag{5.74}$$

we find that

$$\left.\frac{\partial b_t}{\partial t}\right|_{t=0} = -\left.\frac{\partial \boldsymbol{\theta}_t}{\partial t}\right|_{t=0} \boldsymbol{\mu}^* - \int_{\mathcal{Y}} \exp\left(b^* + \boldsymbol{\theta}^{*T}\boldsymbol{y}_1\right) \mathrm{d}(F - F^*)(\boldsymbol{y}_1) \tag{5.75}$$

$$\left.\frac{\partial \boldsymbol{\theta}_t}{\partial t}\right|_{t=0} = -\left[\int_{\mathcal{Y}} (\boldsymbol{y}_1 - \boldsymbol{\mu}^*)^T \exp\left(b^* + \boldsymbol{\theta}^{*T}\boldsymbol{y}_1\right) \mathrm{d}(F - F^*)(\boldsymbol{y}_1)\right] \Sigma_Y^{-1}(\boldsymbol{X}; \boldsymbol{\beta}^*, F^*) \tag{5.76}$$

Thus, evaluating the partial derivatives at $t = 0$, we can express (5.72) as

$$\int_{\mathcal{Y}} h(\boldsymbol{y}) \exp\left(b_t|_{t=0} + (\boldsymbol{\theta}_t|_{t=0})^T\, \boldsymbol{y}\right)\psi(\boldsymbol{y})\, \mathrm{d}F^*(\boldsymbol{y}) = \int_{\mathcal{Y}} h(\boldsymbol{y}) \exp\left(b^* + \boldsymbol{\theta}^{*T}\boldsymbol{y}\right) \mathrm{d}(F - F^*)(\boldsymbol{y})\,. \tag{5.77}$$

Evaluating at $t = 0$, we can express (5.73) as

$$\begin{aligned}
&\int_{\mathcal{Y}} h(\boldsymbol{y}) \left(\left.\frac{\partial b_t}{\partial t}\right|_{t=0} + \left.\frac{\partial \boldsymbol{\theta}_t}{\partial t}\right|_{t=0} \boldsymbol{y}\right) \exp\left(b_t|_{t=0} + (\boldsymbol{\theta}_t|_{t=0})^T\, \boldsymbol{y}\right) (1 + 0 \cdot \psi(\boldsymbol{y}))\, \mathrm{d}F^*(\boldsymbol{y}) \\
&= \int_{\mathcal{Y}} h(\boldsymbol{y}) \left[\left.\frac{\partial \boldsymbol{\theta}_t}{\partial t}\right|_{t=0}\right] (\boldsymbol{y} - \boldsymbol{\mu}^*) \exp\left(b^* + \boldsymbol{\theta}^{*T}\boldsymbol{y}\right) \mathrm{d}F^*(\boldsymbol{y}) \qquad (5.78) \\
&- \int_{\mathcal{Y}} h(\boldsymbol{y}) \left[\int_{\mathcal{Y}} \exp\left(b^* + \boldsymbol{\theta}^{*T}\boldsymbol{y}_1\right) \mathrm{d}(F - F^*)(\boldsymbol{y}_1)\right] \exp\left(b^* + \boldsymbol{\theta}^{*T}\boldsymbol{y}\right) \mathrm{d}F^*(\boldsymbol{y}) \qquad (5.79)
\end{aligned}$$

Subbing in (5.76) into (5.78) and applying Fubini's Theorem to swap the order of integration,

$$\begin{aligned}
&\int_{\mathcal{Y}} h(\boldsymbol{y}) \left[\left.\frac{\partial \boldsymbol{\theta}_t}{\partial t}\right|_{t=0}\right] (\boldsymbol{y} - \boldsymbol{\mu}^*) \exp\left(b^* + \boldsymbol{\theta}^{*T}\boldsymbol{y}\right) \mathrm{d}F^*(\boldsymbol{y}) \\
&= -\int_{\mathcal{Y}} \int_{\mathcal{Y}} h(\boldsymbol{y})(\boldsymbol{y}_1 - \boldsymbol{\mu}^*)^T \Sigma_{\boldsymbol{Y}}^{-1^*} (\boldsymbol{y} - \boldsymbol{\mu}^*) \exp\left(b^* + \boldsymbol{\theta}^{*T}\boldsymbol{y}\right) \exp\left(b^* + \boldsymbol{\theta}^{*T}\boldsymbol{y}_1\right) \mathrm{d}(F - F^*)(\boldsymbol{y}_1)\, \mathrm{d}F^*(\boldsymbol{y}) \\
&= -\int_{\mathcal{Y}} \int_{\mathcal{Y}} h(\boldsymbol{y})(\boldsymbol{y}_1 - \boldsymbol{\mu}^*)^T \Sigma_{\boldsymbol{Y}}^{-1^*} (\boldsymbol{y} - \boldsymbol{\mu}^*) \exp\left(b^* + \boldsymbol{\theta}^{*T}\boldsymbol{y}\right) \exp\left(b^* + \boldsymbol{\theta}^{*T}\boldsymbol{y}_1\right) \mathrm{d}F^*(\boldsymbol{y})\, \mathrm{d}(F - F^*)(\boldsymbol{y}_1).
\end{aligned}$$

Note that

$$(\boldsymbol{y}_1 - \boldsymbol{\mu}^*)^T \Sigma_{\boldsymbol{Y}}^{-1^*} (\boldsymbol{y} - \boldsymbol{\mu}^*) = (\boldsymbol{y} - \boldsymbol{\mu}^*)^T \Sigma_{\boldsymbol{Y}}^{-1^*} (\boldsymbol{y}_1 - \boldsymbol{\mu}^*) = (\boldsymbol{y} - \boldsymbol{\mu}^*)^T \Sigma_{\boldsymbol{Y}}^{-\frac{1}{2}^*} \Sigma_{\boldsymbol{Y}}^{-\frac{1}{2}^*} (\boldsymbol{y}_1 - \boldsymbol{\mu}^*),$$

and the exponential function is a real-valued function. Thus, we can re-arrange (5.78) to become

$$\begin{aligned}
&\int_{\mathcal{Y}} h(\boldsymbol{y}) \left[\left.\frac{\partial \boldsymbol{\theta}_t}{\partial t}\right|_{t=0}\right] (\boldsymbol{y} - \boldsymbol{\mu}^*) \exp\left(b^* + \boldsymbol{\theta}^{*T}\boldsymbol{y}\right) \mathrm{d}F^*(\boldsymbol{y}) \\
&= -\int_{\mathcal{Y}} h(\boldsymbol{y}) \left[\int_{\mathcal{Y}} (\boldsymbol{y}_1 - \boldsymbol{\mu}^*)^T \exp\left(b^* + \boldsymbol{\theta}^{*T}\boldsymbol{y}_1\right) \mathrm{d}(F - F^*)(\boldsymbol{y}_1)\right] \Sigma_{\boldsymbol{Y}}^{-1}(\boldsymbol{X}; \boldsymbol{\beta}^*, F^*)\, (\boldsymbol{y} - \boldsymbol{\mu}^*) \exp\left(b^* + \boldsymbol{\theta}^{*T}\boldsymbol{y}\right) \mathrm{d}F^*(\boldsymbol{y}) \\
&= -\int_{\mathcal{Y}} \int_{\mathcal{Y}} h(\boldsymbol{y})(\boldsymbol{y}_1 - \boldsymbol{\mu}^*)^T \Sigma_{\boldsymbol{Y}}^{-1}(\boldsymbol{X}; \boldsymbol{\beta}^*, F^*)\, (\boldsymbol{y} - \boldsymbol{\mu}^*) \exp\left(b^* + \boldsymbol{\theta}^{*T}\boldsymbol{y}_1\right) \exp\left(b^* + \boldsymbol{\theta}^{*T}\boldsymbol{y}\right) \mathrm{d}(F - F^*)(\boldsymbol{y}_1)\, \mathrm{d}F^*(\boldsymbol{y}) \\
&= -\int_{\mathcal{Y}} \int_{\mathcal{Y}} h(\boldsymbol{y})(\boldsymbol{y} - \boldsymbol{\mu}^*)^T \Sigma_{\boldsymbol{Y}}^{-\frac{1}{2}^*} \Sigma_{\boldsymbol{Y}}^{-\frac{1}{2}^*} (\boldsymbol{y}_1 - \boldsymbol{\mu}^*) \exp\left(b^* + \boldsymbol{\theta}^{*T}\boldsymbol{y}_1\right) \exp\left(b^* + \boldsymbol{\theta}^{*T}\boldsymbol{y}\right) \mathrm{d}F^*(\boldsymbol{y})\, \mathrm{d}(F - F^*)(\boldsymbol{y}_1) \\
&= -\int_{\mathcal{Y}} \int_{\mathcal{Y}} h(\boldsymbol{y})(\boldsymbol{y} - \boldsymbol{\mu}^*)^T \Sigma_{\boldsymbol{Y}}^{-\frac{1}{2}^*} \exp\left(b^* + \boldsymbol{\theta}^{*T}\boldsymbol{y}\right) \Sigma_{\boldsymbol{Y}}^{-\frac{1}{2}^*} (\boldsymbol{y}_1 - \boldsymbol{\mu}^*) \exp\left(b^* + \boldsymbol{\theta}^{*T}\boldsymbol{y}_1\right) \mathrm{d}F^*(\boldsymbol{y})\, \mathrm{d}(F - F^*)(\boldsymbol{y}_1) \\
&= -\int_{\mathcal{Y}} \left[\int_{\mathcal{Y}} h(\boldsymbol{y})(\boldsymbol{y} - \boldsymbol{\mu}^*)^T \Sigma_{\boldsymbol{Y}}^{-\frac{1}{2}^*} \exp\left(b^* + \boldsymbol{\theta}^{*T}\boldsymbol{y}\right) \mathrm{d}F^*(\boldsymbol{y})\right] \Sigma_{\boldsymbol{Y}}^{-\frac{1}{2}^*} (\boldsymbol{y}_1 - \boldsymbol{\mu}^*) \exp\left(b^* + \boldsymbol{\theta}^{*T}\boldsymbol{y}_1\right) \mathrm{d}(F - F^*)(\boldsymbol{y}_1) \\
&= -\int_{\mathcal{Y}} C_{\boldsymbol{\beta}^*, F^*} h(\boldsymbol{x}) \Sigma_{\boldsymbol{Y}}^{-\frac{1}{2}^*} (\boldsymbol{y}_1 - \boldsymbol{\mu}^*) \exp\left(b^* + \boldsymbol{\theta}^{*T}\boldsymbol{y}_1\right) \mathrm{d}(F - F^*)(\boldsymbol{y}_1) \qquad (5.80)
\end{aligned}$$

Applying Fubini's Theorem to (5.79) to swap the order of integration, the expression becomes

$$\begin{aligned}
&\int_{\mathcal{Y}} h(\boldsymbol{y}) \left[ \int_{\mathcal{Y}} \exp\left(b^* + \boldsymbol{\theta}^{*^T} \boldsymbol{y}_1\right) \mathrm{d}(F - F^*)(\boldsymbol{y}_1) \right] \exp\left(b^* + \boldsymbol{\theta}^{*^T} \boldsymbol{y}\right) \mathrm{d}F^*(\boldsymbol{y}) \\
&= \int_{\mathcal{Y}} \int_{\mathcal{Y}} h(\boldsymbol{y}) \exp\left(b^* + \boldsymbol{\theta}^{*^T} \boldsymbol{y}_1\right) \exp\left(b^* + \boldsymbol{\theta}^{*^T} \boldsymbol{y}\right) \mathrm{d}(F - F^*)(\boldsymbol{y}_1) \, \mathrm{d}F^*(\boldsymbol{y}) \\
&= \int_{\mathcal{Y}} \int_{\mathcal{Y}} h(\boldsymbol{y}) \exp\left(b^* + \boldsymbol{\theta}^{*^T} \boldsymbol{y}_1\right) \exp\left(b^* + \boldsymbol{\theta}^{*^T} \boldsymbol{y}\right) \mathrm{d}F^*(\boldsymbol{y}) \, \mathrm{d}(F - F^*)(\boldsymbol{y}_1) \\
&= \int_{\mathcal{Y}} \left[ \int_{\mathcal{Y}} h(\boldsymbol{y}) \exp\left(b^* + \boldsymbol{\theta}^{*^T} \boldsymbol{y}\right) \mathrm{d}F^*(\boldsymbol{y}) \right] \exp\left(b^* + \boldsymbol{\theta}^{*^T} \boldsymbol{y}_1\right) \mathrm{d}(F - F^*)(\boldsymbol{y}_1) \\
&= \int_{\mathcal{Y}} B_{\boldsymbol{\beta}^*, F^*} h(\boldsymbol{x}) \exp\left(b^* + \boldsymbol{\theta}^{*^T} \boldsymbol{y}_1\right) \mathrm{d}(F - F^*)(\boldsymbol{y}_1) \qquad (5.81)
\end{aligned}$$

Noting that $\boldsymbol{y}_1$ is a dummy variable, changing this back to $\boldsymbol{y}$ and subbing (5.77) into (5.72) and (5.80), (5.81) into (5.73) with the evaluation at $t = 0$

$$\begin{aligned}
\frac{\partial}{\partial t} B_{\boldsymbol{\beta}, F_t} h(\boldsymbol{X}) \Big|_{t=0} &= \int_{\mathcal{Y}} h(\boldsymbol{y}) \exp\left(b^* + \boldsymbol{\theta}^{*^T} \boldsymbol{y}\right) \mathrm{d}(F - F^*)(\boldsymbol{y}) \qquad (5.82) \\
&\quad - \int_{\mathcal{Y}} B_{\boldsymbol{\beta}^*, F^*} h(\boldsymbol{x}) \exp\left(b^* + \boldsymbol{\theta}^{*^T} \boldsymbol{y}\right) \mathrm{d}(F - F^*)(\boldsymbol{y}) \qquad (5.83) \\
&\quad - \int_{\mathcal{Y}} C_{\boldsymbol{\beta}^*, F^*} h(\boldsymbol{x}) \Sigma_{\boldsymbol{Y}}^{-\frac{1}{2}^*} (\boldsymbol{y} - \boldsymbol{\mu}^*) \exp\left(b^* + \boldsymbol{\theta}^{*^T} \boldsymbol{y}\right) \mathrm{d}(F - F^*)(\boldsymbol{y}) \qquad (5.84)
\end{aligned}$$

Now it remains to take the expectation of the three terms above. Taking the expectation of (5.82) and applying Fubini's Theorem to change the order of integration,

$$\begin{aligned}
\mathbb{E}^{\mathcal{X}} \left[ \int_{\mathcal{Y}} h(\boldsymbol{y}) \exp\left(b^* + \boldsymbol{\theta}^{*^T} \boldsymbol{y}\right) \mathrm{d}(F - F^*)(\boldsymbol{y}) \right] &= \int_{\mathcal{X}} \int_{\mathcal{Y}} h(\boldsymbol{y}) \exp\left(b^* + \boldsymbol{\theta}^{*^T} \boldsymbol{y}\right) \mathrm{d}(F - F^*)(\boldsymbol{y}) \, \mathrm{d}G_{\boldsymbol{X}}(\boldsymbol{x}) \\
&= \int_{\mathcal{Y}} h(\boldsymbol{y}) \int_{\mathcal{X}} \exp\left(b^* + \boldsymbol{\theta}^{*^T} \boldsymbol{y}\right) \mathrm{d}G_{\boldsymbol{X}}(\boldsymbol{x}) \, \mathrm{d}(F - F^*)(\boldsymbol{y}) \\
&= \int_{\mathcal{Y}} h(\boldsymbol{y}) \omega(\boldsymbol{y}) \, \mathrm{d}(F - F^*)(\boldsymbol{y}) \, , \qquad (5.85)
\end{aligned}$$

where

$$\omega(\boldsymbol{y}) := \int_{\mathcal{X}} \exp\left(b^* + \boldsymbol{\theta}^{*^T} \boldsymbol{y}\right) \mathrm{d}G_{\boldsymbol{X}}(\boldsymbol{x}) \, . \qquad (5.86)$$

Taking the expectation of (5.83), applying Fubini's Theorem and using the Hilbert adjoint operator (5.59) where $g(\boldsymbol{x}) = B_{\boldsymbol{\beta}^*, F^*} h(\boldsymbol{x})$,

$$\begin{aligned}
&\mathbb{E}^{\mathcal{X}} \left[ \int_{\mathcal{Y}} B_{\boldsymbol{\beta}^*, F^*} h(\boldsymbol{x}) \exp\left(b^* + \boldsymbol{\theta}^{*^T} \boldsymbol{y}\right) \mathrm{d}(F - F^*)(\boldsymbol{y}) \right] \\
&= \int_{\mathcal{X}} \int_{\mathcal{Y}} B_{\boldsymbol{\beta}^*, F^*} h(\boldsymbol{x}) \exp\left(b^* + \boldsymbol{\theta}^{*^T} \boldsymbol{y}\right) \mathrm{d}(F - F^*)(\boldsymbol{y}) \, \mathrm{d}G_{\boldsymbol{X}}(\boldsymbol{x}) \\
&= \int_{\mathcal{Y}} \int_{\mathcal{X}} B_{\boldsymbol{\beta}^*, F^*} h(\boldsymbol{x}) \exp\left(b^* + \boldsymbol{\theta}^{*^T} \boldsymbol{y}\right) \mathrm{d}G_{\boldsymbol{X}}(\boldsymbol{x}) \, \mathrm{d}(F - F^*)(\boldsymbol{y}) \\
&= \int_{\mathcal{Y}} B^*_{\boldsymbol{\beta}^*, F^*} B_{\boldsymbol{\beta}^*, F^*} h(\boldsymbol{y}) \, \mathrm{d}(F - F^*)(\boldsymbol{y}) \, . \qquad (5.87)
\end{aligned}$$

Finally, taking the expectation of (5.84) applying Fubini's Theorem and using the Hilbert adjoint operator (5.60) where $g(\boldsymbol{x}) = C_{\boldsymbol{\beta}^*,F^*}h(\boldsymbol{x})$,

$$\begin{aligned}
&\mathbb{E}^{\mathcal{X}}\left[\int_{\mathcal{Y}} C_{\boldsymbol{\beta}^*,F^*}h(\boldsymbol{x})\Sigma_{\boldsymbol{Y}}^{-\frac{1}{2}^*}(\boldsymbol{y}-\boldsymbol{\mu}^*)\exp\left(b^*+\boldsymbol{\theta}^{*^T}\boldsymbol{y}\right)\mathrm{d}(F-F^*)(\boldsymbol{y})\right] \\
&= \int_{\mathcal{X}}\int_{\mathcal{Y}} C_{\boldsymbol{\beta}^*,F^*}h(\boldsymbol{x})\Sigma_{\boldsymbol{Y}}^{-\frac{1}{2}^*}(\boldsymbol{y}-\boldsymbol{\mu}^*)\exp\left(b^*+\boldsymbol{\theta}^{*^T}\boldsymbol{y}\right)\mathrm{d}(F-F^*)(\boldsymbol{y})\,\mathrm{d}G_{\boldsymbol{X}}(\boldsymbol{x}) \\
&= \int_{\mathcal{Y}}\int_{\mathcal{X}} C_{\boldsymbol{\beta}^*,F^*}h(\boldsymbol{x})\Sigma_{\boldsymbol{Y}}^{-\frac{1}{2}^*}(\boldsymbol{y}-\boldsymbol{\mu}^*)\exp\left(b^*+\boldsymbol{\theta}^{*^T}\boldsymbol{y}\right)\mathrm{d}G_{\boldsymbol{X}}(\boldsymbol{x})\,\mathrm{d}(F-F^*)(\boldsymbol{y}) \\
&= \int_{\mathcal{Y}} C^*_{\boldsymbol{\beta}^*,F^*}C_{\boldsymbol{\beta}^*,F^*}h(\boldsymbol{y})\,\mathrm{d}(F-F^*)(\boldsymbol{y})\,. \qquad (5.88)
\end{aligned}$$

Bringing together (5.85),(5.87),(5.88) and subbing into (5.65), we find that

$$\Psi_2(\boldsymbol{\beta},F)-\Psi_2(\boldsymbol{\beta^*},F)\approx -\int_{\mathcal{Y}}\left(\omega(\boldsymbol{y})-B^*_{\boldsymbol{\beta}^*,F^*}B_{\boldsymbol{\beta}^*,F^*}-C^*_{\boldsymbol{\beta}^*,F^*}C_{\boldsymbol{\beta}^*,F^*}\right)h(\boldsymbol{y})\,\mathrm{d}(F-F^*)(\boldsymbol{y})\,. \qquad (5.89)$$

We can formally check calculations by finding the first derivative and bounding the second derivatives as $\boldsymbol{\beta}\to\boldsymbol{\beta^*}, F\to F^*$, but the details are not insightful. Thus, by the definition of the Fréchet derivative, we find that as $\boldsymbol{\beta}\to\boldsymbol{\beta^*}, F\to F^*$,

$$\frac{||\Psi_2(\boldsymbol{\beta},F)-\Psi_2(\boldsymbol{\beta^*},F^*)-\dot{\Psi}_{21}(\boldsymbol{\beta}-\boldsymbol{\beta^*})-\dot{\Psi}_{22}(F-F^*)||}{||\boldsymbol{\beta}-\boldsymbol{\beta^*},F-F^*||}\to 0\,, \qquad (5.90)$$

where from (5.89)

$$\dot{\Psi}_{21}=0 \qquad (5.91)$$

$$\dot{\Psi}_{22}=-\int_{\mathcal{Y}}\left(\omega(\boldsymbol{y})-B^*_{\boldsymbol{\beta}^*,F^*}B_{\boldsymbol{\beta}^*,F^*}-C^*_{\boldsymbol{\beta}^*,F^*}C_{\boldsymbol{\beta}^*,F^*}\right)h(\boldsymbol{y})\,\mathrm{d}(F-F^*)(\boldsymbol{y})\,. \qquad (5.92)$$

Note that this is analogous to the expression found in the supplementary material of Huang (2014), but with all details derived and being able to handle vector responses.

### 5.3.4 Invertibility of Fréchet Derivative

To combine everything we have derived so far, in Lemma 5.1 we have shown that

$$\sqrt{n}\,(\Psi_n-\Psi)(\boldsymbol{\beta^*},F^*)\rightsquigarrow \boldsymbol{Z}$$

By the stochastic approximation (5.14) and the Fréchet derivative $\dot{\Psi}$ at $(\boldsymbol{\beta^*},F^*)$, means that we have shown that

$$-\sqrt{n}\begin{pmatrix}\dot{\Psi}_{11} & 0\\ 0 & \dot{\Psi}_{22}\end{pmatrix}\begin{pmatrix}\hat{\boldsymbol{\beta}}-\boldsymbol{\beta^*}\\ \hat{F}-F^*\end{pmatrix}\rightsquigarrow\begin{pmatrix}Z_1\\ Z_2\end{pmatrix}\,,$$

where $Z_1$ is a mean zero Gaussian random vector with covariance matrix $W_{\boldsymbol{\beta}}$ which is independent of $Z_2$, a mean zero Gaussian process with covariance function $W_F(h_1,h_2)$. Thus, if $\dot{\Psi}_{11}$ and $\dot{\Psi}_{22}$ are continuously invertible, we can conclude that

$$\sqrt{n}\begin{pmatrix}\hat{\boldsymbol{\beta}}-\boldsymbol{\beta^*}\\ \hat{F}-F^*\end{pmatrix}\rightsquigarrow\begin{pmatrix}-\dot{\Psi}_{11}^{-1}Z_1\\ -\dot{\Psi}_{22}^{-1}Z_2\end{pmatrix}\,.$$

Firstly, we require the sample measure $G_{\boldsymbol{X}}$ to be such that

$$\mathbf{A}(\boldsymbol{\beta^*},F^*):=\dot{\Psi}_{11}=-\mathbb{E}^{\mathcal{X}}\left[D(\boldsymbol{X};\boldsymbol{\beta^*})\Sigma_{\boldsymbol{Y}}^{-1}(\boldsymbol{X};\boldsymbol{\beta^*},F^*)D(\boldsymbol{X};\boldsymbol{\beta^*})^T\right] \qquad (5.93)$$

is invertible. We have that under most replications, that this holds which gives us

$$\mathbf{A}(\boldsymbol{\beta}^*, F^*)^{-1} = \left(-\mathbb{E}^{\mathcal{X}}\left[D(\boldsymbol{X};\boldsymbol{\beta}^*)\Sigma_{\boldsymbol{Y}}^{-1}(\boldsymbol{X};\boldsymbol{\beta}^*, F^*)D(\boldsymbol{X};\boldsymbol{\beta}^*)^T\right]\right)^{-1} . \tag{5.94}$$

Therefore, we have that

$$\mathrm{Cov}_{\boldsymbol{\beta}^*,F^*}(-\mathbf{A}(\boldsymbol{\beta}^*, F^*)^{-1}Z_1) = \mathbf{A}(\boldsymbol{\beta}^*, F^*)^{-1}\mathrm{Cov}_{\boldsymbol{\beta}^*,F^*}(Z_1)\mathbf{A}(\boldsymbol{\beta}^*, F^*)^{-1} .$$

Denoting $\mathbf{B}(\boldsymbol{\beta}^*, F^*) = \mathrm{Cov}_{\boldsymbol{\beta}^*,F^*}(Z_1)$, under the assumption that $F^*(\boldsymbol{y}|\boldsymbol{x})$ in the multivariate exponential family, we have that

$$\mathbf{B}(\boldsymbol{\beta}^*, F^*) = -\mathbf{A}(\boldsymbol{\beta}^*, F^*)$$

Therefore, we find that as $n \to \infty$ that

$$\sqrt{n}\,\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right) \to N(0, W_1)$$

where

$$W_1 := -\mathbf{A}(\boldsymbol{\beta}^*, F^*)^{-1}. \tag{5.95}$$

Now, let us consider the invertibility of $\dot{\Psi}_{22}$ where our argument follows similarly to that given in the supplementary material of Huang (2014), which is unique for the given problem. Firstly, we can re-express $\dot{\Psi}_{22}$ as

$$\dot{\Psi}_{22}(\hat{F} - F^*)h = -\int_{\mathcal{Y}} \omega(\boldsymbol{y})\left(I - \omega^{-1}(\boldsymbol{y})B^*_{\boldsymbol{\beta}^*,F^*}B_{\boldsymbol{\beta}^*,F^*} - w^{-1}(\boldsymbol{y})C^*_{\boldsymbol{\beta}^*,F^*}C_{\boldsymbol{\beta}^*,F^*}\right)h(\boldsymbol{y})\,\mathrm{d}(F - F^*)(\boldsymbol{y}) ,$$

where $\omega(\boldsymbol{y}) > 0$ for all $\boldsymbol{y}$ by construction, implying $w(\boldsymbol{y})$ is invertible. The expression above has a defined covariance expression of $W_F(h_1, h_2)$ for $h_1, h_2 \in \mathcal{H}_L$, so it would be useful for the limiting Gaussian process to have a similarly well-defined covariance expression. Thus, if there exists some continuous inverse operator $\mathcal{D}h : \mathcal{H}_L \mapsto \ell^\infty(\mathcal{H}_L)$ such that

$$\sqrt{n}\,(\hat{F} - F^*)h = \sqrt{n}\,\int_{\mathcal{Y}} \omega(\boldsymbol{y})\left(I - \omega^{-1}(\boldsymbol{y})B^*_{\boldsymbol{\beta}^*,F^*}B_{\boldsymbol{\beta}^*,F^*} - w^{-1}(\boldsymbol{y})C^*_{\boldsymbol{\beta}^*,F^*}C_{\boldsymbol{\beta}^*,F^*}\right)\mathcal{D}h(\boldsymbol{y})\,\mathrm{d}(F - F^*)(\boldsymbol{y})$$

then the sequence above converges in distribution in $\ell^\infty(\mathcal{H}_L)$ to a mean zero Gaussian process $Z_2(\mathcal{D}h)$ with covariance function $W_2(h_1, h_2) = W_F(\mathcal{D}h_1, \mathcal{D}h_2)$, provided $\mathcal{D}h$ is square-integrable with respect to $P^*$. Here $Z_2(\mathcal{D}h)$ can be interpreted as a Gaussian random process which is indexed by $\mathcal{D}h$ instead of $h$.

As reinforced by Huang (2011), invertibility is up to equivalence classes defined by the equivalence relation $h_1 \equiv h_2$ if and only if $h_1 = h_2 + \boldsymbol{a}\boldsymbol{y} + c$ where $\boldsymbol{a}$ is a vector of scalars and $c$ is a scalar. This is because $\dot{\Psi}_{22}(h) = \dot{\Psi}_{22}(F - F^*)(h + \boldsymbol{a}\boldsymbol{y} + c)$ which can be seen below.

$$\begin{aligned}
\dot{\Psi}_{22}(F - F^*)(h) &= -\int_{\mathcal{Y}} \left(\omega(\boldsymbol{y}) - B^*_{\boldsymbol{\beta}^*,F^*}B_{\boldsymbol{\beta}^*,F^*} - C^*_{\boldsymbol{\beta}^*,F^*}C_{\boldsymbol{\beta}^*,F^*}\right)h(\boldsymbol{y})\,\mathrm{d}(F - F^*)(\boldsymbol{y}) \\
\dot{\Psi}_{22}(F - F^*)(h + \boldsymbol{a}\boldsymbol{y} + c) &= -\int_{\mathcal{Y}} \left(\omega(\boldsymbol{y}) - B^*_{\boldsymbol{\beta}^*,F^*}B_{\boldsymbol{\beta}^*,F^*} - C^*_{\boldsymbol{\beta}^*,F^*}C_{\boldsymbol{\beta}^*,F^*}\right)(h(\boldsymbol{y}) + \boldsymbol{a}\boldsymbol{y} + c)\,\mathrm{d}(F - F^*)(\boldsymbol{y}) \\
&= \dot{\Psi}_{22}(F - F^*)(h) \\
&\quad - \int_{\mathcal{Y}} \left(\omega(\boldsymbol{y}) - B^*_{\boldsymbol{\beta}^*,F^*}B_{\boldsymbol{\beta}^*,F^*} - C^*_{\boldsymbol{\beta}^*,F^*}C_{\boldsymbol{\beta}^*,F^*}\right)(\boldsymbol{a}\boldsymbol{y} + c)\,\mathrm{d}(F - F^*)(\boldsymbol{y})
\end{aligned}$$

Simplifying above, note that

$$\begin{aligned}
B^*_{\boldsymbol{\beta}^*,F^*}B_{\boldsymbol{\beta}^*,F^*}(\boldsymbol{a}\boldsymbol{y} + c) &= \int_{\mathcal{X}}\left(\int_{\mathcal{Y}}(\boldsymbol{a}\boldsymbol{y}_1 + c)\exp\left(b^* + \boldsymbol{\theta}^{*T}\boldsymbol{y}\right)dF(\boldsymbol{y}_1)\right)\exp\left(b^* + \boldsymbol{\theta}^{*T}\boldsymbol{y}\right)\mathrm{d}G_X(\boldsymbol{x}) \\
&= \omega(\boldsymbol{y})\left(\boldsymbol{a}\boldsymbol{\mu}^* + c\right) ,
\end{aligned}$$

and,

$$
\begin{aligned}
&C^*_{\boldsymbol{\beta^*},F^*}C_{\boldsymbol{\beta^*},F^*}(\boldsymbol{a}\boldsymbol{y}+c)\\
&=\boldsymbol{a}\int_{\mathcal{X}}\left(\int_{\mathcal{Y}}\boldsymbol{y}_1(\boldsymbol{y}_1-\boldsymbol{\mu}^*)^T\Sigma_{\boldsymbol{Y}}^{-\frac{1}{2}^*}\exp\left(b^*+\boldsymbol{\theta}^{*T}\boldsymbol{y}_1\right)dF(\boldsymbol{y}_1)\right)\Sigma_{\boldsymbol{Y}}^{-\frac{1}{2}^*}(\boldsymbol{y}-\boldsymbol{\mu}^*)\exp\left(b^*+\boldsymbol{\theta}^{*T}\boldsymbol{y}\right)\mathrm{d}G_X(\boldsymbol{x})\\
&+c\int_{\mathcal{X}}\left(\int_{\mathcal{Y}}(\boldsymbol{y}_1-\boldsymbol{\mu}^*)^T\Sigma_{\boldsymbol{Y}}^{-\frac{1}{2}^*}\exp\left(b^*+\boldsymbol{\theta}^{*T}\boldsymbol{y}\right)dF(\boldsymbol{y}_1)\right)\Sigma_{\boldsymbol{Y}}^{-\frac{1}{2}^*}(\boldsymbol{y}-\boldsymbol{\mu}^*)\exp\left(b^*+\boldsymbol{\theta}^{*T}\boldsymbol{y}\right)\mathrm{d}G_X(\boldsymbol{x})\\
&=\boldsymbol{a}\int_{\mathcal{X}}(\boldsymbol{y}-\boldsymbol{\mu}^*)\exp\left(b^*+\boldsymbol{\theta}^{*T}\boldsymbol{y}\right)\mathrm{d}G_X(\boldsymbol{x})+c\int_{\mathcal{X}}0\cdot\Sigma_{\boldsymbol{Y}}^{-\frac{1}{2}^*}(\boldsymbol{y}-\boldsymbol{\mu}^*)\exp\left(b^*+\boldsymbol{\theta}^{*T}\boldsymbol{y}\right)\mathrm{d}G_X(\boldsymbol{x})\\
&=\boldsymbol{a}\omega(\boldsymbol{y})(\boldsymbol{y}-\boldsymbol{\mu}^*)\,.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\dot{\Psi}_{22}(F-F^*)(h+\boldsymbol{a}\boldsymbol{y}+c)&=\dot{\Psi}_{22}(F-F^*)(h)\\
&\quad-\int_{\mathcal{Y}}\omega(\boldsymbol{y})\left(\boldsymbol{a}\boldsymbol{y}+c\right)-\omega(\boldsymbol{y})(\boldsymbol{a}\boldsymbol{\mu}^*+c)-\omega(\boldsymbol{y})(\boldsymbol{a}\boldsymbol{y}-\boldsymbol{a}\boldsymbol{\mu}^*)\,\mathrm{d}(F-F^*)(\boldsymbol{y})\\
&=\dot{\Psi}_{22}(F-F^*)(h)
\end{aligned}
$$

Thus, the invertibility of $\dot{\Psi}_{22}$ is equivalent to the existence of this continuous operator $\mathcal{D}:\mathcal{H}_L\mapsto\ell^\infty(\mathcal{H}_L)$ that has the property

$$
\left(\omega-B^*_{\boldsymbol{\beta^*},F^*}B_{\boldsymbol{\beta^*},F^*}-C^*_{\boldsymbol{\beta^*},F^*}C_{\boldsymbol{\beta^*},F^*}\right)\mathcal{D}h=h\,, \tag{5.96}
$$

or equivalently

$$
\omega\left(I-\omega^{-1}B^*_{\boldsymbol{\beta^*},F^*}B_{\boldsymbol{\beta^*},F^*}-\omega^{-1}C^*_{\boldsymbol{\beta^*},F^*}C_{\boldsymbol{\beta^*},F^*}\right)\mathcal{D}h=h\,, \tag{5.97}
$$

as $\omega$ is invertible. Thus, the existence of the continuous inverse operator $\mathcal{D}h$ is equivalent to the operator

$$
\left(I-\omega^{-1}B^*_{\boldsymbol{\beta^*},F^*}B_{\boldsymbol{\beta^*},F^*}-\omega^{-1}C^*_{\boldsymbol{\beta^*},F^*}C_{\boldsymbol{\beta^*},F^*}\right)
$$

being continuously invertible. We use expression (5.97) so we can appeal to the following statement from Fredholm Theory as suggested by Van Der Vaart and Wellner (1996) and Huang (2014). Supposing that $(I-\mathcal{A})$ is a linear operator where $\mathcal{A}$ is a compact operator, if $(I-\mathcal{A})$ is injective up to a set an equivalence classes, then $(I-\mathcal{A})$ is surjective and has bounded inverse, implying that $(I-\mathcal{A})$ is continuously invertible.

Firstly, to show that $\omega^{-1}B^*_{\boldsymbol{\beta^*},F^*}B_{\boldsymbol{\beta^*},F^*}-\omega^{-1}C^*_{\boldsymbol{\beta^*},F^*}C_{\boldsymbol{\beta^*},F^*}$ is a compact operator, we will consider two key properties of compact operators. The first from Rudin (1973) is that if $\mathcal{E},\mathcal{F}$ are Banach spaces, and $\mathcal{R}:\mathcal{E}\to\mathcal{F}$ and $\mathcal{S}:\mathcal{E}\to\mathcal{F}$ are compact operators, then $\mathcal{R}+\mathcal{S}$ is a compact operator. The next is Proposition 3.1.14 in Zimmer (1990), which states that if $\mathcal{R}:\mathcal{E}\to\mathcal{F}$, $\mathcal{S}:\mathcal{F}\to\mathcal{G}$ are bounded linear maps between Banach spaces, then $\mathcal{S}\circ\mathcal{R}$ is compact if either $\mathcal{S}$ or $\mathcal{R}$ is compact. Thus, it suffices to show that $B^*_{\boldsymbol{\beta^*},F^*}$ and $C^*_{\boldsymbol{\beta^*},F^*}$ are compact operators.

From (5.59), we have that as $B^*_{\boldsymbol{\beta^*},F^*}$ is an integral operator defined by

$$
B^*_{\boldsymbol{\beta^*},F}g(\boldsymbol{y})=\int_{\mathcal{X}}g(\boldsymbol{x})\exp\left(b(\boldsymbol{x};\boldsymbol{\beta^*},F^*)+\boldsymbol{\theta}(\boldsymbol{x};\boldsymbol{\beta}^*,F^*)^T\boldsymbol{y}\right)\,\mathrm{d}G_{\boldsymbol{X}}(\boldsymbol{x})\,.
$$

Note that as $\mathcal{X}$ and $\mathcal{Y}$ are compact sets by A.2.1, $\exp\left(b(\boldsymbol{x};\boldsymbol{\beta^*},F^*)+\boldsymbol{\theta}(\boldsymbol{x};\boldsymbol{\beta}^*,F^*)^T\boldsymbol{y}\right)$ is continuous in $\boldsymbol{x},\boldsymbol{y}$ on $\mathcal{X}\times\mathcal{Y}$, and $G_{\boldsymbol{X}}$ has finite measure. Then by Theorem 3.5.1 in Zimmer (1990) (Definition 1.15), the operator $B^*_{\boldsymbol{\beta^*},F^*}$ is compact. Considering the integral operator $C^*_{\boldsymbol{\beta^*},F^*}$

$$
C^*_{\boldsymbol{\beta^*},F^*}g(\boldsymbol{y})=\int_{\mathcal{X}}g(\boldsymbol{x})\Sigma_{\boldsymbol{Y}}^{-\frac{1}{2}^*}\left(\boldsymbol{y}-\boldsymbol{\mu}(\boldsymbol{X}^T\boldsymbol{\beta^*})\right)\exp\left(b(\boldsymbol{x};\boldsymbol{\beta^*},F^*)+\boldsymbol{\theta}(\boldsymbol{x};\boldsymbol{\beta}^*,F^*)^T\boldsymbol{y}\right)\,\mathrm{d}G_{\boldsymbol{X}}(\boldsymbol{x})
$$

we have that

$$\Sigma_{\boldsymbol{Y}}^{-\frac{1}{2}^*}\left(\boldsymbol{y}-\boldsymbol{\mu}(\boldsymbol{X}^T\boldsymbol{\beta}^*)\right)\exp\left(b(\boldsymbol{x};\boldsymbol{\beta}^*,F^*)+\boldsymbol{\theta}(\boldsymbol{x};\boldsymbol{\beta}^*,F^*)^T\boldsymbol{y}\right)$$

is continuous in $\boldsymbol{x},\boldsymbol{y}$ on $\mathcal{X},\times\mathcal{Y}$ as each of it's $K$ components is continuous by A.2.2, A.2.3. Similarly, the operator $C^*_{\boldsymbol{\beta}^*,F^*}$ is compact implying that $\omega^{-1}B^*_{\boldsymbol{\beta}^*,F^*}B_{\boldsymbol{\beta}^*,F^*}-\omega^{-1}C^*_{\boldsymbol{\beta}^*,F^*}C_{\boldsymbol{\beta}^*,F^*}$ is a compact operator.

Next is to show that $(I-\omega^{-1}-B^*_{\boldsymbol{\beta}^*,F^*}B_{\boldsymbol{\beta}^*,F^*}-\omega^{-1}C^*_{\boldsymbol{\beta}^*,F^*}C_{\boldsymbol{\beta}^*,F^*})$ or equivalently $(\omega I-B^*_{\boldsymbol{\beta}^*,F^*}B_{\boldsymbol{\beta}^*,F^*}-C^*_{\boldsymbol{\beta}^*,F^*}C_{\boldsymbol{\beta}^*,F^*})$ is one-to-one up to equivalence classes of functions in $\ell^\infty(\mathcal{H}_L)$. We note that if $h_0$ is a solution to

$$\left(\omega I-B^*_{\boldsymbol{\beta}^*,F^*}B_{\boldsymbol{\beta}^*,F^*}-C^*_{\boldsymbol{\beta}^*,F^*}C_{\boldsymbol{\beta}^*,F^*}\right)h_0=0\,, \tag{5.98}$$

then our operator is injective if and only if $h_0\equiv 0$. This is because if $h_0\equiv 0$ is a trivial solution and if there were other non-equivalent $h_0$ which satisfy the equation, then by definition the operator is not one-to-one.

As seen in Huang (2014), by (5.98) it is true that

$$\begin{aligned}P^*\left[h_0\left(\omega I-B^*_{\boldsymbol{\beta}^*,F^*}B_{\boldsymbol{\beta}^*,F^*}-C^*_{\boldsymbol{\beta}^*,F^*}C_{\boldsymbol{\beta}^*,F^*}\right)h_0\right]&=0\\ P^*(h_0^2\omega)-P^*(h_0B^*_{\boldsymbol{\beta}^*,F^*}B_{\boldsymbol{\beta}^*,F^*}h_0)-P^*(h_0C^*_{\boldsymbol{\beta}^*,F^*}C_{\boldsymbol{\beta}^*,F^*}h_0)&=0\end{aligned} \tag{5.99}$$

Considering the first term of (5.99), applying Fubini's Theorem

$$\begin{aligned}P^*(h_0^2\omega)&=\int_{\mathcal{Y}}h_0(\boldsymbol{y})^2\int_{\mathcal{X}}\exp\left(b^*+\boldsymbol{\theta}^{*^T}\boldsymbol{y}\right)\,\mathrm{d}G_{\boldsymbol{X}}(\boldsymbol{x})\,\mathrm{d}F^*(\boldsymbol{y})\\&=\int_{\mathcal{X}}\int_{\mathcal{Y}}h_0(\boldsymbol{y})^2\exp\left(b^*+\boldsymbol{\theta}^{*^T}\boldsymbol{y}\right)\,\mathrm{d}F^*(\boldsymbol{y})\,\mathrm{d}G_{\boldsymbol{X}}(\boldsymbol{x})\\&=\int_{\mathcal{X}}(B_{\boldsymbol{\beta}^*,F^*}h_0^2)(\boldsymbol{x})\,\mathrm{d}G_{\boldsymbol{X}}(\boldsymbol{x})\,.\end{aligned} \tag{5.100}$$

The second term of (5.99) becomes

$$\begin{aligned}P^*(h_0B^*_{\boldsymbol{\beta}^*,F^*}B_{\boldsymbol{\beta}^*,F^*}h_0)&=\int_{\mathcal{Y}}\int_{\mathcal{X}}h_0(\boldsymbol{y})B_{\boldsymbol{\beta}^*,F^*}h_0(\boldsymbol{x})\exp\left(b^*+\boldsymbol{\theta}^{*^T}\boldsymbol{y}\right)\,\mathrm{d}G_{\boldsymbol{X}}(\boldsymbol{x})\,\mathrm{d}F^*(\boldsymbol{y})\\&=\int_{\mathcal{X}}B_{\boldsymbol{\beta}^*,F^*}h_0(\boldsymbol{x})\left(\int_{\mathcal{Y}}h_0(\boldsymbol{y})\exp\left(b^*+\boldsymbol{\theta}^{*^T}\boldsymbol{y}\right)\,\mathrm{d}F^*(\boldsymbol{y})\right)\,\mathrm{d}G_{\boldsymbol{X}}(\boldsymbol{x})\\&=\int_{\mathcal{X}}\left(B_{\boldsymbol{\beta}^*,F^*}h_0(\boldsymbol{x})\right)^2\,\mathrm{d}G_{\boldsymbol{X}}(\boldsymbol{x})\,.\end{aligned} \tag{5.101}$$

The third term of (5.99) becomes

$$\begin{aligned}P^*(h_0C^*_{\boldsymbol{\beta}^*,F^*}C_{\boldsymbol{\beta}^*,F^*}h_0)&=\int_{\mathcal{Y}}\int_{\mathcal{X}}h_0(\boldsymbol{y})C_{\boldsymbol{\beta}^*,F^*}h_0(\boldsymbol{x})\Sigma_{\boldsymbol{Y}}^{-\frac{1}{2}}(\boldsymbol{y}-\boldsymbol{\mu}^*)\exp\left(b^*+\boldsymbol{\theta}^{*^T}\boldsymbol{y}\right)\,\mathrm{d}G_{\boldsymbol{X}}(\boldsymbol{x})\,\mathrm{d}F^*(\boldsymbol{y})\\&=\int_{\mathcal{X}}C_{\boldsymbol{\beta}^*,F^*}h_0(\boldsymbol{x})\int_{\mathcal{Y}}h_0(\boldsymbol{y})\Sigma_{\boldsymbol{Y}}^{-\frac{1}{2}}(\boldsymbol{y}-\boldsymbol{\mu}^*)\exp\left(b^*+\boldsymbol{\theta}^{*^T}\boldsymbol{y}\right)\,\mathrm{d}F^*(\boldsymbol{y})\,\mathrm{d}G_{\boldsymbol{X}}(\boldsymbol{x})\\&=\int_{\mathcal{X}}C_{\boldsymbol{\beta}^*,F^*}h_0(\boldsymbol{x})\Sigma_{\boldsymbol{Y}}^{-\frac{1}{2}}\mathbb{E}_{\boldsymbol{\beta}^*,F^*}\left[h_0(\boldsymbol{Y})(\boldsymbol{Y}-\boldsymbol{\mu})|\boldsymbol{x}\right]\,\mathrm{d}G_{\boldsymbol{X}}(\boldsymbol{x})\\&=\int_{\mathcal{X}}\left(C_{\boldsymbol{\beta}^*,F^*}h_0(\boldsymbol{x})\right)\left(C_{\boldsymbol{\beta}^*,F^*}h_0(\boldsymbol{x})\right)^T\,\mathrm{d}G_{\boldsymbol{X}}(\boldsymbol{x})\,.\end{aligned} \tag{5.102}$$

Subbing (5.100), (5.101), (5.102) into (5.99), we have the following relationship

$$\int_{\mathcal{X}}(B_{\boldsymbol{\beta}^*,F^*}h_0^2)(\boldsymbol{x})-\left(B_{\boldsymbol{\beta}^*,F^*}h_0(\boldsymbol{x})\right)^2-\left(C_{\boldsymbol{\beta}^*,F^*}h_0(\boldsymbol{x})\right)\left(C_{\boldsymbol{\beta}^*,F^*}h_0(\boldsymbol{x})\right)^T\,\mathrm{d}G_{\boldsymbol{X}}(\boldsymbol{x})=0\,. \tag{5.103}$$

Now, let us consider the score function at $h_0, A_{\boldsymbol{\beta}^*,F^*}h_0$. We can obtain this by using the parametric submodel defined by $dF_t = (1+th_0)dF^*$ which is a path only in the parameter space for $F$ that goes to the true $(\boldsymbol{\beta}^*, F^*)$ when $t=0$. Following the same derivations as in Section 2.3.2, we have that

$$A_{\boldsymbol{\beta}^*,F^*}h_0 = h_0(\boldsymbol{Y}) - \boldsymbol{B}_{\boldsymbol{\beta}^*,F^*}h_0(\boldsymbol{X}) - \boldsymbol{C}_{\boldsymbol{\beta}^*,F^*}h_0(\boldsymbol{X})\Sigma_{\boldsymbol{Y}}^{-\frac{1}{2}^*}\left(\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{X}^T\boldsymbol{\beta}^*)\right) ,$$

where

$$\mathbb{E}^{\mathcal{X}}\left[A_{\boldsymbol{\beta}^*,F^*}h_0|\boldsymbol{X}\right] = 0,\ \mathbb{E}_{\boldsymbol{\beta}^*,F^*}\left[A_{\boldsymbol{\beta}^*,F^*}h_0\right] = 0\ .$$

Considering the variance of the expression, by the law of total variance we have that

$$\mathrm{Var}\left(A_{\boldsymbol{\beta}^*,F^*}h_0\right) = \mathbb{E}^{\mathcal{X}}\left[\mathrm{Var}(A_{\boldsymbol{\beta}^*,F^*}h_0|\boldsymbol{X})\right]\ . \tag{5.104}$$

The conditional variance expression can be evaluated as

$$\begin{aligned}
\mathrm{Var}(A_{\boldsymbol{\beta}^*,F^*}h_0|\boldsymbol{X}) &= \mathbb{E}\left[(A_{\boldsymbol{\beta}^*,F^*}h_0)(A_{\boldsymbol{\beta}^*,F^*}h_0)^T|\boldsymbol{X}\right] \\
&= \mathbb{E}\left[\left(h_0(\boldsymbol{Y}) - \boldsymbol{B}_{\boldsymbol{\beta}^*,F^*}h_0(\boldsymbol{X})\right)^2|\boldsymbol{X}\right] && (5.105)\\
&\quad - 2\mathbb{E}\left[\left(h_0(\boldsymbol{Y}) - \boldsymbol{B}_{\boldsymbol{\beta}^*,F^*}h_0(\boldsymbol{X})\right)C_{\boldsymbol{\beta}^*,F^*}h_0\Sigma_{\boldsymbol{Y}}^{-\frac{1}{2}^*}(\boldsymbol{Y}-\boldsymbol{\mu}^*)|\boldsymbol{X}\right] && (5.106)\\
&\quad + \mathbb{E}\left[C_{\boldsymbol{\beta}^*,F^*}h_0\Sigma_{\boldsymbol{Y}}^{-\frac{1}{2}^*}(\boldsymbol{Y}-\boldsymbol{\mu}^*)(\boldsymbol{Y}-\boldsymbol{\mu}^*)^T\Sigma_{\boldsymbol{Y}}^{-\frac{1}{2}^*}(C_{\boldsymbol{\beta}^*,F^*}h_0)^T|\boldsymbol{X}\right]\ . && (5.107)
\end{aligned}$$

The expression (5.105) becomes

$$\begin{aligned}
\mathbb{E}\left[\left(h_0(\boldsymbol{Y}) - \boldsymbol{B}_{\boldsymbol{\beta}^*,F^*}h_0(\boldsymbol{X})\right)^2|\boldsymbol{X}\right] &= \mathbb{E}\left[h_0(\boldsymbol{Y})^2|\boldsymbol{X}\right] - 2B_{\boldsymbol{\beta}^*,F^*}h_0(\boldsymbol{X})\mathbb{E}\left[h_0(\boldsymbol{Y})|\boldsymbol{X}\right] + B_{\boldsymbol{\beta}^*,F^*}h_0(\boldsymbol{X})^2 \\
&= B_{\boldsymbol{\beta}^*,F^*}h_0^2(\boldsymbol{X}) - B_{\boldsymbol{\beta}^*,F^*}h_0(\boldsymbol{X})^2 && (5.108)
\end{aligned}$$

The expression (5.106) becomes

$$\begin{aligned}
\mathbb{E}\left[\left(h_0(\boldsymbol{Y}) - \boldsymbol{B}_{\boldsymbol{\beta}^*,F^*}h_0(\boldsymbol{X})\right)C_{\boldsymbol{\beta}^*,F^*}h_0\Sigma_{\boldsymbol{Y}}^{-\frac{1}{2}^*}(\boldsymbol{Y}-\boldsymbol{\mu}^*)|\boldsymbol{X}\right] &= \mathbb{E}\left[h_0(\boldsymbol{Y})C_{\boldsymbol{\beta}^*,F^*}h_0\Sigma_{\boldsymbol{Y}}^{-\frac{1}{2}^*}(\boldsymbol{Y}-\boldsymbol{\mu}^*)|\boldsymbol{X}\right] \\
&\quad - \mathbb{E}\left[B_{\boldsymbol{\beta}^*,F^*}h_0(\boldsymbol{X})C_{\boldsymbol{\beta}^*,F^*}h_0\Sigma_{\boldsymbol{Y}}^{-\frac{1}{2}^*}(\boldsymbol{Y}-\boldsymbol{\mu}^*)|\boldsymbol{X}\right] \\
&= C_{\boldsymbol{\beta}^*,F^*}h_0\Sigma_{\boldsymbol{Y}}^{-\frac{1}{2}^*}\mathbb{E}\left[h_0(\boldsymbol{Y})(\boldsymbol{Y}-\boldsymbol{\mu}^*)|\boldsymbol{X}\right] \\
&\quad - B_{\boldsymbol{\beta}^*,F^*}h_0(\boldsymbol{X})C_{\boldsymbol{\beta}^*,F^*}h_0\Sigma_{\boldsymbol{Y}}^{-\frac{1}{2}^*}\mathbb{E}\left[(\boldsymbol{Y}-\boldsymbol{\mu}^*)|\boldsymbol{X}\right] \\
&= (C_{\boldsymbol{\beta}^*,F^*}h_0)(C_{\boldsymbol{\beta}^*,F^*}h_0)^T - 0\ . && (5.109)
\end{aligned}$$

The expression (5.107) becomes

$$C_{\boldsymbol{\beta}^*,F^*}h_0\Sigma_{\boldsymbol{Y}}^{-\frac{1}{2}^*}\mathbb{E}\left(\boldsymbol{Y}-\boldsymbol{\mu}^*)(\boldsymbol{Y}-\boldsymbol{\mu}^*)^T|\boldsymbol{X}\right]\Sigma_{\boldsymbol{Y}}^{-\frac{1}{2}^*}(C_{\boldsymbol{\beta}^*,F^*}h_0)^T = (C_{\boldsymbol{\beta}^*,F^*}h_0)(C_{\boldsymbol{\beta}^*,F^*}h_0)^T\ . \tag{5.110}$$

Thus, subbing in (5.108),(5.109), (5.110), we can express the conditional variance as

$$\mathrm{Var}(A_{\boldsymbol{\beta}^*,F^*}h_0|\boldsymbol{X}) = B_{\boldsymbol{\beta}^*,F^*}h_0^2 - (B_{\boldsymbol{\beta}^*,F^*}h_0)^2 - (C_{\boldsymbol{\beta}^*,F^*}h_0)(C_{\boldsymbol{\beta}^*,F^*}h_0)^T \tag{5.111}$$

Subbing the conditional variance (5.111) back into the law of total variance (5.104),

$$\begin{aligned}
\mathrm{Var}(A_{\boldsymbol{\beta}^*,F^*}h_0) &= \mathbb{E}^{\mathcal{X}}\left[B_{\boldsymbol{\beta}^*,F^*}h_0^2 - (B_{\boldsymbol{\beta}^*,F^*}h_0)^2 - (C_{\boldsymbol{\beta}^*,F^*}h_0)(C_{\boldsymbol{\beta}^*,F^*}h_0)^T\right] \\
&= \int_{\mathcal{X}} B_{\boldsymbol{\beta}^*,F^*}h_0^2(\boldsymbol{x}) - (B_{\boldsymbol{\beta}^*,F^*}h_0(\boldsymbol{x})^2 - (C_{\boldsymbol{\beta}^*,F^*}h_0(\boldsymbol{x}))(C_{\boldsymbol{\beta}^*,F^*}h_0(\boldsymbol{x}))^T\,\mathrm{d}G_{\boldsymbol{X}}(\boldsymbol{x}) \\
&= 0 && (5.112)
\end{aligned}$$

by (5.103). This implies that the score function along the submodel $(\boldsymbol{\beta}^*, F_t)$ in the direction of $h_0$ is 0, and as its expectation is 0, this implies that $h_0$ is such that $A_{\boldsymbol{\beta}^*,F} h_0 = 0$ almost surely. The only solutions that satisfy these conditions for all $(\boldsymbol{\beta}, F)$ are of the form $a\boldsymbol{y} + c$ for some vector of scalars $\boldsymbol{a}$ and for scalar $c$.

$$\begin{aligned}
A_{\boldsymbol{\beta},F}(\boldsymbol{a}\boldsymbol{y} + c) &= \boldsymbol{a}\boldsymbol{y} + c - B_{\boldsymbol{\beta},F}(\boldsymbol{a}\boldsymbol{y} + c) - C_{\boldsymbol{\beta},F}(\boldsymbol{a}\boldsymbol{y} + c)\Sigma_{\boldsymbol{Y}}^{-\frac{1}{2}}(\boldsymbol{y} - \boldsymbol{\mu}) \\
&= \boldsymbol{a}\boldsymbol{y} + c - \boldsymbol{a}\left(\int_{\mathcal{Y}} \boldsymbol{y} \exp\left(b + \boldsymbol{\theta}^T \boldsymbol{y}\right) \mathrm{d}F(\boldsymbol{y})\right) - c - \mathbb{E}_{\boldsymbol{\beta},F}\left[(\boldsymbol{a}\boldsymbol{Y} + c)(\boldsymbol{Y} - \boldsymbol{\mu})\right]^T \Sigma_{\boldsymbol{Y}}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}) \\
&= \boldsymbol{a}(\boldsymbol{y} - \boldsymbol{\mu}) - \boldsymbol{a}\mathbb{E}_{\boldsymbol{\beta},F}\left[\boldsymbol{Y}(\boldsymbol{Y} - \boldsymbol{\mu})\right]^T \Sigma_{\boldsymbol{Y}}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}) - b\mathbb{E}_{\boldsymbol{\beta},F}\left[(\boldsymbol{Y} - \boldsymbol{\mu})\right]^T \Sigma_{\boldsymbol{Y}}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}) \\
&= \boldsymbol{a}(\boldsymbol{y} - \boldsymbol{\mu}) - \boldsymbol{a}\Sigma_Y \Sigma_{\boldsymbol{Y}}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}) - 0 \\
&= 0
\end{aligned}$$

Therefore, $h_0 \equiv 0$ implying that $\left(\omega I - B^*_{\boldsymbol{\beta}^*,F^*} B_{\boldsymbol{\beta}^*,F^*} - C^*_{\boldsymbol{\beta}^*,F^*} C_{\boldsymbol{\beta}^*,F^*}\right)$ is one-to-one and is continuously invertible by Fredholm Theory. Bringing this all together, we have shown that $\dot{\Psi}_{22}$ is continuously invertible, and that $\sqrt{n}\,(\hat{F} - F^*)$ converges in distribution in $\ell^\infty(\mathcal{H}_L)$ to $G_2 = -\Psi_{22}^{-1} Z_2$ where $G_2$ is a mean zero Gaussian random process with co-variance function

$$W_2(h_1, h_2) = W_F(\mathcal{D}h_1, \mathcal{D}h_2)\ ,$$

where $\mathcal{D}$ is implicitly defined by (5.96). We have established the joint asymptotic normality of the MELE as claimed in Proposition 2.1.

## 5.4 Technical Details of Proposition 2.2

To show that the profile log-likelihood behaves asymptotically like a true log-likelihood for $\boldsymbol{\beta}$, we will follow the steps from Huang (2014) and apply a modified version of the main theorem from Murphy and Van Der Vaart (2000).

**Theorem 5.3** *Define $\ell_{\boldsymbol{\beta},F}(t) = \log l(t, F_t(\boldsymbol{\beta}, F))$ as before. Let $\dot{\ell}_{\boldsymbol{\beta},F}(t)$ and $\ddot{\ell}_{\boldsymbol{\beta},F}$ denote the first two derivatives of $\ell_{\boldsymbol{\beta},F}(t)$ with respect to $t$. If the following conditions are satisfied,*

*(i) An approximately least-favourable submodel $(t, F_t(\boldsymbol{\beta}, F))$ exists. This means that the submodel passes through $(\boldsymbol{\beta}, F)$ at $t = \boldsymbol{\beta}$,*

$$F_{\boldsymbol{\beta}}(\boldsymbol{\beta}, F) = F, \quad \forall(\boldsymbol{\beta}, F)$$

*and that it is least favourable $t$ $(\boldsymbol{\beta}^*, F^*)$ for estimating $\boldsymbol{\beta}$, where*

$$\dot{\ell}_{\boldsymbol{\beta}^*,F^*}(\boldsymbol{\beta}^*)(\boldsymbol{X}, \boldsymbol{Y}) = \tilde{S}_{\boldsymbol{\beta}^*,F^*}(\boldsymbol{X}, \boldsymbol{Y})\ .$$

*(ii) For any sequence $\boldsymbol{\beta}_n \xrightarrow{P} \boldsymbol{\beta}^*$, the sequence $\hat{F}(\boldsymbol{\beta}_n)$ where $\hat{F}(\boldsymbol{\beta})$ is the value of $F$ that maximizes the log-likelihood for a fixed $\boldsymbol{\beta}$, should be consistent with respect to the weak topology .*

*(iii) $P^* \dot{\ell}_{\boldsymbol{\beta}^*,\hat{F}(\boldsymbol{\beta}_n)}(\boldsymbol{\beta}^*) = o_p(||\boldsymbol{\beta}_n - \boldsymbol{\beta}^*|| + n^{-1/2})$. In other words, the score function along the approximately least favourable submodel is asymptotically unbiased.*

*(iv) $\dot{\ell}_{\boldsymbol{\beta},F}(t)$ and $\ddot{\ell}_{\boldsymbol{\beta},F}(t)$ are continuous at $(\boldsymbol{\beta}^*, \boldsymbol{\beta}^*, F^*)$ and there exists some neighbourhood $V$ of $(\boldsymbol{\beta}^*, \boldsymbol{\beta}^*, F^*)$ such that the class of functions $\{\dot{\ell}_{\boldsymbol{\beta},F}(t) | (t, \boldsymbol{\beta}, F) \in V\}$ is $P^*$ Donsker and $\{\ddot{\ell}_{\boldsymbol{\beta},F}(t) | (t, \boldsymbol{\beta}, F) \in V\}$ is $P^*$-Glivenko-Cantelli.*

*Then if the maximiser $\hat{\boldsymbol{\beta}}_n$ of $pl_n(\boldsymbol{\beta})$ is consistent for $\boldsymbol{\beta}^*$ then under the null hypothesis, $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}^*$ the profile empirical log-likelihood ratio statistic*

$$-2(pl_n(\boldsymbol{\beta}^*) - pl_n(\hat{\boldsymbol{\beta}}))$$

*is asymptotically $\chi^2_q$ in distribution as $n \to \infty$.*

Note that (ii) follows from Lemma 2.2 since for any consistent sequence $\boldsymbol{\beta}_n$, we can express it as

$$\boldsymbol{\beta}_n = \hat{\boldsymbol{\beta}} + o_p(1)$$

Then, by a first-order Taylor expansion, we have that

$$\hat{F}(\boldsymbol{\beta}_n) = \hat{F}(\hat{\boldsymbol{\beta}}) + o_p(1) \xrightarrow{P} F^*$$

relative to the weak topology, as the partial derivative of $\hat{F}(\boldsymbol{\beta})$ is uniformly bounded for $\boldsymbol{\beta}$ in a sufficiently small neighbourhood of $\boldsymbol{\beta}^*$.

For condition (iv), rigorously checking Donsker and Glivenko-Cantelli properties of the first two derivatives respectively requires calculations similar to that in the proof of Proposition 2.1. For brevity, the details are not included, but as they are made up of uniformly bounded and well-behaved functions, Huang (2011) provides an intuitive argument for the first and second derivative of the log-likelihood along the approximately least favourable submodel is well-behaved in a Donsker and Glivenko-Cantelli sense.

### 5.4.1 Constructing an Approximately Least Favourable Submodel

As we just have to show an approximately least favourable submodel exists, we can construct one. Following the same construction as Huang (2014), for each $(\boldsymbol{\beta}, F) \in \mathbb{R}^q \times \mathcal{F}_\mu$, consider the mapping $(\boldsymbol{\beta}, F) \mapsto (t, F_t(\boldsymbol{\beta}, F))$ indexed by $t$, where $F_t$ has densities is defined by

$$dF_t(\boldsymbol{\beta}, F) = (1 + (t - \boldsymbol{\beta})^T h) dF. \tag{5.113}$$

Note that we do not need to tilt $dF_t(\boldsymbol{\beta}, F)$ to have mean $\boldsymbol{\mu}$ as the log-likelihood function is invariant to an exponential tilting of $F$. The only requirement for $dF_t$ is that for $t$ in a sufficiently small neighbourhood of $\boldsymbol{\beta}$, it is non-negative, which is satisfied if $h$ is a bounded function, for a $q \times 1$-dimensional function $h = h(\boldsymbol{y})$.

For a fixed $(\boldsymbol{\beta}, F)$, the log-likelihood for the parameter $t$ along this mapping is given by

$$l_{\boldsymbol{\beta},F}(t)(\boldsymbol{X}, \boldsymbol{Y}) = \log dF_t(\boldsymbol{\beta}, F)(\boldsymbol{Y}) + b(\boldsymbol{X}; t, F_t(\boldsymbol{\beta}, F)) + \boldsymbol{\theta}(\boldsymbol{X}; t, F_t(\boldsymbol{\beta}, F))^T \boldsymbol{Y}. \tag{5.114}$$

Differentiating (5.114) with respect to $t$ and evaluating at $t = \boldsymbol{\beta}$, the score function along this submodel is given by

$$\dot{l}_{\boldsymbol{\beta},F}(\boldsymbol{\beta})(\boldsymbol{X}, \boldsymbol{Y}) = h(\boldsymbol{Y}) + \left.\frac{\partial b_t}{\partial t}\right|_{t=\boldsymbol{\beta}} + \left.\frac{\partial \boldsymbol{\theta}_t}{\partial t}\right|_{t=\boldsymbol{\beta}} \boldsymbol{Y} \tag{5.115}$$

We can find these partial derivatives by similar working to Section 2.3.2, but we will detail some of the steps where it differs as in this context, $t$ is a $q \times 1$ dimensional vector.

Consider finding the partial derivative of $\boldsymbol{\theta}$ with respect to $t$ along our submodel, the mean constraint on the submodel becomes

$$\mathbf{0} = \int_{\mathcal{Y}} \big(\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{X}^T t)\big) \exp\big(b_t + \boldsymbol{\theta}_t^T \boldsymbol{y}\big)(1 + (t - \boldsymbol{\beta})^T h(\boldsymbol{y}))\, \mathrm{d}F(\boldsymbol{y})$$

where $b_t = b(\boldsymbol{X}; t, F_t(\boldsymbol{\beta}, F)), \boldsymbol{\theta}_t = \boldsymbol{\theta}(\boldsymbol{X}; t, F_t(\boldsymbol{\beta}, F))$. Differentiating with respect to $t$ and using matrix derivative identity, $\dfrac{\partial(v(\boldsymbol{x})\boldsymbol{u}(\boldsymbol{x}))}{\partial \boldsymbol{x}} = v(\boldsymbol{x})\dfrac{\partial \boldsymbol{u}(\boldsymbol{x})}{\partial \boldsymbol{x}} + \dfrac{\partial v(\boldsymbol{x})}{\partial \boldsymbol{x}}(\boldsymbol{u}(\boldsymbol{x}))^T$ where $v(\boldsymbol{x}) = \exp\big(b_t + \boldsymbol{\theta}_t^T \boldsymbol{y}\big)(1 + (t - \boldsymbol{\beta})^T h(\boldsymbol{y})$ and $u(\boldsymbol{x}) = (\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{X}^T t))$,

$$\mathbf{0} = \int_{\mathcal{Y}} -\frac{\partial \boldsymbol{\mu}(\boldsymbol{X}^T t)}{\partial t} \exp\big(b_t + \boldsymbol{\theta}_t^T \boldsymbol{y}\big)(1 + (t - \boldsymbol{\beta})^T h(\boldsymbol{y}))\, \mathrm{d}F(\boldsymbol{y}) \tag{5.116}$$

$$+ \int_{\mathcal{Y}} \frac{\partial}{\partial t} \left(\exp\big(b_t + \boldsymbol{\theta}_t^T \boldsymbol{y}\big)(1 + (t - \boldsymbol{\beta})^T h(\boldsymbol{y}))\right) \big(\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{X}^T t)\big)^T \, \mathrm{d}F(\boldsymbol{y}) \tag{5.117}$$

Simplifying (5.116), similar to working in Section 2.3.1 where $t$ is split up to match the dimensions of the $K'$ mean models,

$$\left.\frac{\partial \boldsymbol{\mu}(\boldsymbol{X}^T t)}{\partial t}\right|_{t=\boldsymbol{\beta}} = D(\boldsymbol{X}; \boldsymbol{\beta}),$$

which simplifies (5.116) at this evaluation to

$$-D(\boldsymbol{X}; \boldsymbol{\beta}) \int_{\mathcal{Y}} \left(\exp\big(b + \boldsymbol{\theta}^T \boldsymbol{y}\big)(1 + (\boldsymbol{\beta} - \boldsymbol{\beta})^T h(\boldsymbol{y}))\right) \mathrm{d}F(\boldsymbol{y}) = -D(\boldsymbol{X}; \boldsymbol{\beta})$$

Simplifying (5.117),

$$\begin{aligned}
&\int_{\mathcal{Y}} \frac{\partial}{\partial t} \left(\exp\big(b_t + \boldsymbol{\theta}_t^T \boldsymbol{y}\big)(1 + (t - \boldsymbol{\beta})^T h(\boldsymbol{y}))\right) \big(\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{X}^T t)\big)^T \, \mathrm{d}F(\boldsymbol{y}) \\
&= \int_{\mathcal{Y}} \frac{\partial \boldsymbol{\theta}}{\partial t} \boldsymbol{y} \exp\big(b_t + \boldsymbol{\theta}_t^T \boldsymbol{y}\big)(1 + (t - \boldsymbol{\beta})^T h(\boldsymbol{y}))\big(\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{X}^T t)\big)^T \, \mathrm{d}F(\boldsymbol{y}) \\
&+ \int_{\mathcal{Y}} \exp\big(b_t + \boldsymbol{\theta}_t^T \boldsymbol{y}\big) h(\boldsymbol{y}) \big(\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{X}^T t)\big)^T \, \mathrm{d}F(\boldsymbol{y})
\end{aligned}$$

Evaluating this at $t = \boldsymbol{\beta}$, the expression becomes

$$\begin{aligned}
&= \left.\frac{\partial \boldsymbol{\theta}}{\partial t}\right|_{t=\boldsymbol{\beta}} \int_{\mathcal{Y}} \boldsymbol{y} \big(\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{X}^T \boldsymbol{\beta})\big)^T \exp\big(b + \boldsymbol{\theta}^T \boldsymbol{y}\big)\, \mathrm{d}F(\boldsymbol{y}) + \int_{\mathcal{Y}} \exp\big(b + \boldsymbol{\theta}^T \boldsymbol{y}\big) h(\boldsymbol{y}) \big(\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{X}^T \boldsymbol{\beta})\big)^T \, \mathrm{d}F(\boldsymbol{y}) \\
&= \left.\frac{\partial \boldsymbol{\theta}}{\partial t}\right|_{t=\boldsymbol{\beta}} \Sigma_{\boldsymbol{Y}} + \mathbb{E}\left[h(\boldsymbol{Y})(\boldsymbol{Y} - \boldsymbol{\mu})^T | \boldsymbol{X}\right]
\end{aligned}$$

Therefore,

$$\mathbf{0} = -D(\boldsymbol{X}; \boldsymbol{\beta}) + \left.\frac{\partial \boldsymbol{\theta}}{\partial t}\right|_{t=\boldsymbol{\beta}} \Sigma_{\boldsymbol{Y}} + \mathbb{E}\left[h(\boldsymbol{Y})(\boldsymbol{Y} - \boldsymbol{\mu})^T | \boldsymbol{X}\right]$$

$$\left.\frac{\partial \boldsymbol{\theta}}{\partial t}\right|_{t=\boldsymbol{\beta}} = \left(D(\boldsymbol{X}; \boldsymbol{\beta}) - \mathbb{E}\left[h(\boldsymbol{Y})(\boldsymbol{Y} - \boldsymbol{\mu})^T | \boldsymbol{X}\right]\right) \Sigma_{\boldsymbol{Y}}^{-1} \tag{5.118}$$

Now, by the same working as in Section 2.3.2 with the new submodel,

$$\left.\frac{\partial b}{\partial t}\right|_{t=\boldsymbol{\beta}} = -\left(\left.\frac{\partial \boldsymbol{\theta}}{\partial t}\right|_{t=\boldsymbol{\beta}}\right) \boldsymbol{\mu} - \mathbb{E}\left[h(\boldsymbol{Y}) | \boldsymbol{X}\right] \tag{5.119}$$

Subbing back into (5.115), we find that

$$i_{\boldsymbol{\beta}, F}(\boldsymbol{\beta})(\boldsymbol{X}, \boldsymbol{Y}) = h(\boldsymbol{Y}) - \mathbb{E}\left[h(\boldsymbol{Y}) | \boldsymbol{X}\right] + D(\boldsymbol{X}; \boldsymbol{\beta}) \Sigma_{\boldsymbol{Y}}^{-1} (\boldsymbol{Y} - \boldsymbol{\mu}) - \mathbb{E}\left[h(\boldsymbol{Y})(\boldsymbol{Y} - \boldsymbol{\mu})^T | \boldsymbol{X}\right] \Sigma_Y^{-1} (\boldsymbol{Y} - \boldsymbol{\mu})$$

Now, if we consider letting $h(\boldsymbol{y}) = \mathbf{1}\boldsymbol{y}$, where $\mathbf{1}_{q\times K}$ denotes a $q \times K$ matrix of 1s, then the score function along this mapping for any $(\boldsymbol{\beta}, F)$ becomes

$$\begin{aligned}
\dot{l}_{\boldsymbol{\beta},F}(\boldsymbol{\beta})(\boldsymbol{X},\boldsymbol{Y}) &= \mathbf{1}(\boldsymbol{Y}-\boldsymbol{\mu}) + D(\boldsymbol{X};\boldsymbol{\beta})\Sigma_{\boldsymbol{Y}}^{-1}(\boldsymbol{Y}-\boldsymbol{\mu}) - \mathbf{1}\mathbb{E}[\boldsymbol{Y}(\boldsymbol{Y}-\boldsymbol{\mu})^T|\boldsymbol{X}]\Sigma_{\boldsymbol{Y}}^{-1}(\boldsymbol{Y}-\boldsymbol{\mu}) \\
&= \mathbf{1}(\boldsymbol{Y}-\boldsymbol{\mu}) + D(\boldsymbol{X};\boldsymbol{\beta})\Sigma_{\boldsymbol{Y}}^{-1}(\boldsymbol{Y}-\boldsymbol{\mu}) - \mathbf{1}(\boldsymbol{Y}-\boldsymbol{\mu}) \\
&= D(\boldsymbol{X};\boldsymbol{\beta})\Sigma_{\boldsymbol{Y}}^{-1}(\boldsymbol{Y}-\boldsymbol{\mu}) \\
&= \tilde{S}_{\boldsymbol{\beta},F}(\boldsymbol{X},\boldsymbol{Y}) \,,
\end{aligned}$$

which is the efficient score function (5.3). Therefore, we satisfy the least favourable condition at the true parameters $(\boldsymbol{\beta}^*, F^*)$. Finally, we need that $h(\boldsymbol{y}) = \mathbf{1}\boldsymbol{y}$ is bounded, but this holds under Assumption A.2.1. Therefore, the mapping $(\boldsymbol{\beta}, F) \mapsto (t, F_t(\boldsymbol{\beta}, F))$ is an approximately least favourable submodel.

### 5.4.2 Asymptotic Unbiasedness of the Approximately Least Favourable Submodel Score Function

We now wish to show that

$$\mathbb{E}_{\boldsymbol{\beta}^*,F^*}\left[\dot{l}_{\boldsymbol{\beta}^*,\hat{F}(\boldsymbol{\beta}_n)}(\boldsymbol{\beta}^*)\right] = o_P(||\boldsymbol{\beta}_n - \boldsymbol{\beta}^*|| + n^{-1/2}) \tag{5.120}$$

for all random sequences $\boldsymbol{\beta}_n$ such that $\boldsymbol{\beta}_n \xrightarrow{P} \boldsymbol{\beta}^*$, which is satisfied for our submodel. Using similar derivations to Section 2.3.2 and 5.4.1, we can show that the score function evaluated at $t = \boldsymbol{\beta}^*$ along the least favourable submodel is given by the expression

$$\begin{aligned}
\dot{l}_{\boldsymbol{\beta},F}(\boldsymbol{\beta}^*)(\boldsymbol{X},\boldsymbol{Y}) = \frac{\mathbf{1}\boldsymbol{Y}}{1+(\boldsymbol{\beta}^*-\boldsymbol{\beta})^T\mathbf{1}\boldsymbol{Y}} + \frac{\partial}{\partial t}\boldsymbol{\theta}\left(\boldsymbol{X};\boldsymbol{\beta}^*,F_{\boldsymbol{\beta}^*}(\boldsymbol{\beta},F)\right)(\boldsymbol{Y}-\boldsymbol{\mu}) \\
- \frac{\mathbf{1}\int_{\mathcal{Y}}\boldsymbol{y}\exp\left\{\boldsymbol{\theta}\left(\boldsymbol{X};\boldsymbol{\beta}^*,F_{\boldsymbol{\beta}^*}(\boldsymbol{\beta},F)\right)^T\boldsymbol{y}\right\}\mathrm{d}F(\boldsymbol{y})}{\int_{\mathcal{Y}}\exp\left\{\boldsymbol{\theta}\left(\boldsymbol{X};\boldsymbol{\beta}^*,F_{\boldsymbol{\beta}^*}(\boldsymbol{\beta},F)\right)^T\boldsymbol{y}\right\}(1+(\boldsymbol{\beta}^*-\boldsymbol{\beta})^T\mathbf{1}\boldsymbol{y})\,\mathrm{d}F(\boldsymbol{y})} \,.
\end{aligned}$$

Now, taking the expectation under the true values $(\boldsymbol{\beta}^*, F^*)$, we get

$$\mathbb{E}_{\boldsymbol{\beta}^*,F^*}\left[\dot{l}_{\boldsymbol{\beta},F}(\boldsymbol{\beta}^*)(\boldsymbol{X},\boldsymbol{Y})\right] = \mathbb{E}^{\mathcal{X}}\left[\mathbb{E}_{\boldsymbol{\beta}^*,F^*}\left[\dot{l}_{\boldsymbol{\beta},F}(\boldsymbol{\beta}^*)(\boldsymbol{X},\boldsymbol{Y})|\boldsymbol{X}\right]\right]$$

The inner expectation then becomes

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\beta}^*,F^*}\left[\dot{l}_{\boldsymbol{\beta},F}(\boldsymbol{\beta}^*)(\boldsymbol{X},\boldsymbol{Y})|\boldsymbol{X}\right] &= \mathbf{1}\left(\int_{\mathcal{Y}}\frac{\boldsymbol{y}\exp\left\{b(\boldsymbol{X};\boldsymbol{\beta}^*,F^*)+\boldsymbol{\theta}(\boldsymbol{X};\boldsymbol{\beta}^*,F^*)^T\boldsymbol{y}\right\}}{1+(\boldsymbol{\beta}^*-\boldsymbol{\beta})^T\mathbf{1}\boldsymbol{y}}\,\mathrm{d}F^*(\boldsymbol{y})\right) \\
&\quad - \mathbf{1}\left(\frac{\int_{\mathcal{Y}}\boldsymbol{y}\exp\left\{\boldsymbol{\theta}\left(\boldsymbol{X};\boldsymbol{\beta}^*,F_{\boldsymbol{\beta}^*}(\boldsymbol{\beta},F)\right)^T\boldsymbol{y}\right\}\mathrm{d}F(\boldsymbol{y})}{\int_{\mathcal{Y}}\exp\left\{\boldsymbol{\theta}\left(\boldsymbol{X};\boldsymbol{\beta}^*,F_{\boldsymbol{\beta}^*}(\boldsymbol{\beta},F)\right)^T\boldsymbol{y}\right\}(1+(\boldsymbol{\beta}^*-\boldsymbol{\beta})^T\mathbf{1}\boldsymbol{y})\,\mathrm{d}F(\boldsymbol{y})}\right) + 0
\end{aligned}$$

Since $\boldsymbol{\beta}^*$ is non-random, and in particular not a function of $\boldsymbol{X}$, we can evaluate $\boldsymbol{\beta}$ at $\boldsymbol{\beta}^*$ in the inner expectation which gives us

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\beta}^*,F^*}\left[\dot{l}_{\boldsymbol{\beta},F}(\boldsymbol{\beta}^*)(\boldsymbol{X},\boldsymbol{Y})|\boldsymbol{X}\right]\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*} &= \mathbf{1}\left(\int_{\mathcal{Y}} \frac{\boldsymbol{y}\exp\left\{b(\boldsymbol{X};\boldsymbol{\beta}^*,F^*)+\boldsymbol{\theta}(\boldsymbol{X};\boldsymbol{\beta}^*,F^*)^T\boldsymbol{y}\right\}}{1+(\boldsymbol{\beta}^*-\boldsymbol{\beta}^*)^T\mathbf{1}\boldsymbol{y}}\,\mathrm{d}F^*(\boldsymbol{y})\right) \\
&\quad - \mathbf{1}\left(\frac{\int_{\mathcal{Y}}\boldsymbol{y}\exp\left\{\boldsymbol{\theta}\left(\boldsymbol{X};\boldsymbol{\beta}^*,F_{\boldsymbol{\beta}^*}(\boldsymbol{\beta}^*,F)\right)^T\boldsymbol{y}\right\}\mathrm{d}F(\boldsymbol{y})}{\int_{\mathcal{Y}}\exp\left\{\boldsymbol{\theta}\left(\boldsymbol{X};\boldsymbol{\beta}^*,F_{\boldsymbol{\beta}^*}(\boldsymbol{\beta}^*,F)\right)^T\boldsymbol{y}\right\}(1+(\boldsymbol{\beta}^*-\boldsymbol{\beta}^*)^T\mathbf{1}\boldsymbol{y})\,\mathrm{d}F(\boldsymbol{y})}\right) \\
&= \mathbf{1}\left(\int_{\mathcal{Y}}\boldsymbol{y}\exp\left\{b(\boldsymbol{X};\boldsymbol{\beta}^*,F^*)+\boldsymbol{\theta}(\boldsymbol{X};\boldsymbol{\beta}^*,F^*)^T\boldsymbol{y}\right\}\mathrm{d}F^*(\boldsymbol{y})\right) \\
&\quad - \mathbf{1}\left(\frac{\int_{\mathcal{Y}}\boldsymbol{y}\exp\left\{b(\boldsymbol{X};\boldsymbol{\beta}^*,F)+\boldsymbol{\theta}\left(\boldsymbol{X};\boldsymbol{\beta}^*,F\right)^T\boldsymbol{y}\right\}\mathrm{d}F(\boldsymbol{y})}{\int_{\mathcal{Y}}\exp\left\{b(\boldsymbol{X};\boldsymbol{\beta}^*,F)+\boldsymbol{\theta}\left(\boldsymbol{X};\boldsymbol{\beta}^*,F\right)^T\boldsymbol{y}\right\}\mathrm{d}F(\boldsymbol{y})}\right) \\
&= \mathbf{1}\left(\int_{\mathcal{Y}}\boldsymbol{y}\exp\left\{b(\boldsymbol{X};\boldsymbol{\beta}^*,F^*)+\boldsymbol{\theta}(\boldsymbol{X};\boldsymbol{\beta}^*,F^*)^T\boldsymbol{y}\right\}\mathrm{d}F^*(\boldsymbol{y})\right) \\
&\quad - \mathbf{1}\left(\int_{\mathcal{Y}}\boldsymbol{y}\exp\left\{b(\boldsymbol{X};\boldsymbol{\beta}^*,F)+\boldsymbol{\theta}\left(\boldsymbol{X};\boldsymbol{\beta}^*,F\right)^T\boldsymbol{y}\right\}\mathrm{d}F(\boldsymbol{y})\right)\,.
\end{aligned}$$

as $F_{\boldsymbol{\beta}^*}(\boldsymbol{\beta}^*,F)=F$. Note that by the definition of $\boldsymbol{\theta}(\boldsymbol{X};\boldsymbol{\beta},F)$, we have that

$$\boldsymbol{\mu}=\int_{\mathcal{Y}}\boldsymbol{y}\exp\left\{b(\boldsymbol{X};\boldsymbol{\beta},F)+\boldsymbol{\theta}(\boldsymbol{X};\boldsymbol{\beta},F)^T\boldsymbol{y}\right\}\mathrm{d}F(\boldsymbol{y}) \tag{5.121}$$

for any $F$. As a result, we have that

$$\mathbb{E}_{\boldsymbol{\beta}^*,F^*}\left[\dot{l}_{\boldsymbol{\beta},F}(\boldsymbol{\beta}^*)(\boldsymbol{X},\boldsymbol{Y})|\boldsymbol{X}\right]\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*}\equiv\mathbf{1}(\mathbf{0})=\mathbf{0}\in\mathbb{R}^q. \tag{5.122}$$

Therefore, the asymptotic

$$\mathbb{E}_{\boldsymbol{\beta}^*,F^*}\left[\dot{l}_{\boldsymbol{\beta},F}(\boldsymbol{\beta}^*)(\boldsymbol{X},\boldsymbol{Y})\right]=\mathbf{0}$$

for any $F$ and so in particular for $F=\hat{F}(\boldsymbol{\beta}_n)$ which shows the asymptotic unbiasedness condition (5.120) is satisfied.

Therefore, using Lemma 2.2 to establish consistency, we can apply Theorem 5.3 to establish the asymptotic $\chi^2$ distribution of the profile empirical log-likelihood ratio statistic claimed in Proposition 2.2.

# Chapter 6

# Discussion and Conclusion

In this thesis, a Vector Semiparametric Generalized Linear Model (VSPGLM) is proposed, which is a vector response generalisation of the models proposed by Huang (2014) and Rathouz and Gao (2009). The key innovation of VSPGLM is that it removes the requirement for the parametric distribution of the underlying response distribution conditional on the covariates. This is done by leaving the distribution $F$ unspecified and jointly estimating it with the usual mean model parameters $\boldsymbol{\beta}$ to overcome the challenge of model misspecification in a vector response context. Thus, the proposed framework is general and flexible to a variety of problems with different data types and structures, which usually require a specific parametric model for each case. Example usages to showcase the flexibility of VSPGLM were given in Chapter 4 where the model produced comparable and interpretable results to existing parametric and semiparametric VGLMs, highlighting it as an appropriate modelling approach.

The thesis also showed key asymptotic properties of VSPGLM which extend on the properties for the univariate model proposed by Huang (2014) as well as retain the properties from the standard VGLM framework. It was also shown that inference can be formed on the parameters $\boldsymbol{\beta}$ using both Wald-tests and Likelihood Ratio Tests, which provide asymptotically valid inference provided the underlying distribution is from a multivariate exponential family. VSPGLM still performs well for finite samples given it attains the semiparametric efficiency bound. It provides consistent estimates and standard errors which converge from below to the true standard errors due to the convex hull constraint as a result of utilising empirical likelihood to estimate $(\boldsymbol{\beta}, F)$. The key advantage of VSPGLM over competing frameworks is the minimal assumptions needed to fit the model while maintaining asymptotically valid estimates and inference of $\boldsymbol{\beta}$, which becomes more useful for problems with a complex structure that is difficult to express parametrically.

Currently, VSPGLM has a few drawbacks that require further investigation, as well as potential extensions to improve the model. The first is that the computational implementation of VSPGLM is currently performing constrained optimisation over $q + n(K + 2)$ parameters subject to $n(K + 1)$ constraints resulting in slow computation for problems with a larger number of response components or a large number of samples $n$ as highlighted in Chapter 3. This drawback existed in the originally proposed SPGLM by Huang (2014), so the different numerical procedures and improvements made to the computation implementation of SPGLM could also be investigated for VSPGLM. This includes the methods explored by Wurm and Rathouz (2018) which implement the SPGLM in the **R** package **gldrm** using an iterative approach that alternates between optimising $\boldsymbol{\beta}$ and $\boldsymbol{p}$ by marginally optimizing over one set of parameters while holding the other fixed in each step. The optimization of $\boldsymbol{p}$ uses the

Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm which would be the same for a multivariate setting given $\boldsymbol{p}$ is a set of probability masses, and the optimization of $\boldsymbol{\beta}$ is done using Fisher scoring which is equivalent to iteratively re-weighted least squares (IRLS). The key suggestion here is that at each update step for $F$, BFGS is iterated till convergence while only a single IRLS iteration is used in between the $F$ updates. Another alternative proposed by Puang-Ngern (2018) is to express the optimisation problem using Lagrange multipliers and then apply a proposed MI-scoring algorithm to estimate $\boldsymbol{p}$ and $\boldsymbol{\beta}$ which is an extension of the Newton-Mi algorithm proposed by Ma (2010). The algorithm updates $\boldsymbol{p}$ using the MI algorithm (Ma 2010) and then estimates half step iterates for $\boldsymbol{\theta}$ to update $\boldsymbol{\beta}$ using the Fisher scoring algorithm, where the estimates for $F$ is done by using a piecewise constant approximation. Although this may pose a potential coding challenge, the algorithms above could improve the computational performance of VSPGLM and are a key direction for future investigation. Another suggestion to reduce the number of parameters is to only consider putting probability masses on unique observations similar to what is done in the multivariate density ratio model by Marchese and Diao (2017), rather than accumulating the probabilities onto the unique values after optimising over the $n$ probability masses. In the event that the suggestions above prove to be too challenging to extend to VSPGLM, future research into a tailor-made optimization algorithm for VSPGLM could be explored to efficiently perform the joint estimation while reducing the number of parameters and constraints. There may also be performance advantages in implementing VSPGLM in other programming languages such as Python, C++ or **R**, while also increasing the accessibility of the model to a wider audience.

Furthermore, if extra information about the response distribution is known a priori, alterations can be made to the VSPGLM framework to improve the efficiency of the model computationally for finite samples. For discrete responses where the response's support is known such as the Sorbinil example in Section 4.6, probabilities masses could be placed on the entire support rather than only on the observed support, allowing for probability density to be allocated to points which are feasible but not observed. This also aids the estimation procedure as VSPGLM may have trouble converging for datasets with a very small number of samples due to the convex hull constraint. In the event that the response's support is unknown, probability masses could be placed on the Cartesian product of the set of observed responses, which can be seen as a naive form of generating new observations, although the parameters $\boldsymbol{\beta}$ would still be optimised only over the observed values. Other generative methods for expanding the observed support could be employed for continuous cases, but this may require a continuous distribution estimate.

The suggestions above all stay within the empirical likelihood approach, which although computationally expensive has desirable asymptotic properties and is a universal method for all types of vector response data as it only places probability masses on the observed support. The semiparametric framework currently relies on exponential tilting and empirical likelihood, but the framework does allow for any estimator of the underlying distribution $F$ to be used in the likelihood. As a result, other parametric or non-parametric estimates for $F$ can be considered within the VSPGLM framework. An alternative estimator of $F$ could be through the use of a high dimensional kernel density estimator, where the choice of kernel function depends on a priori knowledge of the response. A similar estimator to this in the continuous case could be through the use of a finite mixture of multivariate normal distributions where the number of mixtures would need to be selected optimally for a given problem. An investigation of how the estimation procedure fits into the framework and asymptotic theory would be required, but although these estimators remove the convex hull constraint for finite samples, they cannot be used for distribution estimation in the case where the response components are of mixed types. Furthermore, primarily only estimation and inference of $\boldsymbol{\beta}$ is of interest so it's important that any alternative estimation of $F$ is not at the cost of the estimation

of $\boldsymbol{\beta}$. Other methods for finding a continuous distribution of the data include performing kernel smoothing on the probability masses after fitting the model rather than while estimating the model as this generally tends to perform better for higher dimensional cases. However, to do this the support of the distribution needs to be known prior which would result in further assumptions to the model.

Furthermore, if we are considering a longitudinal study a downside of VSPGLM is that it currently cannot handle an unbalanced study where observations with missing response components have to be removed. Methods for handling this data through such as stratification seen in Han et al. (2014), GEEs and linear-mixed models could be considered but again would require the framework to be altered to specifically handle longitudinal data.

The main advantage of VSPGLM is its applicability to any vector regression problem with minimal assumptions, so a more problem-specific VSPGLM framework does take away from the model's generality which was the initial motivation of the proposed model. However, the suggestions above are interesting topics for future research as they do resolve some drawbacks of VSPGLM.

Finally, VSPGLM has asymptotically valid standard errors when $F$ is a part of the multivariate exponential family or a distribution with a second moment that is a function of the first moment. Although this encompasses a wide class of multivariate distributions, if the variance is a function of the covariates then VSPGLM will be misspecified and thus requires an adjustment to be asymptotically valid. As a late addition, the sandwich estimator was proposed which is asymptotically valid under misspecification, but how the other asymptotic properties and the joint asymptotic normality with $F$ behave under misspecification was not explored in this thesis and is a topic for future investigation. Along with this, a wider suite of simulations under misspecification could also be performed to explore the efficiency of the sandwich estimator, given that for univariate regression problems, it may perform poorly for even moderate sample sizes (Kauermann and Carroll 2001).

Currently, there is no inference performed on the distribution $F$, but as the covariance function of $F$ is challenging to work with in practice, inference on $F$ could be performed by utilising bootstrapping. This can be done nonparametrically by generating pairs $(\boldsymbol{X}, \boldsymbol{Y})$ with replacement, although this would require a large number of samples to ensure the observed support is representative of the true support of the response. As an alternative, given that VSPGLM generates a fitted distribution for each combination of covariates, a form of parametric bootstrapping could be performed by generating the responses from the fitted distribution, or residual bootstrapping could also be employed as it retains the information from the covariates.

# Bibliography


Abraham, R., J.E. Marsden, and T. Ratiu (1988). *Manifolds, Tensor Analysis, and Applications*. 2nd ed. 1988, New York: Springer New York.

Cantelli, F.P (1933). "Sulla determinazione empirica delle leggi di probabilita". *Giorn. Ist. Ital. Attuari* 4, pp. 221–424.

Chiou, J. M. and H. G. Müller (1999). "Nonparametric Quasi-Likelihood". *The Annals of Statistics* 27, pp. 33–64.

Crowder, M. (1986). "On consistency and inconsistency of estimating equations". *Econometric Theory* 2, pp. 591–597.

Dennis, Gabriel (2021). "Semiparametric Vector Generalized Models: Computation and Estimation". Honours's Thesis. The University of Queensland.

Dewanji, A. and LP. Zhao (2002). "An optimal estimating equation with unspecified variances". *Sankhyā* 64, pp. 95–108.

Diciccio, T., P. Hall, and J. Romano (1991). "Empirical Likelihood is Bartlett-Correctable". *The Annals of statistics* 19, pp. 1053–1061.

Donsker, M. D. (1952). "Justification and extension of Doob's heuristic approach to the Komogorov-Smirnov theorems". *Ann. Maths. Statistics* 23, pp. 277–281.

Dunstan, P K., Foster S D., and Darnell R. (2011). "Model based grouping of species across environmental gradients". *Ecological Modelling* 222, pp. 955–963.

Fahrmeir, L. and T. Gerhard (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models*. New York, NY: Springer New York.

Fan, J. and I. Gijbels (1996). *Local polynomial modelling and its applications.* . London: Chapman and Hall.

Fitzmaurice, G M. and N M. Laird (1995). "'Regression Models for a Bivariate Discrete and Continuous Outcome with Clustering". *, Journal of the American Statistical Association* 90, pp. 845–852.

Gill, R. D., Y. Vardi, and J. A Wellner (1988). "Large Sample Theory of Empirical Distributions in Biased Sampling Models". *The Annals of statistics* 16, pp. 1069–1112.

Glivenko, V (1933). "Sulla determinazione empirica delle leggi di probabilita". *Giorn. Ist. Ital. Attuari* 4, pp. 92–99.

Gungor, Ayse Dilek. (2010). "Erratum to "An upper bound for the condition number of a matrix in spectral norm" [J. Comput. Appl. Math. 143 (2002) 141–144]". *Journal of computational and applied mathematics* 234, p. 316.

Han, Peisong., Peter X-K. Song, and Wang Lu (2014). "Longitudinal data analysis using the conditional empirical likelihood method". *The Canadian Journal of Statistics* 42, pp. 404–422.

Hiejima, Y. (1997). "Interpretation of the quasi-likelihood via the tilted exponential family". *Journal of the Japan Statistical Society* 27, pp. 157–164.

Huang, A. (2011). "An exponential tilt approach to generalized linear models". PhD thesis. The University of Chicago.

— (2014). "Joint Estimation of the Mean and Error Distribution in Generalized Linear Models". *Journal of the American Statistical Association* 109, pp. 186–196.

— (2017). "On generalised estimating equations for vector regression". *Australian New Zealand Journal of Statistics* 59, pp. 195–213.

Huang, A. and P. Rathouz (2012). "Propotional likelihood ratio models for mean regression". *Biometrika* 99, pp. 223–229.

— (2017). "Orthogonality of the Mean and Error Distribution in Generalized Linear Models". *Communications in Statistics-Theory and Methods* 46, pp. 3290–3296.

Hui, Francis. et al. (2013). "To mix or not to mix: comparing the predictive performance of mixture models vs. separate species distribution models". *Ecology* 94, pp. 1913–1919.

Jørgensen, B. (1987). "Exponential Dispersion Models". *Journal of the Royal Statistical Society. Series B* 49, pp. 127–162.

Jørgensen, B. and R.S. Labouriau (1992). *Famílias exponenciais e inferência teórica*. Rio de Janeiro: Instituto de Matemática Pura e Aplicada (IMPA).

Kauermann, G. and R. Carroll (2001). "A Note on the Efficiency of Sandwich Covariance Estimation". *Journal of the American Statistical Association* 96, pp. 1387–1396.

Kosorok, M R. (2008). *Introduction to Empirical Processes and Semiparametric Inference*. New York, NY: Springer-Verlag.

Liang, Kung-Yee. and Scott. Zeger (1986). "Longitudinal data analysis using generalized linear models". *Biometrika* 73, pp. 13–22.

Luo, Xiadong. and Wei Yann Tsai (2012). "A proportional likelihood ratio model". *Biometrika* 99, pp. 211–222.

Ma, J. (2010). "Positively constrained multiplicative iterative algorithm for maximum penalized likelihood tomographic reconstruction". *IEEE Transactions on Nuclear Science* 57, pp. 181–192.

Marchese, Scott. and Guoqing. Diao (2017). "Density ratio model for multivariate outcomes". *Journal of multivariate analysis* 154, pp. 249–261.

McCullagh, P. and J. Nelder (1989). *Generalized Linear Models*. London: Chapman & Hall.

McLachlan, G J. and D. Peel (2000). *Finite mixture models*. New York, USA: Wiley, New York.

Morris, C. (1982). "Natural Exponential Families with Quadratic Variance Functions". *The Annals of Statistics,* 10, pp. 65–80.

Murphy, S. A. and A. W. Van Der Vaart (2000). "On profile likelihood (with comments and a rejoinder by the authors)". *Journal of American Statistical Analysis* 95, pp. 449–485.

— (2001). "Semiparametric Mixtures in Case-Control Studies". *Journal of Multivariate Analysis* 79, pp. 1–32.

Nelder, J. and R. Wedderburn (1972). "Generalized Linear Models". *Journal of the Royal Statistical Society* 135, pp. 370–384.

Neumann, C G. et al. (2003). "Animal source foods improve dietary quality, micronutrient status, growth and cognitive function in Kenyan School Children: Background, Study Design and Baseline Findings." *The Journal of Nutrition* 133-11, 3941S–3949S.

Oliver, J C., K L. Prudic, and K. Collinge S (2006). "Boulder County Open Space Butterfly Diversity and Abundance". *Ecological Archives" E087-061". Ecology (Durham)* 87, pp. 1066–1066.

Owen, Art B. (2001). *Empirical Likelihood*. Boca Raton, Fla. : Chapman and Hall/CRC.

Puang-Ngern, Busayasachee . (2018). "The estimation of Semiparametric Generalized Linear Models". PhD thesis. Macquarie University.

Rathouz, P. and L. Gao (2009). "Generalized linear models with unspecified reference distribution". *Biostatistics* 10, pp. 205–218.

Raven, J C., J H. Court, and J. Raven (1995). *Manual for Raven's Progressive Matrices and Vocabulary Scales Summary of Contents of All Sections.* Oxford Psychologist Press: Oxford, UK.

Rosner, Bernard., Robert. Glynn, and Mei-Ling Lee (2006). "Extension of the rank sum test for clustered data: Two-group comparisons with group membership defined at the subunit level". *Biometrics* 62, pp. 1251–1259.

Rudin, W. (1973). *Functional Analysis*. New York : McGraw-Hill.

Song, Peter (2007). *Correlated data analysis: modeling, analytics, and applications.* Springer Science Business Media.

Trial, Sorbinil Retinopathy (1990). "A randomized trial of Sorbinil, an aldose reductose inhibitor, in diabetic retinopathy". *Archives of Ophthalmology* 108, pp. 1234–1244.

Tsiatis, AA. (2006). *Semiparametric Theory and Missing Data*. New York, USA: Springer.

Van Der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge, England: Cambridge University Press.

Van Der Vaart, A. W. and J.A. Wellner (1996). *Weak Convergence and Empirical Processes With Applications to Statistics*. 2nd ed. 2023. New York, USA: Cham : Springer International Publishing.

Vardi, Y. (1985). "Empirical Distributions in Selection Bias Models". *The Annals of statistics* 13, pp. 178–203.

Wald, A. (1949). "Note on the Consistency of the Maximum Likelihood Estimate". *The Annals of Mathematical Statistics* 20, pp. 595–601.

Wechsler, D. (2003). *Wechsler intelligence scale for children - Fourth edition (WISC-IV)*. San Antonio, TX: The Psychological Corporation.

Wedderburn, R. W. (1974). "Quasi-Likelihood Functions, Generalized Linear Models, and the Gauss-Newton Method". *Biometrika* 61, pp. 439–447.

Weiss, Robert E. (2005). *Modeling Longitudinal Data*. Springer.

Wellner, J A. (1981). "A Glivenko-Cantelli Theorem for Empirical Measures of Independent but Nonidentically Distributed Random Variables,"" *Stochastic Processes and Their Applications* 11, pp. 309–312.

Whaley, S E. et al. (2003). "The Impact of Dietary Intervention on the Cognitive Development of Kenyan School Children". *The Journal of Nutrition* 133-11, 3965S–3971S.

Wurm, M. and P. Rathouz (2018). "Semiparametric Generalized Linear Models with the gldrm Package". *The R journal* 10, p. 288.

Yee, T. (2015). *Vector generalized linear and additive models: with an implementation.* New York, NY : Springer New York : Imprint: Springer.

Zimmer, R.J. (1990). *Essential results of functional analysis*. Chicago: University of Chicago Press: Chicago Lectures in Mathematics.

# Appendix A

# Appendix

## A.1 Notation

Vectorising the Semiparametric GLM requires a choice of notation which aims to provide clarity while also simplifying arithmetic with vectors and matrices.

### A.1.1 Vector Notation

We need a way to differentiate when we are looking at a particular component or mean model for our given problem. Generally, vectors are denoted with bold variables and elements in $\mathbb{R}$ are denoted by non-bold variables.

Suppose $(\boldsymbol{X}_i, \boldsymbol{Y}_i) \in \mathcal{X} \times \mathcal{Y} \subset \mathbb{R}^q \times \mathbb{R}^K$ for $i = 1, 2, \ldots, n$. Components of $\boldsymbol{Y}_i$ is denoted by the vector

$$\boldsymbol{Y}_i = \begin{bmatrix} Y_{i(1)} \\ Y_{i(2)} \\ \vdots \\ Y_{i(K)} \end{bmatrix},$$

where $Y_{i(k)}$ is the $k$-th component of the $i$-th replication for $k = 1, 2, \ldots, K$, $i = 1, 2, \ldots, n$. The parenthesis in the subscript in this instance represents a particular component.

Note that between components, we can have the same mean model with the same coefficients. Therefore, we have $K' \leq K$ mean models to consider and let $q$ denote the total number of mean-model parameters. As a result, $\boldsymbol{\beta} \in \mathbb{R}^q$ is a vector of mean-model parameters. To split this up we will denote $\boldsymbol{\beta}_{(k')} \in \mathbb{R}^{q_k}$ to represent the coefficients associated with the $k'$-th mean model for $k' = 1, 2, \ldots K'$, where $\sum_{k=1}^{K'} q_k = q$. In other words

$$\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_{(1')} \\ \boldsymbol{\beta}_{(2')} \\ \vdots \\ \boldsymbol{\beta}_{(K')} \end{bmatrix} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_q \end{bmatrix}.$$

In the situation that $K = K'$, the $'$ notation is dropped in the index, but for explicit examples the appropriate index of the mean-model will be used.

Regarding the covariates, we have that $\boldsymbol{X} = (X_1, X_2, \ldots, X_q)^T \in \mathbb{R}^q$. Let $\boldsymbol{X}_{(k)}$ denote the covariates associated with the $k$-th component. Note that $\boldsymbol{X}_{(k)} \in \mathbb{R}^{q_k}$, matching the dimensionality of the parameter in the component's mean model $\boldsymbol{\beta}_{(k')}$. Importantly, $\boldsymbol{X} \neq (\boldsymbol{X}_{(1)}, \boldsymbol{X}_{(2)}, \ldots \boldsymbol{X}_{(K)})^T$ because there may be overlapping covariates between the components, implying that there are no covariates repeated in $\boldsymbol{X}$.

When we have a multivariate response, the mean function or user-specified inverse link function vectorized can be expressed as

$$\boldsymbol{\mu}_i = \boldsymbol{\mu}\left(\boldsymbol{X}_i^T\ \boldsymbol{\beta}\right) = \begin{bmatrix} \mu_{(1)}\left(\boldsymbol{X}_{i(1)}^T\ \boldsymbol{\beta}_{(1')}\right) \\ \mu_{(2)}\left(\boldsymbol{X}_{i(2)}^T\ \boldsymbol{\beta}_{(2')}\right) \\ \vdots \\ \mu_{(K)}\left(\boldsymbol{X}_{i(K)}^T\ \boldsymbol{\beta}_{(K')}\right) \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu}_{(1')} \\ \boldsymbol{\mu}_{(2')} \\ \vdots \\ \boldsymbol{\mu}_{(K')} \end{bmatrix}.$$

The first matrix form looks at the $K$ components, but it's important to note that the covariates $\boldsymbol{X}$ are with respect to the component, but the coefficients $\boldsymbol{\beta}$ are with respect to the components mean model, whose index may not necessarily match the components. In the second matrix form it groups together the mean functions which are a part of the same mean model (shared coefficients). This is detailed below and will be useful for certain derivations.

Although messy, for components belonging to the $k'$-th mean model for $k' = 1, 2, \ldots, K'$, they are given an index in some index set $I_{k'}$ of size $M_{k'}$ using the $k'$-th mean model for $k' = 1, 2, \ldots, K'$, then

$$\boldsymbol{\mu}_{(k')} = \begin{bmatrix} \mu_{(k',1)}\left(\boldsymbol{X}_{(k',1)}^T \boldsymbol{\beta}_{(k')}\right) \\ \mu_{(k',2)}\left(\boldsymbol{X}_{(k',2)}^T \boldsymbol{\beta}_{(k')}\right) \\ \vdots \\ \mu_{(k',M_{k'})}\left(\boldsymbol{X}_{(k',M_{k'})}^T \boldsymbol{\beta}_{(k')}\right) \end{bmatrix}.$$

If the arguments for any mean variables $\boldsymbol{\mu}$ are omitted, it is assumed that the mean function is evaluated at the correct arguments relative to it's mean model and component. This also holds for other variables, where an omission of it's argument implies evaluation at the usual parameters. If we are evaluating a variable at the true parameters, a * will be used on the superscript.

Finally, let us denote $\boldsymbol{X}_{(k')}$ to be the matrix of covaraites associated with the $k'$-th mean model for $k' = 1, 2, \ldots, K'$. By construction this is a $q_k \times M_{k'}$ matrix given by

$$\boldsymbol{X}_{(k')} = \left[\boldsymbol{X}_{(k',1)}, \boldsymbol{X}_{(k',2)}, \ldots, \boldsymbol{X}_{(k',M_{k'})}\right]$$

This notation is particularly useful for deriving the score function for $\boldsymbol{\beta}$ and note that each column may be the same or different. Note that then, $\boldsymbol{X'} = (\boldsymbol{X}_{(1')}, \boldsymbol{X}_{(2')}, \ldots \boldsymbol{X}_{(K')}) \in \mathbb{R}^Q$ where $Q \geq q$. $Q$ can be thought of as the total number of covariates to be plugged into our model, and is finite by construction.

Generally, the use of a single apostrophe on subscripts means the variable is associated with the mean model, and otherwise are associated with the components. Often instead of considering the $i$-th replication $(\boldsymbol{X}_i, \boldsymbol{Y}_i)$, we consider a generic $(\boldsymbol{X}, \boldsymbol{Y})$ copy and drop the $i$ sub-script in our notation.

### A.1.2 Vector and Matrix Derivatives

Note that there are two main conventions for vector and matrix derivatives, namely the numerator and denominator notation. This thesis will be consistent with Huang (2014) which considers vectors

as column vectors. For any matrix calculus, we adopt the denominator notation. We will explicitly outline the notation below.

**A.1.2.a.** For a scalar $y \in \mathbb{R}$ and vector $\boldsymbol{x} \in \mathbb{R}^n$

$$\frac{\partial y}{\partial \boldsymbol{x}} = \begin{bmatrix} \dfrac{\partial y}{\partial x_1} \\ \vdots \\ \dfrac{\partial y}{\partial x_n} \end{bmatrix} \in \mathbb{R}^{n \times 1}$$

**A.1.2.b.** For a vector $\boldsymbol{y} \in \mathbb{R}^m$ and a scalar $x \in \mathbb{R}$

$$\frac{\partial \boldsymbol{y}}{\partial x} = \begin{bmatrix} \dfrac{\partial y_1}{\partial x}, \ldots, \dfrac{\partial y_m}{\partial x} \end{bmatrix} = \begin{bmatrix} \dfrac{\partial y_1}{\partial x} \\ \vdots \\ \dfrac{\partial y_m}{\partial x} \end{bmatrix}^T \in \mathbb{R}^{1 \times n}$$

**A.1.2.c.** For a vector $\boldsymbol{y} \in \mathbb{R}^m$ and a vector $\boldsymbol{x} \in \mathbb{R}^n$

$$\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}} = \begin{bmatrix} \dfrac{\partial y_1}{\partial x_1} & \dfrac{\partial y_2}{\partial x_1} & \cdots & \dfrac{\partial y_m}{\partial x_1} \\ \dfrac{\partial y_1}{\partial x_2} & \dfrac{\partial y_2}{\partial x_2} & \cdots & \dfrac{\partial y_m}{\partial x_2} \\ \vdots & \vdots & \cdots & \vdots \\ \dfrac{\partial y_1}{\partial x_n} & \dfrac{\partial y_2}{\partial x_n} & \cdots & \dfrac{\partial y_m}{\partial x_n} \end{bmatrix} \in \mathbb{R}^{n \times m}$$

**A.1.2.d.** For a scalar $y \in \mathbb{R}$ and a matrix $\boldsymbol{X} \in \mathbb{R}^{p \times q}$

$$\frac{\partial y}{\partial \boldsymbol{X}} = \begin{bmatrix} \dfrac{\partial y}{\partial x_{11}} & \dfrac{\partial y}{\partial x_{21}} & \cdots & \dfrac{\partial y}{\partial x_{p1}} \\ \dfrac{\partial y}{\partial x_{12}} & \dfrac{\partial y}{\partial x_{22}} & \cdots & \dfrac{\partial y}{\partial x_{p2}} \\ \vdots & \vdots & \cdots & \vdots \\ \dfrac{\partial y}{\partial x_{1q}} & \dfrac{\partial y}{\partial x_{2q}} & \cdots & \dfrac{\partial y}{\partial x_{pq}} \end{bmatrix} \in \mathbb{R}^{p \times q}$$

**A.1.2.e.** For a scalar $y \in \mathbb{R}$ and a vector $\boldsymbol{x} \in \mathbb{R}^n$, the Hessian is given by

$$\begin{aligned} \frac{\partial y}{\partial \boldsymbol{x} \partial \boldsymbol{x}^T} &= \frac{\partial}{\partial \boldsymbol{x}} \begin{bmatrix} \dfrac{\partial y}{\partial \boldsymbol{x}} \end{bmatrix} \\ &= \begin{bmatrix} \frac{\partial y}{\partial x_1 \partial x_1} & \frac{\partial y}{\partial x_1 \partial x_2} & \cdots & \frac{\partial y}{\partial x_1 \partial x_n} \\ \frac{\partial y}{\partial x_2 \partial x_1} & \frac{\partial y}{\partial x_2 \partial x_2} & \cdots & \frac{\partial y}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y}{\partial x_n \partial x_1} & \frac{\partial y}{\partial x_n \partial x_2} & \cdots & \frac{\partial y}{\partial x_n \partial x_n} \end{bmatrix} \in \mathbb{R}^{n \times n} \end{aligned}$$

### A.1.3 Vector Integral Notation

A vector of integrals can be represented as an integral of vectors. Suppose $u(\boldsymbol{y})$ is some real-valued function,

$$\int_{\mathcal{Y}} \boldsymbol{y} \; u(\boldsymbol{y}) \; dF(\boldsymbol{y}) = \begin{bmatrix} \int_{\mathcal{Y}} y_1 u(\boldsymbol{y}) \; dF(\boldsymbol{y}) \\ \vdots \\ \int_{\mathcal{Y}} y_K u(\boldsymbol{y}) \; dF(\boldsymbol{y}) \end{bmatrix}$$

### A.1.4 Norms

We differentiate between the norms used with the following notation

**A.1.4.a.** For a vector $\boldsymbol{x} \in \mathbb{R}^n$, $||\boldsymbol{x}||_2$ is the Euclidean Norm where $||\boldsymbol{x}||_2 = \sqrt{\sum_{j=1}^{n} x_j^2}$

**A.1.4.b.** For a matrix $A \in \mathbb{R}^{n \times m}$, $||A||_F$ is the Frobenius norm where $||A||_F = \sqrt{\sum_{p=1}^{n}\sum_{q=1}^{m} |a_{pq}|^2}$

**A.1.4.c.** For a matrix $A \in \mathbb{R}^{n \times m}$, $||A||_2$ is the Spectral norm where $||A||_2 = \sigma_{\max}(A)$ is the largest singular value.

### A.1.5 Notation for the Different Distributions

For clarity let us summarize the notation for the different distributions used throughout the thesis. We denote the joint distribution of the vector response variable conditional on the covariates generally by $F(\boldsymbol{y}|\boldsymbol{x})$, where the true distribution is denoted as $F^*(\boldsymbol{y}|\boldsymbol{x})$.

VSPGLM supposes that $F^*(\boldsymbol{y}|\boldsymbol{x})$ is in the multivariate exponential family and thus can be re-expressed using an exponential tilt representation $F_{\boldsymbol{\theta}}(\boldsymbol{y})$ with densities of the form

$$dF_{\boldsymbol{\theta}}(\boldsymbol{y}) = \exp\left(b(\boldsymbol{X};\boldsymbol{\beta},F) + \boldsymbol{\theta}(\boldsymbol{X};\boldsymbol{\beta},F)^T\boldsymbol{y}\right) dF(\boldsymbol{y})$$

where $F(\boldsymbol{y})$ is some underlying multivariate reference distribution which we estimate using $\hat{F}$. Note that conditioning on the covariates is implicitly here as $\boldsymbol{\theta}$ is a function of the covariates $\boldsymbol{X}$. Therefore, in the case where $F^*(\boldsymbol{y}|\boldsymbol{x})$ is indeed in the multivariate exponential family, then it has densities of the form

$$dF^*(\boldsymbol{y}|\boldsymbol{x}) = \exp\left(b(\boldsymbol{X};\boldsymbol{\beta}^*,F^*) + \boldsymbol{\theta}(\boldsymbol{X};\boldsymbol{\beta}^*,F^*)^T\boldsymbol{y}\right) dF^*(\boldsymbol{y})$$

Thus, $\hat{F}$ is an estimate of the underlying distribution $F^*(\boldsymbol{y})$ or $F^*$.

## A.2 Model Assumptions

Similar to Huang (2014), the asymptotic results derived require the following three conditions.

**Condition A.2.1.** The response space $\mathcal{Y}$ is contained in a closed finite hyperrectangle in $\mathbb{R}^K$, where $Y_k \in [L_k, U_k]$ in $\mathbb{R}$ for $k = 1, 2, \ldots, K$, and $\mathcal{X}$ is contained in a closed finite hyperrectangle in $\mathbb{R}^q$.

Most covariate and response variables have natural bounds, so in practice, we claim this assumption isn't too restrictive. As a result, all probability measures on compact spaces are necessarily tight.

**Condition A.2.2.** There exists $\delta_1 > 0$ such that $\boldsymbol{\mu}$ maps onto $\mathcal{Y}$, both the first and second derivatives of $\boldsymbol{\mu}$ exist and are continuous on the space

$$\mathcal{X} \times \{\boldsymbol{\beta} \in \mathbb{R}^q : ||\boldsymbol{\beta} - \boldsymbol{\beta}^*||_2 \leq \delta_1\}$$

This concerns the link function, stating that the mean $\boldsymbol{\mu}$ maps into the response space $\mathcal{Y}$ and is sufficiently smooth in some neighbourhood close to the true parameters. This restriction of only considering neighbourhoods of true mean models allows for locally, but not globally well-behaved link functions (eg. inverse link which is discontinuous at the origin).

As the space above is compact because $\mathcal{X}$ is assumed to be compact, by the Extreme Value Theorem we have that the first and second derivative must be bounded by some constant on this space.

**A.2.2.a.** $|\mu_k'(\boldsymbol{x}_{(k)}^T \boldsymbol{\beta}_{(k')})|$ are bounded by some constant $M_{1k} < \infty$ on $\mathcal{X} \times \{\boldsymbol{\beta} \in \mathbb{R}^q : ||\boldsymbol{\beta} - \boldsymbol{\beta^*}||_2 \leq \delta_1\}$ for $k = 1, 2, \ldots, K$. Let $M_1 = \max_{1 \leq k \leq K} M_{1k} < \infty$

**A.2.2.b.** $|\mu_k''(\boldsymbol{x}_{(k)}^T \boldsymbol{\beta}_{(k')})|$ are bounded by some constant $M_{2i} < \infty$ on $\mathcal{X} \times \{\boldsymbol{\beta} \in \mathbb{R}^q : ||\boldsymbol{\beta} - \boldsymbol{\beta^*}||_2 \leq \delta_1\}$ for $k = 1, 2, \ldots, K$. Let $M_2 = \max_{1 \leq k \leq K} M_{2k} < \infty$

**Condition A.2.3.** There exists $\delta_2, V_1 > 0, V_{2k} > 0$ for $k = 1, 2, \ldots, K$ such that $|\det(\Sigma_Y)| \geq V_1$ and $\mathrm{Var}(Y_k) > V_{2k}$ on $\mathcal{X} \times \{(\boldsymbol{\beta}, F) \in \mathbb{R}^q \times \mathcal{F}_u : ||\boldsymbol{\beta} - \boldsymbol{\beta}^*, F - F^*||_2 \leq \delta_2\}$ for $k = 1, 2, \ldots, K$.

This ensures that the conditional covariance matrix is invertible and to ensure that no covariate value leads to a degenerative model where the conditional variance function gets arbitrarily small for any given component.

## A.3 Examples for Score Expression and its Covariance for $\beta$

As found in section 2.3.1, we found that the score function for $\boldsymbol{\beta}$ is given by

$$S_{\boldsymbol{\beta},F}(\boldsymbol{X},\boldsymbol{Y}) = D(\boldsymbol{X};\boldsymbol{\beta})\Sigma_{\boldsymbol{Y}}^{-1}(\boldsymbol{X};\boldsymbol{\beta},F)\left(\boldsymbol{Y}-\boldsymbol{\mu}\right)$$
$$\text{Cov}\left(S_{\boldsymbol{\beta},F}(\boldsymbol{X},\boldsymbol{Y}), S_{\boldsymbol{\beta},F}(\boldsymbol{X},\boldsymbol{Y})\right) = \mathbb{E}^{\mathcal{X}}\left[D(\boldsymbol{X};\boldsymbol{\beta})\Sigma_{\boldsymbol{Y}}^{-1}(\boldsymbol{X};\boldsymbol{\beta},F)D(\boldsymbol{X};\boldsymbol{\beta})^T\right]$$

Let's explore some generic examples for $K=3$ components to see how these expressions are evaluated.

### A.3.1 Case where no coefficients are shared, $K=3$ components

Suppose that $K' = K = 3$, meaning that no coefficients and mean models are shared between the components. Thus, for a generic $(\boldsymbol{X},\boldsymbol{Y})$ data pair, supposing that $\boldsymbol{X}_{(k)}, \boldsymbol{\beta}_{(k)} \in \mathbb{R}^{q_k}$ for some integer $q_k \geq 1$, for $k = 1,2,3$

$$Y_{(1)} = \mu_{(1)}\left(\boldsymbol{X}_{(1)}^T\boldsymbol{\beta}_{(1)}\right)$$
$$Y_{(2)} = \mu_{(2)}\left(\boldsymbol{X}_{(2)}^T\boldsymbol{\beta}_{(2)}\right)$$
$$Y_{(3)} = \mu_{(3)}\left(\boldsymbol{X}_{(3)}^T\boldsymbol{\beta}_{(3)}\right)$$

Here note that there may be covariates in each $\boldsymbol{X}_{(k)}$ which are shared, but denoting $q = q_1+q_2+q_3$, we have $q$ unknown mean model parameters $\boldsymbol{\beta}$ to be estimated. Here we will consider two cases, first when there is covariance between the responses and the seocnd when all the responses are independent.

Note that we need to take an inverse of a $3\times 3$ matrix. Let

$$\Sigma_{\boldsymbol{Y}} = \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix}$$

$$\Sigma_{\boldsymbol{Y}}^{-1} = \frac{1}{a(ei-fh)-b(di-fg)-(dh-eg)}\begin{bmatrix} ei-fh & ch-bi & cf-ce \\ fg-di & ai-cg & cd-af \\ dh-eg & bg-ah & ae-bd \end{bmatrix} = \begin{bmatrix} \Sigma_{11}^{-1} & \Sigma_{12}^{-1} & \Sigma_{13}^{-1} \\ \Sigma_{21}^{-1} & \Sigma_{22}^{-1} & \Sigma_{23}^{-1} \\ \Sigma_{31}^{-1} & \Sigma_{32}^{-1} & \Sigma_{33}^{-1} \end{bmatrix}$$

Therefore, the score function for $\boldsymbol{\beta}$ is given by

$$S_{\boldsymbol{\beta},F}(\boldsymbol{X},\boldsymbol{Y}) = \begin{bmatrix} \boldsymbol{X}_{(1)}\mu'_{(1)} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{X}_{(2)}\mu'_{(2)} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{X}_{(3)}\mu'_{(3)} \end{bmatrix}\begin{bmatrix} \Sigma_{11}^{-1} & \Sigma_{12}^{-1} & \Sigma_{13}^{-1} \\ \Sigma_{21}^{-1} & \Sigma_{22}^{-1} & \Sigma_{23}^{-1} \\ \Sigma_{31}^{-1} & \Sigma_{32}^{-1} & \Sigma_{33}^{-1} \end{bmatrix}\begin{bmatrix} Y_{(1)}-\mu_{(1)} \\ Y_{(2)}-\mu_{(2)} \\ Y_{(3)}-\mu_{(3)} \end{bmatrix}$$
$$= \begin{bmatrix} \left(\boldsymbol{X}_{(1)}\mu'_{(1)}\right)\sum_{k=1}^3 \Sigma_{1k}^{-1}\left(Y_{(k)}-\mu_{(k)}\right) \\ \left(\boldsymbol{X}_{(2)}\mu'_{(2)}\right)\sum_{k=1}^3 \Sigma_{2k}^{-1}\left(Y_{(k)}-\mu_{(k)}\right) \\ \left(\boldsymbol{X}_{(3)}\mu'_{(3)}\right)\sum_{k=1}^3 \Sigma_{3k}^{-1}\left(Y_{(k)}-\mu_{(k)}\right) \end{bmatrix}$$

The expression $D(\boldsymbol{X};\boldsymbol{\beta})\Sigma_{\boldsymbol{Y}}^{-1}(\boldsymbol{X};\boldsymbol{\beta},F)D(\boldsymbol{X};\boldsymbol{\beta})^T$ becomes

$$= \begin{bmatrix} \boldsymbol{X}_{(1)}\mu'_{(1)} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{X}_{(2)}\mu'_{(2)} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{X}_{(3)}\mu'_{(3)} \end{bmatrix}\begin{bmatrix} \Sigma_{11}^{-1} & \Sigma_{12}^{-1} & \Sigma_{13}^{-1} \\ \Sigma_{21}^{-1} & \Sigma_{22}^{-1} & \Sigma_{23}^{-1} \\ \Sigma_{31}^{-1} & \Sigma_{32}^{-1} & \Sigma_{33}^{-1} \end{bmatrix}\begin{bmatrix} \boldsymbol{X}_{(1)}\mu'_{(1)} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{X}_{(2)}\mu'_{(2)} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{X}_{(3)}\mu'_{(3)} \end{bmatrix}^T$$
$$= \begin{bmatrix} \boldsymbol{X}_{(1)}\boldsymbol{X}_{(1)}^T(\mu'_{(1)})^2\Sigma_{11}^{-1} & \boldsymbol{X}_{(1)}\boldsymbol{X}_{(2)}^T\mu'_{(1)}\mu'_{(2)}\Sigma_{12}^{-1} & \boldsymbol{X}_{(1)}\boldsymbol{X}_{(3)}^T\mu'_{(1)}\mu'_{(3)}\Sigma_{13}^{-1} \\ \boldsymbol{X}_{(2)}\boldsymbol{X}_{(1)}^T\mu'_{(1)}\mu'_{(2)}\Sigma_{21}^{-1} & \boldsymbol{X}_{(2)}\boldsymbol{X}_{(2)}^T(\mu'_{(2)})^2\Sigma_{22}^{-1} & \boldsymbol{X}_{(2)}\boldsymbol{X}_{(3)}^T\mu'_{(2)}\mu'_{(3)}\Sigma_{23}^{-1} \\ \boldsymbol{X}_{(3)}\boldsymbol{X}_{(1)}^T\mu'_{(1)}\mu'_{(3)}\Sigma_{31}^{-1} & \boldsymbol{X}_{(3)}\boldsymbol{X}_{(2)}^T\mu'_{(2)}\mu'_{(3)}\Sigma_{32}^{-1} & \boldsymbol{X}_{(3)}\boldsymbol{X}_{(3)}^T(\mu'_{(3)})^2\Sigma_{33}^{-1} \end{bmatrix}$$

Therefore, the expression $\text{Cov}\left(S_{\boldsymbol{\beta},F}(\boldsymbol{X},\boldsymbol{Y}), S_{\boldsymbol{\beta},F}(\boldsymbol{X},\boldsymbol{Y})\right)$ is given by

$$= \begin{bmatrix} \mathbb{E}^{\mathcal{X}}\left(\boldsymbol{X}_{(1)}\boldsymbol{X}_{(1)}^T(\mu'_{(1)})^2\Sigma_{11}^{-1}\right) & \mathbb{E}^{\mathcal{X}}\left(\boldsymbol{X}_{(1)}\boldsymbol{X}_{(2)}^T\mu'_{(1)}\mu'_{(2)}\Sigma_{12}^{-1}\right) & \mathbb{E}^{\mathcal{X}}\left(\boldsymbol{X}_{(1)}\boldsymbol{X}_{(3)}^T\mu'_{(1)}\mu'_{(3)}\Sigma_{13}^{-1}\right) \\ \mathbb{E}^{\mathcal{X}}\left(\boldsymbol{X}_{(2)}\boldsymbol{X}_{(1)}^T\mu'_{(1)}\mu'_{(2)}\Sigma_{21}^{-1}\right) & \mathbb{E}^{\mathcal{X}}\left(\boldsymbol{X}_{(2)}\boldsymbol{X}_{(2)}^T(\mu'_{(2)})^2\Sigma_{22}^{-1}\right) & \mathbb{E}^{\mathcal{X}}\left(\boldsymbol{X}_{(2)}\boldsymbol{X}_{(3)}^T\mu'_{(2)}\mu'_{(3)}\Sigma_{23}^{-1}\right) \\ \mathbb{E}^{\mathcal{X}}\left(\boldsymbol{X}_{(3)}\boldsymbol{X}_{(1)}^T\mu'_{(1)}\mu'_{(3)}\Sigma_{31}^{-1}\right) & \mathbb{E}^{\mathcal{X}}\left(\boldsymbol{X}_{(3)}\boldsymbol{X}_{(2)}^T\mu'_{(2)}\mu'_{(3)}\Sigma_{32}^{-1}\right) & \mathbb{E}^{\mathcal{X}}\left(\boldsymbol{X}_{(3)}\boldsymbol{X}_{(3)}^T(\mu'_{(3)})^2\Sigma_{33}^{-1}\right) \end{bmatrix}$$

Note that in the case where all components of $\boldsymbol{Y}$ are **independent**, then

$$\Sigma_{ii}^{-1} = \left[\text{Var}Y_{(i)}|\boldsymbol{X}\right]^{-1},\ i = 1,2,3, \quad \Sigma_{ij}^{-1} = 0,\ i \neq j.$$

Thus, in the case of independence,

$$S_{\boldsymbol{\beta},F}(\boldsymbol{X},\boldsymbol{Y}) = \begin{bmatrix} \boldsymbol{X}_{(1)}\mu'_{(1)}\left[\text{Var}(Y_{(1)}|\boldsymbol{X})\right]^{-1}\left(Y_{(1)} - \mu_{(1)}\right) \\ \boldsymbol{X}_{(2)}\mu'_{(2)}\left[\text{Var}(Y_{(2)}|\boldsymbol{X})\right]^{-1}\left(Y_{(2)} - \mu_{(2)}\right) \\ \boldsymbol{X}_{(3)}\mu'_{(3)}\left[\text{Var}(Y_{(3)}|\boldsymbol{X})\right]^{-1}\left(Y_{(3)} - \mu_{(3)}\right) \end{bmatrix}$$

Note that we recover the score function for $\boldsymbol{\beta}$ of the usual least squares form in each component. The expression $D(\boldsymbol{X};\boldsymbol{\beta})\Sigma_{\boldsymbol{Y}}^{-1}(\boldsymbol{X};\boldsymbol{\beta},F)D(\boldsymbol{X};\boldsymbol{\beta})^T$ becomes

$$= \begin{bmatrix} \boldsymbol{X}_{(1)}\boldsymbol{X}_{(1)}^T(\mu'_{(1)})^2\left[\text{Var}(Y_{(1)}|\boldsymbol{X})\right]^{-1} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{X}_{(2)}\boldsymbol{X}_{(2)}^T(\mu'_{(2)})^2\left[\text{Var}(Y_{(2)}|\boldsymbol{X})\right]^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{X}_{(3)}\boldsymbol{X}_{(3)}^T(\mu'_{(3)})^2\left[\text{Var}(Y_{(3)}|\boldsymbol{X})\right]^{-1} \end{bmatrix}$$

Therefore, the expression $\text{Cov}\left(S_{\boldsymbol{\beta},F}(\boldsymbol{X},\boldsymbol{Y}), S_{\boldsymbol{\beta},F}(\boldsymbol{X},\boldsymbol{Y})\right)$ is given by

$$= \begin{bmatrix} \mathbb{E}^{\mathcal{X}}\left[\boldsymbol{X}_{(1)}\boldsymbol{X}_{(1)}^T(\mu'_{(1)})^2\left[\text{Var}(Y_{(1)}|\boldsymbol{X})\right]^{-1}\right] & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \mathbb{E}^{\mathcal{X}}\left[\boldsymbol{X}_{(2)}\boldsymbol{X}_{(2)}^T(\mu'_{(2)})^2\left[\text{Var}(Y_{(2)}|\boldsymbol{X})\right]^{-1}\right] & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \mathbb{E}^{\mathcal{X}}\left[\boldsymbol{X}_{(3)}\boldsymbol{X}_{(3)}^T(\mu'_{(3)})^2\left[\text{Var}(Y_{(3)}|\boldsymbol{X})\right]^{-1}\right] \end{bmatrix}$$

Again, note as expected for each component, we recover the covariance function given in Lemma 1.1 of Huang (2014). For this particular case it's simple so can see how it extends to arbitrary $K$ components. This implies that if the responses are independent, jointly estimating the coefficients of each component is equivalent to applying the univariate method in Huang (2014) to the coefficients in each component.

### A.3.2 Case where all coefficients are shared, $K = 3$ components

Suppose that $K' = 1, K = 3$, meaning that all coefficients and mean models are shared between the components. Thus, for a generic $(\boldsymbol{X},\boldsymbol{Y})$ data pair, supposing that $\boldsymbol{X}_{(k)}, \boldsymbol{\beta}_{(k)} \in \mathbb{R}^{q_1}$ for some integer $q_1 \geq 1$, for $k = 1,2,3$

$$\begin{aligned} Y_{(1)} &= \mu_{(1)}\left(\boldsymbol{X}_{(1)}^T\boldsymbol{\beta}\right) \\ Y_{(2)} &= \mu_{(2)}\left(\boldsymbol{X}_{(2)}^T\boldsymbol{\beta}\right) \\ Y_{(3)} &= \mu_{(3)}\left(\boldsymbol{X}_{(3)}^T\boldsymbol{\beta}\right) \end{aligned}$$

Here all covariates in each $\boldsymbol{X}_{(k)}$ are shared and we have $q_1$ unknown mean model parameters $\boldsymbol{\beta}$ to be estimated. Here we will consider two cases, first when there is covariance between the responses and the second when all the responses are independent.

Therefore, the score function for $\boldsymbol{\beta}$ is given by

$$\begin{aligned} S_{\boldsymbol{\beta},F}(\boldsymbol{X},\boldsymbol{Y}) &= \left[\boldsymbol{X}_{(1')}\right] \begin{bmatrix} \mu'_{(1)} & 0 & 0 \\ 0 & \mu'_{(2)} & 0 \\ 0 & 0 & \mu'_{(3)} \end{bmatrix} \begin{bmatrix} \Sigma_{11}^{-1} & \Sigma_{12}^{-1} & \Sigma_{13}^{-1} \\ \Sigma_{21}^{-1} & \Sigma_{22}^{-1} & \Sigma_{23}^{-1} \\ \Sigma_{31}^{-1} & \Sigma_{32}^{-1} & \Sigma_{33}^{-1} \end{bmatrix} \begin{bmatrix} Y_{(1)} - \mu_{(1)} \\ Y_{(2)} - \mu_{(2)} \\ Y_{(3)} - \mu_{(3)} \end{bmatrix} \\ &= \begin{bmatrix} \boldsymbol{X}_{(1)}\mu'_{(1)} & \boldsymbol{X}_{(2)}\mu'_{(2)} & \boldsymbol{X}_{(3)}\mu'_{(3)} \end{bmatrix} \begin{bmatrix} \Sigma_{11}^{-1} & \Sigma_{12}^{-1} & \Sigma_{13}^{-1} \\ \Sigma_{21}^{-1} & \Sigma_{22}^{-1} & \Sigma_{23}^{-1} \\ \Sigma_{31}^{-1} & \Sigma_{32}^{-1} & \Sigma_{33}^{-1} \end{bmatrix} \begin{bmatrix} Y_{(1)} - \mu_{(1)} \\ Y_{(2)} - \mu_{(2)} \\ Y_{(3)} - \mu_{(3)} \end{bmatrix} \\ &= \left[\sum_{j=1}^{3} \left(\boldsymbol{X}_{(j)}\mu'_{(j)}\right) \sum_{k=1}^{3} \Sigma_{jk}^{-1} \left(Y_{(k)} - \mu_{(k)}\right)\right] \end{aligned}$$

The expression $D(\boldsymbol{X};\boldsymbol{\beta})\Sigma_{\boldsymbol{Y}}^{-1}(\boldsymbol{X};\boldsymbol{\beta},F)D(\boldsymbol{X};\boldsymbol{\beta})^T$ becomes

$$\begin{aligned} &= \begin{bmatrix} \boldsymbol{X}_{(1)}\mu'_{(1)} & \boldsymbol{X}_{(2)}\mu'_{(2)} & \boldsymbol{X}_{(3)}\mu'_{(3)} \end{bmatrix} \begin{bmatrix} \Sigma_{11}^{-1} & \Sigma_{12}^{-1} & \Sigma_{13}^{-1} \\ \Sigma_{21}^{-1} & \Sigma_{22}^{-1} & \Sigma_{23}^{-1} \\ \Sigma_{31}^{-1} & \Sigma_{32}^{-1} & \Sigma_{33}^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{X}_{(1)}\mu'_{(1)} & \boldsymbol{X}_{(2)}\mu'_{(2)} & \boldsymbol{X}_{(3)}\mu'_{(3)} \end{bmatrix}^T \\ &= \left[\sum_{j=1}^{3} \sum_{k=1}^{3} \boldsymbol{X}_{(j)}\boldsymbol{X}_{(k)}^T \mu'_{(j)}\mu'_{(k)}\Sigma_{jk}^{-1}\right] \end{aligned}$$

Therefore, the expression $\mathrm{Cov}\left(S_{\boldsymbol{\beta},F}(\boldsymbol{X},\boldsymbol{Y}), S_{\boldsymbol{\beta},F}(\boldsymbol{X},\boldsymbol{Y})\right)$ is given by

$$\mathrm{Cov}\left(S_{\boldsymbol{\beta},F}(\boldsymbol{X},\boldsymbol{Y}), S_{\boldsymbol{\beta},F}(\boldsymbol{X},\boldsymbol{Y})\right) = \left[\sum_{j=1}^{3} \sum_{k=1}^{3} \mathbb{E}^{\mathcal{X}}\left(\boldsymbol{X}_{(j)}\boldsymbol{X}_{(k)}^T \mu'_{(j)}\mu'_{(k)}\Sigma_{jk}^{-1}\right)\right]$$

We can see that the contribution of each component is additive. Similar to the prior, if we consider the case of **independence**,

$$\begin{aligned} S_{\boldsymbol{\beta},F}(\boldsymbol{X},\boldsymbol{Y}) &= \left[\sum_{k=1}^{3} \boldsymbol{X}_{(k)}\mu'_{(k)}\Sigma_{kk}^{-1}\left(Y_{(k)} - \mu_{(k)}\right)\right] \\ \mathrm{Cov}\left(S_{\boldsymbol{\beta},F}(\boldsymbol{X},\boldsymbol{Y}), S_{\boldsymbol{\beta},F}(\boldsymbol{X},\boldsymbol{Y})\right) &= \left[\sum_{k=1}^{3} \mathbb{E}^{\mathcal{X}}\left(\boldsymbol{X}_{(k)}\boldsymbol{X}_{(k)}^T (\mu'_{(k)})^2\Sigma_{kk}^{-1}\right)\right] \end{aligned}$$

### A.3.3 Case where some coefficients are shared, $K = 3$ components

Suppose that $K' = 2, K = 3$, where WLOG assume the first two components share the same components and mean model. Thus, for a generic $(\boldsymbol{X},\boldsymbol{Y})$ data pair, supposing that $\boldsymbol{X}_{(1)}, \boldsymbol{X}_{(2)}\boldsymbol{\beta}_{(1')} \in \mathbb{R}^{q_1}$ and $\boldsymbol{X}_{(3)}, \boldsymbol{\beta}_{(2')} \in \mathbb{R}^{q_2}$,

$$\begin{aligned} Y_{(1)} &= \mu_{(1)}\left(\boldsymbol{X}_{(1)}^T\boldsymbol{\beta}_{(1)}\right) \\ Y_{(2)} &= \mu_{(2)}\left(\boldsymbol{X}_{(2)}^T\boldsymbol{\beta}_{(1)}\right) \\ Y_{(3)} &= \mu_{(3)}\left(\boldsymbol{X}_{(3)}^T\boldsymbol{\beta}_{(2)}\right) \end{aligned}$$

The total number of unknown mean parameters in $\boldsymbol{\beta}$ is $q = q_1 + q_2$. Again, first consider the case where there is covariance between the responses. The score function for $\boldsymbol{\beta}$ is given by

$$\begin{aligned}
S_{\boldsymbol{\beta},F}(\boldsymbol{X},\boldsymbol{Y}) &= \begin{bmatrix} \boldsymbol{X}_{(1')} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{X}_{(2')} \end{bmatrix} \begin{bmatrix} \mu'_{(1)} & 0 & 0 \\ 0 & \mu'_{(2)} & 0 \\ 0 & 0 & \mu'_{(3)} \end{bmatrix} \begin{bmatrix} \Sigma_{11}^{-1} & \Sigma_{12}^{-1} & \Sigma_{13}^{-1} \\ \Sigma_{21}^{-1} & \Sigma_{22}^{-1} & \Sigma_{23}^{-1} \\ \Sigma_{31}^{-1} & \Sigma_{32}^{-1} & \Sigma_{33}^{-1} \end{bmatrix} \begin{bmatrix} Y_{(1)} - \mu_{(1)} \\ Y_{(2)} - \mu_{(2)} \\ Y_{(3)} - \mu_{(3)} \end{bmatrix} \\
&= \begin{bmatrix} \boldsymbol{X}_{(1)} & \boldsymbol{X}_{(2)} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{X}_{(3)} \end{bmatrix} \begin{bmatrix} \mu'_{(1)} & 0 & 0 \\ 0 & \mu'_{(2)} & 0 \\ 0 & 0 & \mu'_{(3)} \end{bmatrix} \begin{bmatrix} \Sigma_{11}^{-1} & \Sigma_{12}^{-1} & \Sigma_{13}^{-1} \\ \Sigma_{21}^{-1} & \Sigma_{22}^{-1} & \Sigma_{23}^{-1} \\ \Sigma_{31}^{-1} & \Sigma_{32}^{-1} & \Sigma_{33}^{-1} \end{bmatrix} \begin{bmatrix} Y_{(1)} - \mu_{(1)} \\ Y_{(2)} - \mu_{(2)} \\ Y_{(3)} - \mu_{(3)} \end{bmatrix} \\
&= \begin{bmatrix} \boldsymbol{X}_{(1)}\mu'_{(1)} & \boldsymbol{X}_{(2)}\mu'_{(2)} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{X}_{(3)}\mu'_{(3)} \end{bmatrix} \begin{bmatrix} \Sigma_{11}^{-1} & \Sigma_{12}^{-1} & \Sigma_{13}^{-1} \\ \Sigma_{21}^{-1} & \Sigma_{22}^{-1} & \Sigma_{23}^{-1} \\ \Sigma_{31}^{-1} & \Sigma_{32}^{-1} & \Sigma_{33}^{-1} \end{bmatrix} \begin{bmatrix} Y_{(1)} - \mu_{(1)} \\ Y_{(2)} - \mu_{(2)} \\ Y_{(3)} - \mu_{(3)} \end{bmatrix} \\
&= \begin{bmatrix} \boldsymbol{X}_{(1)}\mu'_{(1)} & \boldsymbol{X}_{(2)}\mu'_{(2)} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{X}_{(3)}\mu'_{(3)} \end{bmatrix} \begin{bmatrix} \sum_{k=1}^{3} \Sigma_{1k}^{-1} \left(Y_{(k)} - \mu_{(k)}\right) \\ \sum_{k=1}^{3} \Sigma_{2k}^{-1} \left(Y_{(k)} - \mu_{(k)}\right) \\ \sum_{k=1}^{3} \Sigma_{3k}^{-1} \left(Y_{(k)} - \mu_{(k)}\right) \end{bmatrix} \\
&= \begin{bmatrix} \sum_{j=1}^{2} \boldsymbol{X}_{(j)}\mu'_{(j)} \sum_{k=1}^{3} \Sigma_{jk}^{-1} \left(Y_{(k)} - \mu_{(k)}\right) \\ \boldsymbol{X}_{(3)}\mu'_{(3)} \sum_{k=1}^{3} \Sigma_{3k}^{-1} \left(Y_{(k)} - \mu_{(k)}\right) \end{bmatrix}
\end{aligned}$$

The above can be extended to a more general case, as we see that for components that share coefficients, their contribution is the sum of the weighted least squared expression.

The expression $D(\boldsymbol{X};\boldsymbol{\beta})\Sigma_{\boldsymbol{Y}}^{-1}(\boldsymbol{X};\boldsymbol{\beta},F)D(\boldsymbol{X};\boldsymbol{\beta})^T$ becomes

$$\begin{aligned}
&= \begin{bmatrix} \boldsymbol{X}_{(1)}\mu'_{(1)} & \boldsymbol{X}_{(2)}\mu'_{(2)} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{X}_{(3)}\mu'_{(3)} \end{bmatrix} \begin{bmatrix} \Sigma_{11}^{-1} & \Sigma_{12}^{-1} & \Sigma_{13}^{-1} \\ \Sigma_{21}^{-1} & \Sigma_{22}^{-1} & \Sigma_{23}^{-1} \\ \Sigma_{31}^{-1} & \Sigma_{32}^{-1} & \Sigma_{33}^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{X}_{(1)}\mu'_{(1)} & \boldsymbol{X}_{(2)}\mu'_{(2)} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{X}_{(3)}\mu'_{(3)} \end{bmatrix}^T \\
&= \begin{bmatrix} \boldsymbol{X}_{(1)}\mu'_{(1)} & \boldsymbol{X}_{(2)}\mu'_{(2)} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{X}_{(3)}\mu'_{(3)} \end{bmatrix} \begin{bmatrix} \Sigma_{11}^{-1} & \Sigma_{12}^{-1} & \Sigma_{13}^{-1} \\ \Sigma_{21}^{-1} & \Sigma_{22}^{-1} & \Sigma_{23}^{-1} \\ \Sigma_{31}^{-1} & \Sigma_{32}^{-1} & \Sigma_{33}^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{X}_{(1)}^T\mu'_{(1)} & \mathbf{0} \\ \boldsymbol{X}_{(2)}^T\mu'_{(2)} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{X}_{(3)}^T\mu'_{(3)} \end{bmatrix} \\
&= \begin{bmatrix} \sum_{j=1}^{2} \left(\boldsymbol{X}_{(1)}\mu'_{(1)}\Sigma_{1j}^{-1} + \boldsymbol{X}_{(2)}\mu'_{(2)}\Sigma_{2j}^{-1}\right) \boldsymbol{X}_{(j)}^T\mu'_{(j)} & \left(\boldsymbol{X}_{(1)}\mu'_{(1)}\Sigma_{13}^{-1} + \boldsymbol{X}_{(2)}\mu'_{(2)}\Sigma_{23}^{-1}\right) \boldsymbol{X}_{(3)}^T\mu'_3 \\ \boldsymbol{X}_{(3)}\mu'_{(3)} \left(\boldsymbol{X}_{(1)}^T\mu'_{(1)}\Sigma_{31}^{-1} + \boldsymbol{X}_{(2)}^T\mu'_{(2)}\Sigma_{32}^{-1}\right) & \boldsymbol{X}_{(3)}\boldsymbol{X}_{(3)}^T(\mu'_3)^2\Sigma_{33}^{-1} \end{bmatrix}
\end{aligned}$$

Taking the expectation yields the expression for $\mathrm{Cov}\left(S_{\boldsymbol{\beta},F}(\boldsymbol{X},\boldsymbol{Y}), S_{\boldsymbol{\beta},F}(\boldsymbol{X},\boldsymbol{Y})\right)$

$$= \begin{bmatrix} \sum_{j=1}^{2} \mathbb{E}^{\mathcal{X}}\left[\left(\boldsymbol{X}_{(1)}\mu'_{(1)}\Sigma_{1j}^{-1} + \boldsymbol{X}_{(2)}\mu'_{(2)}\Sigma_{2j}^{-1}\right) \boldsymbol{X}_{(j)}^T\mu'_{(j)}\right] & \mathbb{E}^{\mathcal{X}}\left[\left(\boldsymbol{X}_{(1)}\mu'_{(1)}\Sigma_{13}^{-1} + \boldsymbol{X}_{(2)}\mu'_{(2)}\Sigma_{23}^{-1}\right) \boldsymbol{X}_{(3)}^T\mu'_3\right] \\ \mathbb{E}^{\mathcal{X}}\left[\boldsymbol{X}_{(3)}\mu'_{(3)} \left(\boldsymbol{X}_{(1)}^T\mu'_{(1)}\Sigma_{31}^{-1} + \boldsymbol{X}_{(2)}^T\mu'_{(2)}\Sigma_{32}^{-1}\right)\right] & \mathbb{E}^{\mathcal{X}}\left[\boldsymbol{X}_{(3)}\boldsymbol{X}_{(3)}^T(\mu'_3)^2\Sigma_{33}^{-1}\right] \end{bmatrix}$$

In the case of **independence**, similar to above we find that

$$S_{\boldsymbol{\beta},F}(\boldsymbol{X},\boldsymbol{Y}) = \begin{bmatrix} \boldsymbol{X}_{(1)}\mu'_{(1)}\Sigma_{11}^{-1} \left(Y_{(1)} - \mu_{(1)}\right) + \boldsymbol{X}_{(2)}\mu'_{(2)}\Sigma_{22}^{-1} \left(Y_{(2)} - \mu_{(2)}\right) \\ \boldsymbol{X}_{(3)}\mu'_{(3)}\Sigma_{33}^{-1} \left(Y_{(3)} - \mu_{(3)}\right) \end{bmatrix}$$

$$\mathrm{Cov}\left(S_{\boldsymbol{\beta},F}(\boldsymbol{X},\boldsymbol{Y}), S_{\boldsymbol{\beta},F}(\boldsymbol{X},\boldsymbol{Y})\right) = \begin{bmatrix} \mathbb{E}^{\mathcal{X}}\left[\boldsymbol{X}_{(1)}(\boldsymbol{X}_{(1)}^T(\mu'_{(1)})^2\Sigma_{11}^{-1}\right] + \mathbb{E}^{\mathcal{X}}\left[\boldsymbol{X}_{(2)}(\boldsymbol{X}_{(2)}^T(\mu'_{(2)})^2\Sigma_{22}^{-1}\right] & \mathbf{0} \\ \mathbf{0} & \mathbb{E}^{\mathcal{X}}\left[\boldsymbol{X}_{(3)}\boldsymbol{X}_{(3)}^T(\mu'_3)^2\Sigma_{33}^{-1}\right] \end{bmatrix}$$

Therefore, we see that this is a hybrid of the two cases seen prior.

## A.4 Simulation Plots

Below we will include the histograms for each estimated parameter from the simulations in Section 4.1

### A.4.1 Unconstrained Bivariate Normal Distribution, 4 coefficients

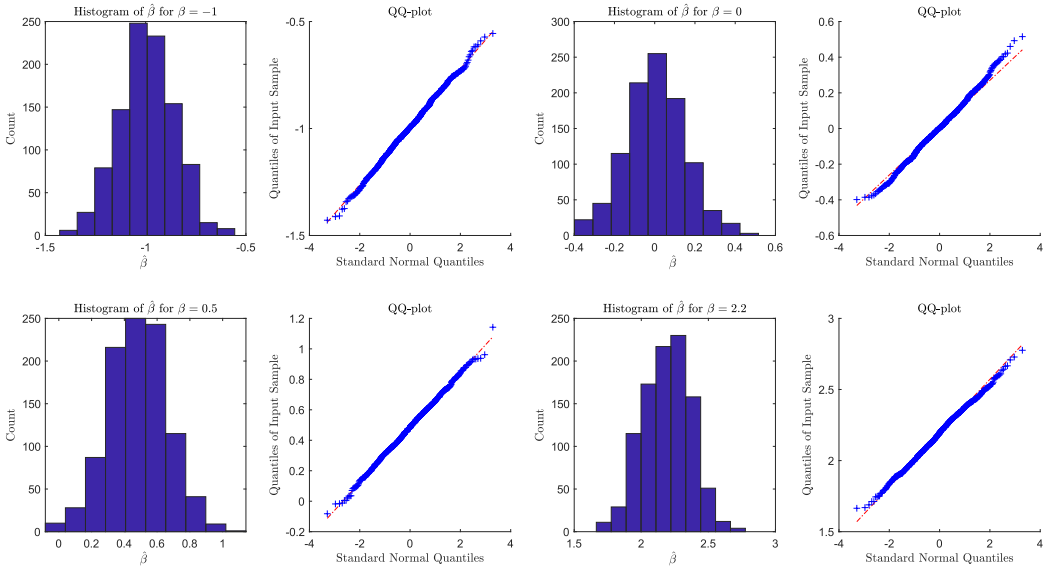


**Figure A.1:** Histogram of $\hat{\boldsymbol{\beta}}$ estimates in Table 4.1 for $n = 100$

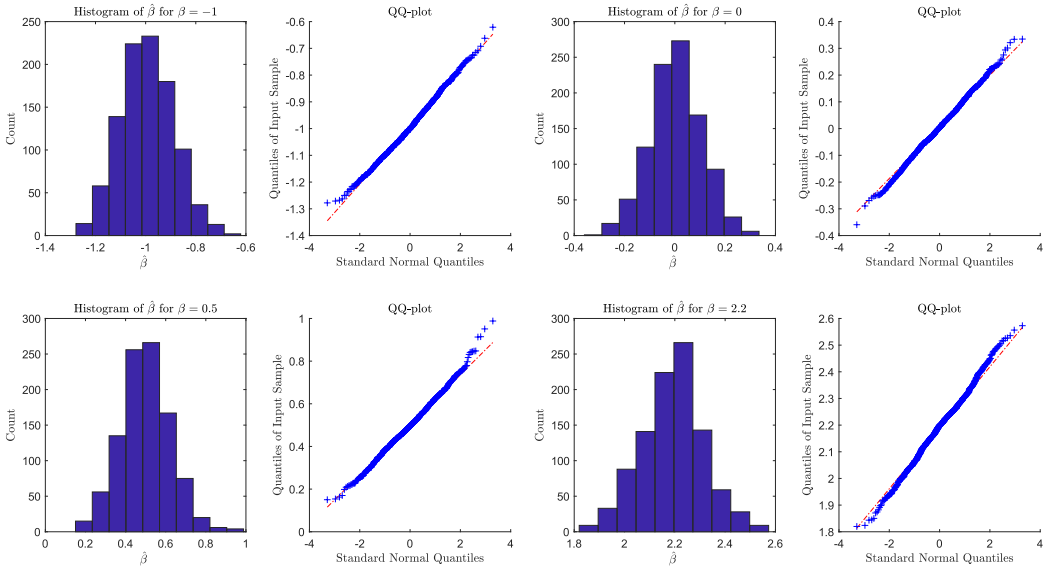


**Figure A.2:** Histogram of $\hat{\boldsymbol{\beta}}$ estimates in Table 4.1 for $n = 200$

### A.4.2 Unconstrained Bivariate Normal Distribution, 6 coefficients

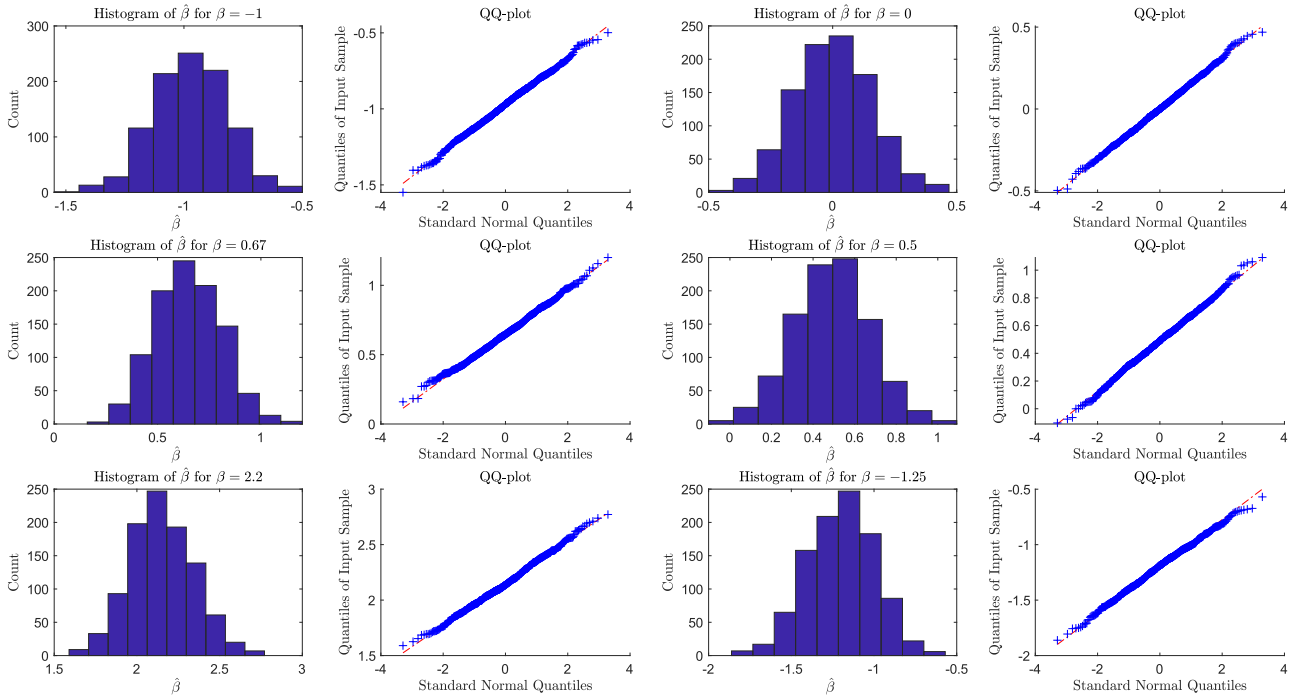


**Figure A.3:** Histogram of $\hat{\boldsymbol{\beta}}$ estimates in Table 4.2 for $n = 100$

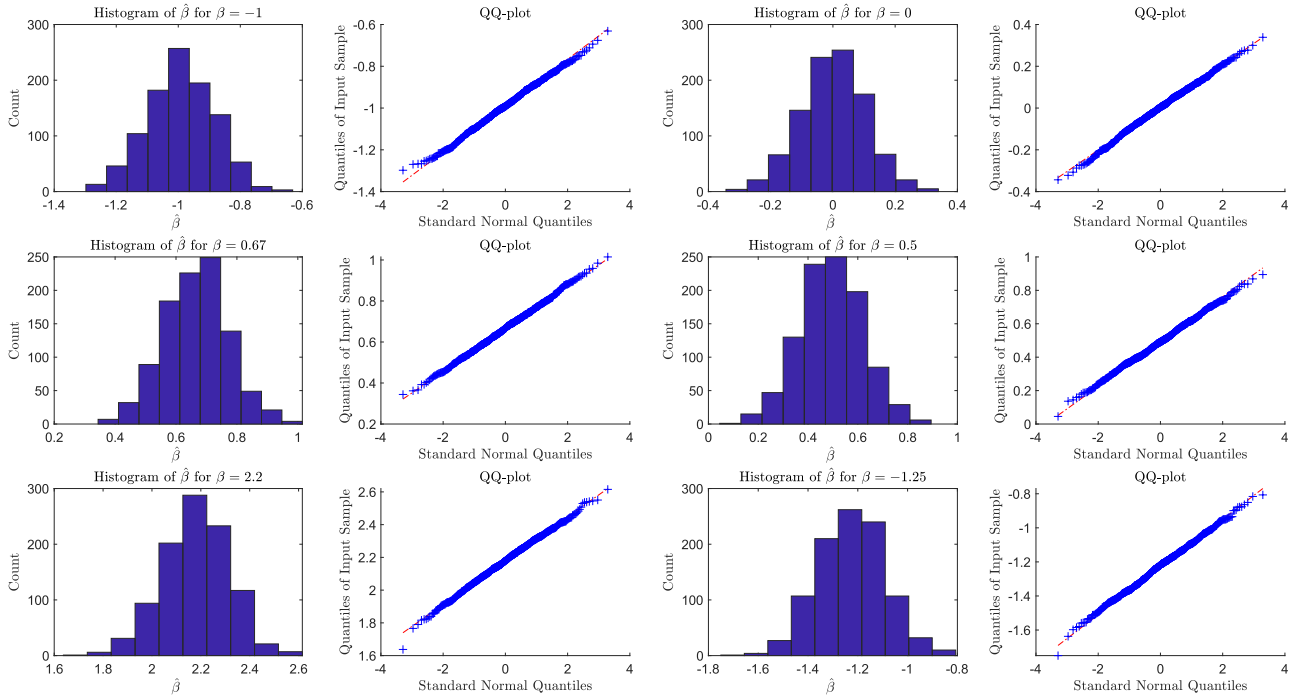


**Figure A.4:** Histogram of $\hat{\boldsymbol{\beta}}$ estimates in Table 4.2 for $n = 200$

### A.4.3 Constrained Bivariate Normal Distribution, 3 coefficients

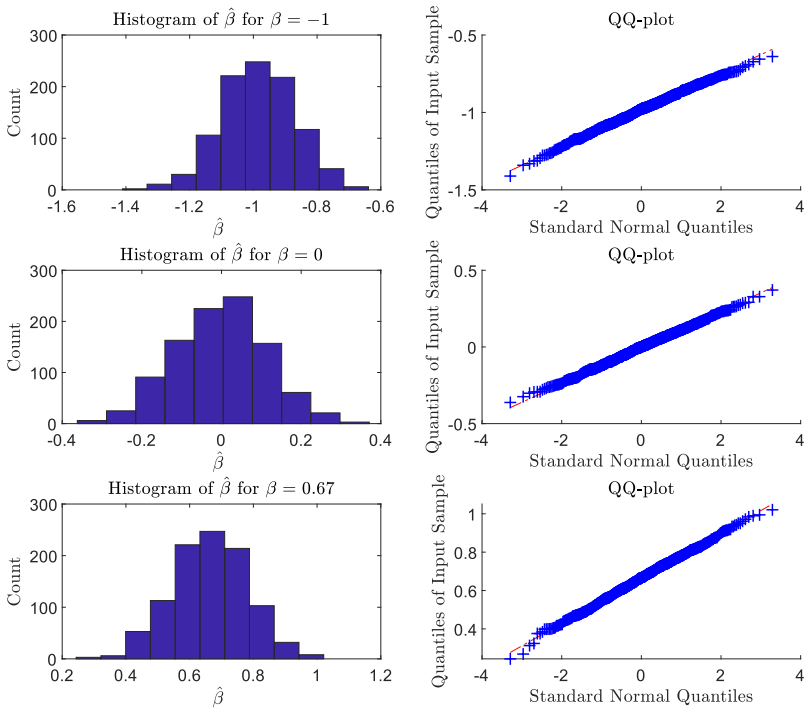


**Figure A.5:** Histogram of $\hat{\boldsymbol{\beta}}$ estimates in Table 4.3 for $n = 100$

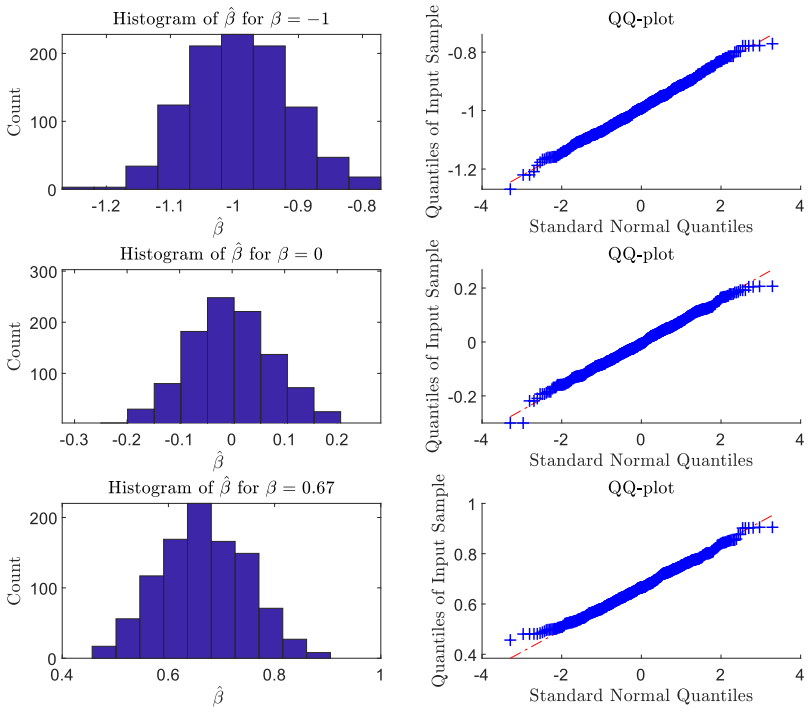


**Figure A.6:** Histogram of $\hat{\boldsymbol{\beta}}$ estimates in Table 4.3 for $n = 200$

## A.5 Butterfly Dataset Application

### A.5.1 Correlation Matrix of the dataset

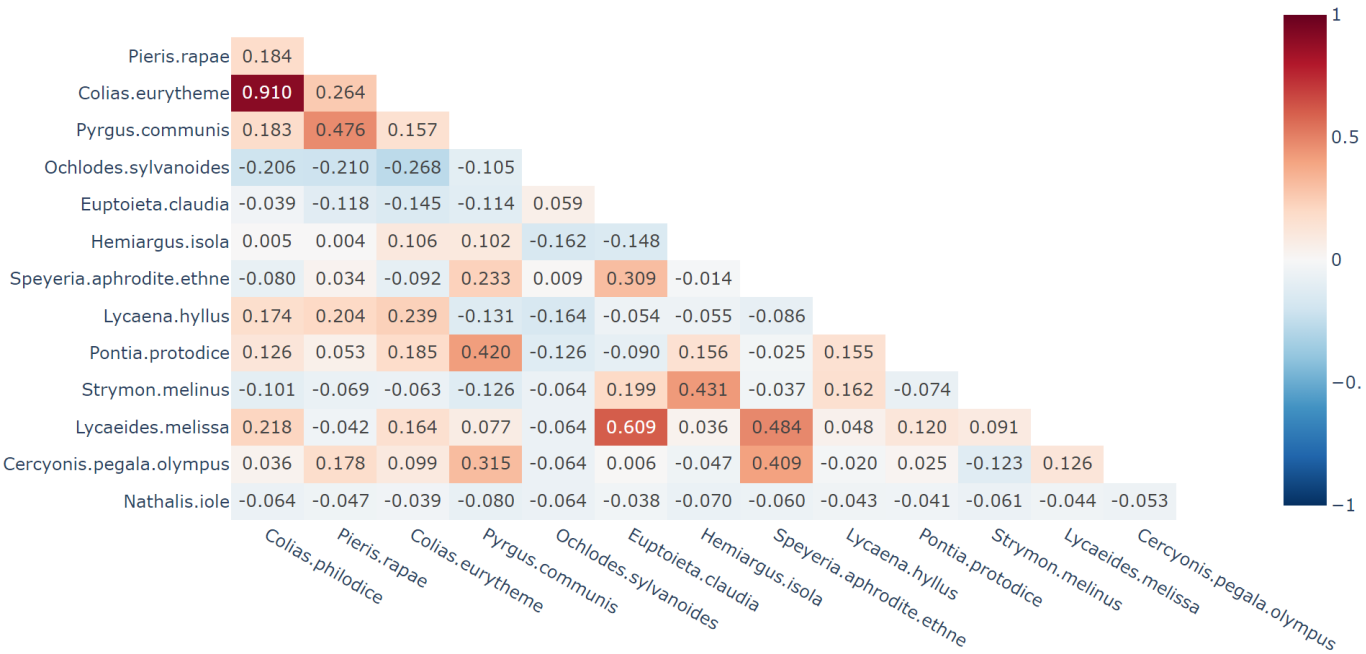


**Figure A.7:** Correlation matrix for the 14 most populous species in the butterfly, ordered by their total observed counts in the dataset.

### A.5.2 Butterfly Counts

| No. | Species | Count | No. | Species | Count |
|---|---|---|---|---|---|
| 1 | Colias philodice | 383 | 18 | Polites themistocles | 5 |
| 2 | Pieris rapae | 317 | 19 | Papilio polyxenes | 5 |
| 3 | Colias eurytheme | 256 | 20 | Limenitis archippus | 5 |
| 4 | Pyrgus communis | 59 | 21 | Vanessa cardui | 4 |
| 5 | Ochlodes sylvanoides | 54 | 22 | Danaus gilippus | 4 |
| 6 | Euptoieta claudia | 42 | 23 | Speyeria edwardsii | 3 |
| 7 | Hemiargus isola | 41 | 24 | Papilio rutulus | 2 |
| 8 | Speyeria aphrodite ethne | 39 | 25 | Junonia coenia | 2 |
| 9 | Lycaena hyllus | 38 | 26 | Pholisora catullus | 1 |
| 10 | Pontia protodice | 28 | 27 | Hesperia uncas uncas | 1 |
| 11 | Strymon melinus | 26 | 28 | Hesperia ottoe | 1 |
| 12 | Lycaeides melissa | 15 | 29 | Polites peckius | 1 |
| 13 | Cercyonis pegala olympus | 12 | 30 | Polites mystic | 1 |
| 14 | Nathalis iole | 11 | 31 | Papilio multicaudatus | 1 |
| 15 | Hesperia leonardus pawnee | 10 | 32 | Everes comyntas | 1 |
| 16 | Danaus plexippus | 10 | 33 | Plebejus acmon | 1 |
| 17 | Phyciodes campestris | 8 | | | |

**Table A.1:** Total counts for all 33 butterfly species across the 66 locations in Boulder County Colorado.

### A.5.3 Fitted Regression Coefficients of Butterfly Model.

| Species | Intercept | Building | Urban Vegetation | Mixed | Short | Tall |
|---|---|---|---|---|---|---|
| Colias philodice | 3.12 | $1.6 \times 10^{-3}$ | $-1.0 \times 10^{-2}$ | -2.10 | -3.91 | -1.51 |
| Pieris rapae | 2.02 | 0.11 | $2.1 \times 10^{-2}$ | -2.92 | -1.62 | -0.34 |
| Colias eurytheme | 2.47 | $3.1 \times 10^{-3}$ | $-1.0 \times 10^{-2}$ | -1.99 | -2.47 | -0.62 |
| Pyrgus communis | 0.14 | 0.08 | $2.8 \times 10^{-2}$ | -1.23 | -1.45 | -0.53 |
| Ochlodes sylvanoides | -2.61 | -0.13 | $4.8 \times 10^{-2}$ | 2.56 | 1.99 | 1.61 |
| Euptoieta claudia | -1.60 | -0.10 | $-1.6 \times 10^{-2}$ | 2.89 | 0.22 | -0.49 |
| Hemiargus isola | -0.03 | -0.04 | $-2.9 \times 10^{-2}$ | -1.18 | -0.74 | 0.27 |
| Speyeria aphrodite ethne | -1.25 | -0.05 | $-2.4 \times 10^{-3}$ | 1.83 | 0.42 | 0.51 |
| Lycaena hyllus | -0.53 | -0.19 | $1.9 \times 10^{-2}$ | -2.15 | $-3288.3$ | 1.47 |
| Pontia protodice | -0.55 | -0.08 | $4.9 \times 10^{-2}$ | -2.95 | -2.46 | 0.83 |
| Strymon melinus | -1.27 | -0.26 | $-2.5 \times 10^{-3}$ | 0.77 | -0.08 | 0.59 |
| Lycaeides melissa | -1.63 | -0.08 | $-8.4 \times 10^{-3}$ | 0.86 | -4.81 | 0.92 |
| Cercyonis pegala olympus | -2.65 | -0.07 | $5.2 \times 10^{-2}$ | 1.01 | 0.80 | 1.00 |
| Nathalis iole | $-2.3 \times 10^{5}$ | 0.16 | $3.5 \times 10^{-2}$ | $2.3 \times 10^{5}$ | $2.2 \times 10^{5}$ | $-1.7 \times 10^{5}$ |
| **VSPGLM Log-Likelihood ($\ell$)** | | -170.9158 | | | | |

**Table A.2:** Table of fitted regression coefficients of the 14-dimensional butterfly species model. Note that Mixed, Short and Tall refers to the covariates of the type of habitat.

| Species | Coefficient | Estimate | S.E | $T$ | $p$ |
|---|---|---|---|---|---|
| *Colias Philodice* | Intercept | 2.98 | 0.30 | 9.96 | $2.51 \times 10^{-14}$ |
| | Building | -0.06 | 0.05 | -1.26 | 0.2138 |
| | Urban.Vegetation | 0.01 | 0.01 | 0.64 | 0.5246 |
| | Habitat.Mixed | -1.93 | 0.42 | -4.56 | $2.60 \times 10^{-5}$ |
| | Habitat.Short | -4.04 | 0.57 | -7.09 | $1.77 \times 10^{-9}$ |
| | Habitat.Tall | -1.66 | 0.42 | -3.93 | $2.225 \times 10^{-4}$ |
| *Pieris Rapae* | Intercept | 1.83 | 0.38 | 4.84 | $9.49 \times 10^{-6}$ |
| | Building | 0.13 | 0.07 | 1.82 | 0.0734 |
| | Urban.Vegetation | 0.03 | 0.02 | 1.31 | 0.1966 |
| | Habitat.Mixed | -3.88 | 0.78 | -4.95 | $6.38 \times 10^{-6}$ |
| | Habitat.Short | -2.47 | 1.01 | -2.44 | 0.01770 |
| | Habitat.Tall | -0.26 | 0.47 | -0.56 | 0.5777 |
| *Colias Eurytheme* | Intercept | 2.35 | 0.30 | 7.72 | $1.467 \times 10^{-10}$ |
| | Building | 0.02 | 0.04 | 0.72 | 0.4756 |
| | Urban.Vegetation | -0.01 | 0.01 | -0.37 | 0.7115 |
| | Habitat.Mixed | -2.18 | 0.43 | -5.13 | $3.30 \times 10^{-6}$ |
| | Habitat.Short | -3.00 | 0.49 | -6.08 | $9.21 \times 10^{-6}$ |
| | Habitat.Tall | -0.86 | 0.33 | -2.61 | 0.0114 |
| **VSPGLM Log-Likelihood ($\ell$)** | | -228.871 | | | |

**Table A.3:** Coefficient summary for the 3 species Butterfly Model. Observe that all building and urban vegetation coefficients using individual Wald tests are found to be not significant, and interestingly all habitat coefficients are negative.

## A.6 Hunua 6 Plant Species Fitted Model

| Component | Coefficient | Estimate | S.E (Adjusted) | $T$ (Adjusted) | $p$-value (Adjusted) |
|---|---|---|---|---|---|
| Cyadea | Intercept | -0.7289 | 0.1787 (0.1792) | -4.08 (-4.07) | $5.45 \times 10^{-5}$ ($5.72 \times 10^{-5}$) |
| | Altitude ($\times 10^3$) | 0.0895 | 0.8787 (0.8797) | 0.10 ( 0.10) | 0.91893 (0.91903) |
| Beitaw | Intercept | -1.0281 | 0.1835 (0.2084) | -5.60 (-4.93) | $3.99 \times 10^{-8}$ ($1.19 \times 10^{-6}$) |
| | Altitude ($\times 10^3$) | 3.9708 | 0.9140 (1.0990) | 4.34 ( 3.61) | $1.78 \times 10^{-5}$ ($3.42 \times 10^{-4}$) |
| Kniexc | Intercept | -0.0452 | 0.1743 (0.1838) | -0.26 (-0.25) | 0.79525 (0.80559) |
| | Altitude ($\times 10^3$) | 2.4853 | 0.9155 (1.0460) | 2.71 ( 2.38) | 0.00693 (0.01798) |
| Kuneri | Intercept | 0.2709 | 0.1905 (0.2429) | 1.43 ( 1.12) | 0.15567 (0.26540) |
| | Altitude ($\times 10^3$) | -6.1320 | 1.1707 (1.8460) | -5.24 (-3.32) | $2.65 \times 10^{-7}$ ($9.77 \times 10^{-4}$) |
| Daccup | Intercept | -0.6334 | 0.1826 (0.1847) | -3.47 (-3.43) | 0.00058 (0.00067) |
| | Altitude ($\times 10^3$) | -1.1637 | 0.9416 (0.9995) | -1.24 (-1.16) | 0.21726 (0.24502) |
| Cyamed | Intercept | -1.3163 | 0.1937 (0.1994) | -6.80 (-6.60) | $4.02 \times 10^{-11}$ ($1.33 \times 10^{-10}$) |
| | Altitude ($\times 10^3$) | 1.7547 | 0.8999 (0.8839) | 1.95 ( 1.99) | 0.05190 (0.047819) |
| **Log-Likelihood** ($\ell$) | | -2317.2 | | | |

**Table A.4:** Fitted coefficients and inference for VSPGLM model fitted on 6 Hunua plant species.

## A.7 Kenyan School Children Dietary Intervention

### A.7.1 Observation time plot

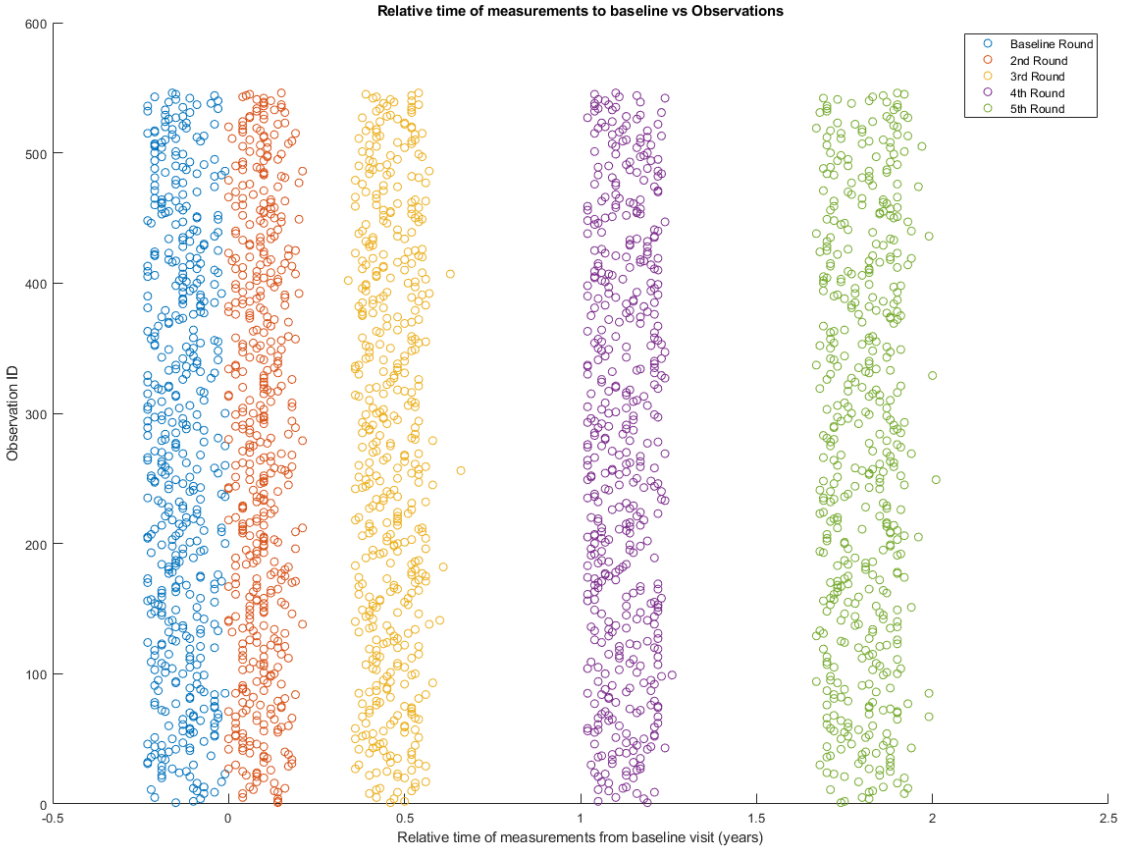


**Figure A.8:** Plot of times of observations against observation ID number

### A.7.2 VSPGLM Estimated Correlation Plot

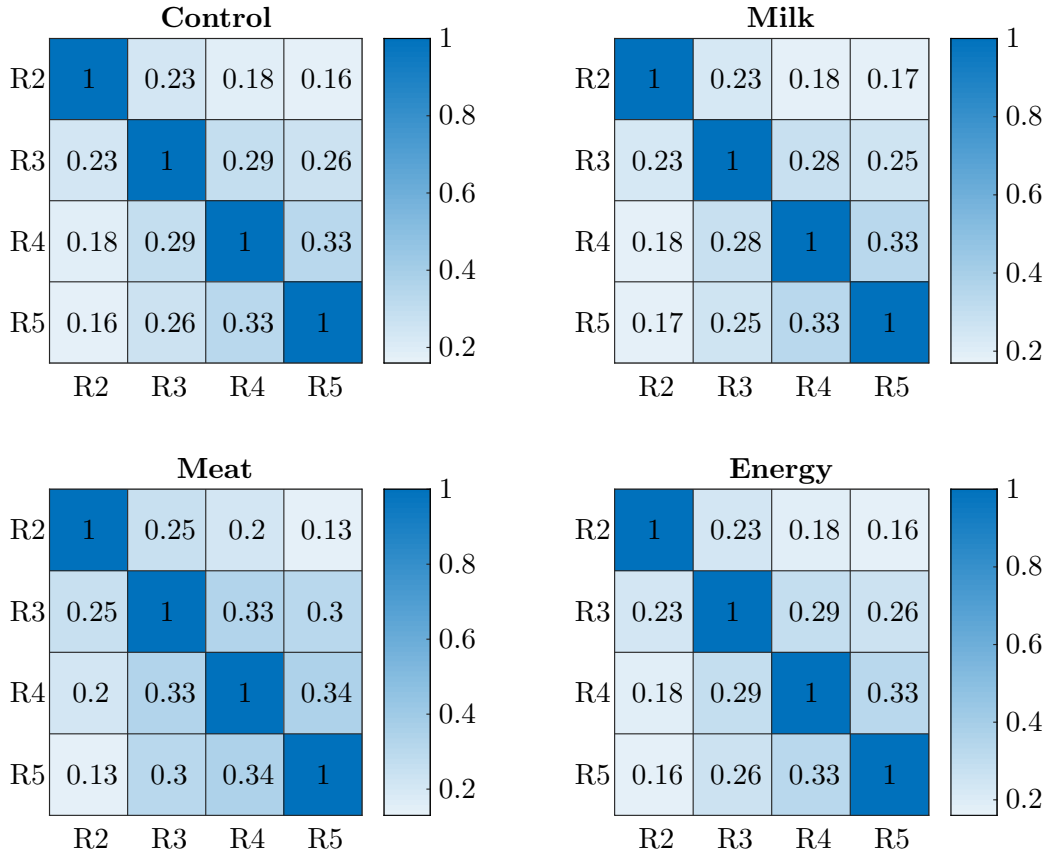


**Figure A.9:** Estimated Correlation of Raven Scores at Rounds 2-5 for an average male observation in each treatment group (averaging over the covariates) using an average time interaction. Scores that are further away in time are less correlated, but there are no auto-regressive or shared correlation structures.

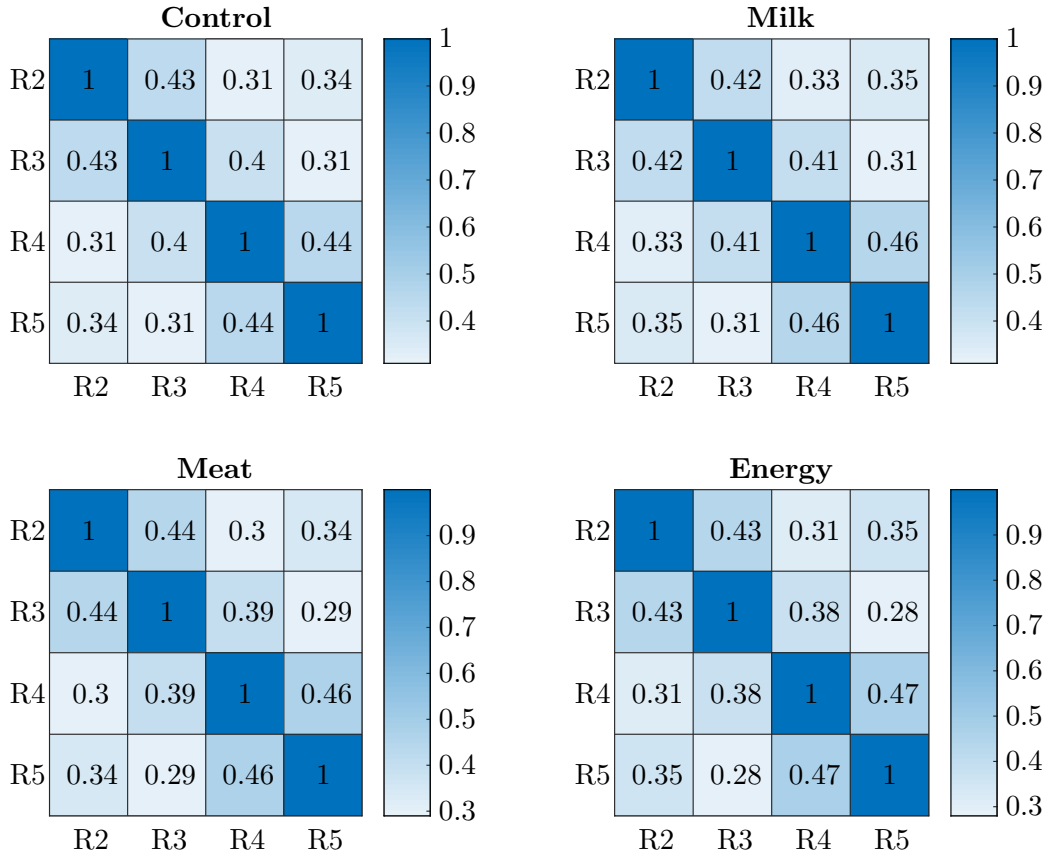


**Figure A.10:** Estimated Correlation of Arithmetic Scores at Rounds 2-5 for an average male observation in each treatment group (averaging over the covariates) using a relative time interaction. Similar to the Raven scores, time points further away are less correlated and there are no auto-regressive or shared correlation structures.