THE UNIVERSITY OF QUEENSLAND

AUSTRALIA

Semiparametric Vector Generalized Models: Computation and Estimation

Gabriel Dennis

BMath

A thesis submitted for the degree of BMath (Honours) at

The University of Queensland in 2021

School of Mathematics and Physics

Acknowledgements

I would like to thank my supervisor, Dr Alan Huang, who remained extremely patient and encouraging throughout this year, and was a great source of knowledge and direction.

Abstract

This thesis introduces a vector semiparametric generalized linear model (VSPGLM), which is a generalization of a semiparametric generalized linear model (SPGLM), proposed by Huang (2014), for data with multiple responses. This model originates from a re-writing of the data's reference distribution using an exponential tilt mechanism, which allows for simultaneous estimation of the models regression parameters and an arbitrary nonparametric reference distribution. The advantage of this approach is that it does not require any distributional assumptions about the structure of the data, as these are absorbed into the estimated reference distribution. Therefore, this VSPGLM can be applied to a wide variety of differently structured data. This greatly reduces the risk of model mispecification, a common thorn in the side of parametric modelling.

This thesis has the following structure, firstly, Chapter 1 contains some background on generalized linear models (GLMs) and vector generalized linear models (VGLMs), and also reviews several parametric and semiparametric modeling approaches. Chapter 2 then outlines the mathematical basis of the VSPGLM, its estimation procedure using maximum empirical likelihood estimation (MELE), and how inference can be conducted on the mean model parameters using both a Wald and empirical likelihood ratio test (ELRT). Chapter 3 provides an outline of the computational methods currently used for fitting the VSPGLM in MATLAB, and Chapter 4 shows several applications of the VSPGLM fit to a variety of differently structured datasets. Some of these datasets have previously published analysis using both parametric and semiparametric modelling, which allows for easy comparison of the VSPGLM's results. Chapter 5 then contains a variety of simulation studies to verify the asymptotic results for the models estimators presented in Chapter 2, and show that the power achieved by the VSPGLM is comparable to other competing methods. Finally, Chapter 6 concludes this thesis with a discussion of the proposed model, and mentions areas where future research could be conducted, and improvements and changes made to the code currently used to fit the model.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Generalized linear models (GLMs) (McCullagh and Nelder 1989) have become ubiquitous in modern applied statistics, with applications ranging across a variety of domains, such as agriculture, economics, bio-medicine, and the natural sciences. The wide range of uses for GLMs is related to the versatility of the modelling framework, due to their relationship with exponential families, home to some of the most well known probability distributions.

One constraint on the applicability of GLMs to an even wider variety of contexts is that they require a complete parametric specification of the "error" distribution from which the data was generated, with misspecification leading to inefficient parameter estimation and asymptotically incorrect inferences on the models regression parameters. The problem of model mispecification has lead to investigations and the proposals of many semiparametric methods of fitting GLMs. In general these methods admit a similar sets of estimating equations as classical GLMs, but also reduce the number of assumptions necessary to obtain an appropriate model by not requiring specification of an error distribution for the data. Examples of semiparametric methods which can be found commonly across GLM literature include quasi-likelihood (QL) methods (Wedderburn 1974), and generalized estimating equations (GEEs) (Liang and Zeger 1986), both of which obviate the constraints parametric methods place on the data by allowing it originate from an arbitrary error distribution, which, in the case of the semiparametric model proposed in Chapter 2 of this thesis, can be efficiently estimated alongside the models regression parameters using nonparametric empirical likelihood estimation (MELE).

Other disadvantages of the classical GLM framework arises in how it deals with multivariate response data, which can potentially have components containing data of differing types (*e.g.*, continuous, binary, count, ordinal). With mixed data of this type being particularly common across a number of the domains mentioned previously, particularly in health, bio-medicine and agriculture, where repeated or multiple measures per unit are common. Current parametric methods which have been proposed to deal with multivariate data of this type are, in general, relatively problem-specific (Song 2007; Yee 2015), and also suffer from the same issues of model misspecification present in classical GLMs.

In this chapter we introduce the classical GLM framework, as well as some ways in which adaptions have been made, both parametric and semiparametric, to combat the problem of model mispecification. We will then go on to discuss the generalization of this framework to multivariate data via vector generalized linear models (VGLMs), and how people have dealt with the difficulties and challenges that come when attempting to model multiple responses. As was the case with GLMs, there is the problem of model mispecification in the marginal distribution of each component, but now there is also the issue of modelling any association structure or mixed data types within the response vector.

Several motivating examples for the problems discussed in this chapter are the Burn Injury, Butterfly and Cochlear datasets which are fit and analysed in Chapter 4, using the model introduced in Chapter 2. In the case of the Burn Injury dataset, each response pair contains data of separate types, in this case, binary and continuous. This makes it a particularly interesting example, as previous attempts at vector regression have sometimes found it difficult to produce a valid models in this case. Similarly, the Butterfly dataset shows the model's performance when dealing with a large number of response components, which in this case are counts of 14 butterfly species across 66 locations. These species may also have both positive and negative correlations, (Figure A.1), which is another reason why this example is interesting, as generalizations of the Poisson distribution to deal with multiple responses typically do not allow negative correlations between components. The last motivating application that will be mentioned here, is the Cochlear dataset. This is a 12 dimensional dataset consisting of 12 repeated measures on 12 subjects after 6 different treatments. This study has a rather unusual crossover design, with measurements collected twice over two testing periods, with treatments randomized during the first testing period and reversed in order for the second. Data of this type are commonly modelled using repeated measures multiple analysis of variance (MANOVA), however, this model may be

a viable alternate approach. Each of these examples shows different aspects of the general nature, and advantages, of the proposed model, in contrast to existing VGLM methods. More analysis, results and discussion of these examples can be found in Sections 4.2, 4.3 and 4.7.

1.1 Generalised Linear Models (GLMs)

Definition 1.1 (Exponential Family)**.** *A family of probability distributions $F(y;\theta)$ for a random variable Y and indexed by a vector of parameters $\theta \in \Theta$, belongs to a single parameter exponential family if its density $dF(y;\theta)$ is of the form*

$$dF(y;\theta) = dF(y) \exp\left\{ c(\theta) \cdot a(y) + b(\theta) \right\}. \tag{1.1}$$

Here, $dF(y)$ is called a reference distribution for y, $a(y)$ the sufficient statistic, $c(\theta)$ the natural parameter and $b(\theta)$ the log-partition, or normalising, function, and is defined

$$b(\theta) = -\log\left\{ \int_{\mathcal{Y}} dF(y) \exp\left\{ c(\theta) \cdot a(y) \right\} \right\}. \tag{1.2}$$

Here, $b(\theta)$ is also known as the distribution's cumulant generating function (c.g.f), and in the case where $c(\theta) \equiv \theta$, the distribution is said to be in its "canonical form", or if $a(y) \equiv y$, a "linear exponential family". It should be noted that any exponential family can be made to be a linear, by modelling the sufficient statistic $a(y)$ directly. A distribution which is both linear and in canonical form is called a "natural" exponential family.

Suppose the data $\{(\boldsymbol{X}_i, Y_i) \in \mathcal{X} \times \mathcal{Y} \subset \mathbb{R}^q \times \mathbb{R} |\ i = 1, \ldots, n\}$ is observed, where each response Y_i is independently sampled conditional on the covariates $\boldsymbol{X}_i$, which could also be a random sample from some probability distribution or fixed by design. If the conditional distribution, or error distribution, of Y_i belongs to a single parameter exponential family (1.1) then it can be modeled using a generalized linear model (GLM). Here the density of Y_i is

$$dF(y;\theta,\phi) = \exp\left\{ \frac{a(y)c(\theta) + b(\theta)}{\phi} \right\} dF(y,\phi), \tag{1.3}$$

where the parameter ϕ models the dispersion present in the data, often one of the most difficult properties to accurately estimate. GLMs typically have the following two "systematic" components: firstly, the conditional distribution of $Y_i|\boldsymbol{X}_i$ depends on $\boldsymbol{X}_i$ only via a linear predictor $\boldsymbol{X}_i^T\boldsymbol{\beta}$, where $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_q)^T$ is a vector of parameters (*i.e.*,

regression coefficients), which are in most cases unknown and to be estimated from the data. Secondly, the conditional mean of each response is connected to the linear predictor via

$$\mathbb{E}[Y_i|\boldsymbol{X}_i] = \mu(\boldsymbol{X}_i^T\beta), \tag{1.4}$$

where the mean function $\mu(\cdot)$ is the inverse of the link function $g(\cdot)$, and is chosen by the modeller. For the remainder of this thesis, both $\mu(\boldsymbol{X}_i^T\boldsymbol{\beta})$ and μ_i will be used interchangeably, dependent on the context, with suppression of the subscript i when appropriate.

The most common example of a GLM is a normal linear model, which assumes the error distribution of Y_i follows a Gaussian, or normal, distribution which has a constant variance σ^2, and for which

$$\mu(\boldsymbol{X}_i^T\boldsymbol{\beta}) = \boldsymbol{X}_i^T\boldsymbol{\beta}\ , \tag{1.5}$$

i.e., the linear predictor $\boldsymbol{X}_i^T\boldsymbol{\beta}$ is linked directly to the conditional mean of the error distribution. Another example of a GLM is Poisson regression model, typically applied to count data, and assumes a Poisson error distribution. This results in the relationship

$$\mu(\boldsymbol{X}_i^T\boldsymbol{\beta}) = \mathbb{E}[Y_i|\boldsymbol{X}_i] = \mathrm{Var}[Y_i|\boldsymbol{X}_i], \tag{1.6}$$

where the log-link function,

$$\ln(\mu_i) = \boldsymbol{X}_i^T\boldsymbol{\beta} \iff \mu(\boldsymbol{X}_i^T\boldsymbol{\beta}) = \exp(\boldsymbol{X}_i^T\boldsymbol{\beta}), \tag{1.7}$$

is most often used to ensure non-negativity of the conditional mean.

1.1.1 Common properties of GLMs

One important property of any link function $g(\cdot)$, which ensures its invertibility, is that it is strictly increasing on some open interval given by $(m, M) \subset \mathbb{R}$, where $m = \inf\{\mathcal{Y}\}$ and $M = \sup\{\mathcal{Y}\}$ are the extremal values of the response's possible support. As the link function is chosen by the modeller there is sometimes some freedom for which link function is most appropriate for the data. The main restriction being that a link function must produce a valid model by constraining the conditional mean to take values in (m, M). This was seen in the Poisson regression example, where the log-link was used to ensure $\mu > 0$. For cases where the error distribution is both linear, $a(y) \equiv y$, and in its canonical form $c(\theta) \equiv \theta$, often the most appropriate link function to choose is the "canonical" link function, which links the conditional mean directly to the canonical parameter $g(\mu_i) = \theta = \boldsymbol{X}_i^T\boldsymbol{\beta}$.

Canonical links for common exponential families include the identity-link $\mu_i = \boldsymbol{X}_i^T\boldsymbol{\beta}$ (normal), log-link $\log(\mu_i) = \boldsymbol{X}_i^T\boldsymbol{\beta}$ (Poisson), inverse-link $\mu_i^{-1} = \boldsymbol{X}_i^T\boldsymbol{\beta}$ (gamma), and logit-link $\log(\mu_i/(1-\mu_i)) = \boldsymbol{X}_i^T\boldsymbol{\beta}$ (Bernoulli).

Two useful properties of natural exponential families, which are often used in GLMs, is that if Y originates from a natural exponential family with density

$$dF(y;\theta) = \exp\left\{\frac{y\theta + b(\theta)}{\phi}\right\} dF(y), \tag{1.8}$$

then

$$\mathbb{E}[Y] = -b'(\theta) \tag{1.9}$$

$$\mathrm{Var}[Y] = -b''(\theta)\phi\ . \tag{1.10}$$

To show (1.9), the log-likelihood for θ given y is

$$\begin{aligned}\ell(\theta;y) &= \ln dF(y;\theta) \\ &= \frac{y\theta + b(\theta)}{\phi} + \ln dF(y)\ .\end{aligned} \tag{1.11}$$

Which has the score

$$\begin{aligned}S(\theta;y) &= \frac{\partial \ell}{\partial \theta} \\ &= \frac{y + b'(\theta)}{\phi}\ .\end{aligned} \tag{1.12}$$

A property of the efficient score $S(\theta)$, is that $\mathbb{E}[S(\theta)] = 0$ for natural exponential families when taking the expectation with respect to the true value of θ (Section 2.6). Therefore,

$$\begin{aligned}0 = \mathbb{E}[S(\theta)] &= \mathbb{E}[Y] + \mathbb{E}[b'(\theta)] \text{ as } \phi \neq 0 \\ \implies \mathbb{E}[Y] &= -b'(\theta),\end{aligned} \tag{1.13}$$

which is (1.9). One can also show that $\mathbb{E}\Big[\partial_\theta S(\theta)\Big] = -\mathbb{E}\Big[S(\theta)^2\Big]$ (Section 2.6). Therefore, using (1.9) the variance for Y is

$$\begin{aligned}0 &= \mathbb{E}\Big[\partial_\theta S(\theta)\Big] + \mathbb{E}\Big[S(\theta)^2\Big] \\ &= \frac{1}{\phi}\mathbb{E}[b''(\theta)] + \frac{1}{\phi^2}\mathbb{E}\Big[(y + b'(\theta))^2\Big] \\ &= \phi b''(\theta) + \mathrm{Var}[Y] \\ \implies \mathrm{Var}[Y] &= -b''(\theta)\phi.\end{aligned} \tag{1.14}$$

These properties then imply that, when modelling with GLMs, the conditional variance $\mathrm{Var}[Y_i|\boldsymbol{X}_i]$ can often be expressed as a function of the conditional mean, that is $\mathrm{Var}[Y_i|\boldsymbol{X}_i] =$

$\mathbf{V}(\mu_i)$ for some non-negative variance function $V(\cdot)$. This is referred to as the GLM's mean-variance relationship, a linear example of which was shown previously in (1.6) for Poisson regression. This relationship is often one of the most challenging aspects to correctly specify in a GLM, partially due to the difficulties in estimating the dispersion parameter ϕ which is present in (1.10).

1.1.2 Fitting GLMs and parameter inference

To fit classical GLMs the score $\boldsymbol{S}(\boldsymbol{\beta})$ is used to obtain the maximum likelihood estimates (MLEs) for $\boldsymbol{\beta}$. For a single observation, $\boldsymbol{S}(\boldsymbol{\beta})$ has the form

$$\boldsymbol{S}(\boldsymbol{\beta}) = \frac{Y - \mu(\boldsymbol{X}^T\boldsymbol{\beta})}{\text{Var}[Y|\boldsymbol{X}]}\mu'(\boldsymbol{X}^T\boldsymbol{\beta})\boldsymbol{X} \ . \tag{1.15}$$

To show this, the previous result is used that $\mu(\boldsymbol{X}^T\boldsymbol{\beta}) = \mathbb{E}[Y|\boldsymbol{X}] = -b'(\theta)$. Hence,

$$\frac{\partial\theta}{\partial\mu} = \frac{1}{-b''(\theta)} \ . \tag{1.16}$$

$\boldsymbol{S}(\boldsymbol{\beta})$ can then be derived using the chain rule.

$$\begin{aligned}
\boldsymbol{S}(\boldsymbol{\beta}) = \nabla_{\boldsymbol{\beta}}\ell = \ & \frac{\partial\ell}{\partial\theta}\frac{\partial\theta}{\partial\mu}\frac{\partial\mu}{\partial\boldsymbol{\beta}} \\
&= \Big(\frac{Y + b'(\theta)}{\phi}\Big)\Big(\frac{1}{-b''(\theta)}\Big)\Big(\mu'(\boldsymbol{X}^T\boldsymbol{\beta})\boldsymbol{X}\Big) \\
&= \Big(\frac{Y + b'(\theta)}{-b''(\theta)\phi}\Big)\mu'(\boldsymbol{X}^T\boldsymbol{\beta})\boldsymbol{X} \\
&= \Big(\frac{Y - \mu(\boldsymbol{X}^T\boldsymbol{\beta})}{\text{Var}[Y|\boldsymbol{X}]}\Big)\mu'(\boldsymbol{X}^T\boldsymbol{\beta})\boldsymbol{X}
\end{aligned} \tag{1.17}$$

For n independent observations $Y_1, \ldots, Y_n$ the estimate $\hat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ are the solution to the, typically non-linear, set of equations

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

which are usually solved using numerical methods. A standard example of the score for a parametric GLM is for logistic regression on Bernoulli data. This uses a logit-link function, and has the following mean-variance relationship

$$\mathbb{E}[Y_i|\boldsymbol{X}_i] = \frac{\exp(\boldsymbol{X}_i^T\boldsymbol{\beta})}{1 + \exp(\boldsymbol{X}_i^T\boldsymbol{\beta})} = \mu(\boldsymbol{X}_i^T\boldsymbol{\beta}), \quad \text{Var}[Y_i|\boldsymbol{X}_i] = \mu_i(1 - \mu_i) \ . \tag{1.18}$$

The resulting score equations are then

$$\begin{aligned} 0 &= \sum_{i=1}^{n} \Big(\frac{Y_i - \mu(\boldsymbol{X}_i^T \boldsymbol{\beta})}{\mathrm{Var}[Y_i|\boldsymbol{X}_i]} \Big) \mu'(\boldsymbol{X}_i^T \boldsymbol{\beta}) \boldsymbol{X}_i \\ &= \sum_{i=1}^{n} \Big(\frac{Y_i - \mu(\boldsymbol{X}_i^T \boldsymbol{\beta})}{\mu(\boldsymbol{X}_i^T \boldsymbol{\beta})(1 - \mu(\boldsymbol{X}_i^T \boldsymbol{\beta}))} \Big) \mu(\boldsymbol{X}_i^T \boldsymbol{\beta})(1 - \mu(\boldsymbol{X}_i^T \boldsymbol{\beta})) \boldsymbol{X}_i \\ &= \sum_{i=1}^{n} \Big(Y_i - \mu(\boldsymbol{X}_i^T \boldsymbol{\beta}) \Big) \boldsymbol{X}_i \ . \end{aligned} \tag{1.19}$$

The estimate $\hat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ that the score equations admits has the following asymptotic properties, due to the good behaviour of exponential families (Appendix A.1).

Theorem 1.1 (Asymptotic Normality of $\hat{\boldsymbol{\beta}}$)**.** *As the sample size $n \to \infty$*

$$\sqrt{n}\ (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \to \mathcal{N}(\mathbf{0}, \Sigma_{\boldsymbol{\beta}}) \tag{1.20}$$

in distribution, where the asymptotic covariance matrix $\Sigma_{\boldsymbol{\beta}}$ is given by

$$\Sigma_{\boldsymbol{\beta}} = \Big[\mathbb{E} \Big(\frac{\mu'(\boldsymbol{X}^T \boldsymbol{\beta})^2 \boldsymbol{X} \boldsymbol{X}^T}{\mathrm{Var}[Y|\boldsymbol{X}]} \Big) \Big]^{-1} \tag{1.21}$$

The asymptotic covariance matrix is simply the inverse Fisher information matrix, which for exponential families is

$$\begin{aligned} \mathcal{I}(\boldsymbol{\beta}) &= \mathbb{E}[\boldsymbol{S}(\boldsymbol{\beta}) \boldsymbol{S}(\boldsymbol{\beta})^T] \\ &= \frac{1}{\mathrm{Var}[Y|\boldsymbol{X}]^2} \mathbb{E} \Big[(Y - \mu(\boldsymbol{X}^T \boldsymbol{\beta}))^2 \mu'(\boldsymbol{X}^T \boldsymbol{\beta})^2 \boldsymbol{X} \boldsymbol{X}^T \Big] \\ &= \frac{1}{\mathrm{Var}[Y|\boldsymbol{X}]^2} \mathbb{E} \Big[(Y - \mu(\boldsymbol{X}^T \boldsymbol{\beta}))^2 \Big] \mu'(\boldsymbol{X}^T \boldsymbol{\beta})^2 \boldsymbol{X} \boldsymbol{X}^T \\ &= \frac{\mu'(\boldsymbol{X}^T \boldsymbol{\beta})^2 \boldsymbol{X} \boldsymbol{X}^T}{\mathrm{Var}[Y|\boldsymbol{X}]}. \end{aligned} \tag{1.22}$$

For n independent observations of Y_i, a natural estimator of the asymptotic covariance matrix then is

$$\hat{\Sigma}_{\boldsymbol{\beta}} = \left[\frac{1}{n} \sum_{i=1}^{n} \frac{\mu'(\boldsymbol{X}_i^T \hat{\boldsymbol{\beta}})^2 \boldsymbol{X}_i \boldsymbol{X}_i^T}{\widehat{\mathrm{Var}}[Y_i|\boldsymbol{X}_i]} \right]^{-1}. \tag{1.23}$$

The results of Theorem 1.1 allow parameter inference to be conducted on single regression parameter's estimate $\hat{\beta}_i$, using a Wald test. As, under the null hypothesis $H_0 : \beta_i = 0$,

$$\frac{\hat{\beta}_i}{\sqrt{\hat{\Sigma}_{\boldsymbol{\beta}(ii)}}} \sim t_{n-q}, \tag{1.24}$$

where n is the sample size and q the number of regression parameters in the model. Inference on nested models can also be conducted using the generalized likelihood ratio statistic, or deviance of the models, which asymptotically follows a χ^2_r distribution where r is the difference in the number of parameters between the two models.

1.1.3 Semiparametric methods

A downside of modelling within the classical GLM framework is that the correct parametric specification of a model can be difficult to obtain when modelling responses for which there is a lack of theory, or observation, of the underlying mechanism from which the data was generated. This is often a problem in parametric GLMs, as there are some necessary "researcher degrees of freedom" relating to the correct specification of the error distribution, as well as the the link function and mean-variance relationship. One way to avoid specifying a parametric error distribution for the data is to instead use a semiparametric approach to fitting GLMs. As with any change in approach, semiparametric methods have advantages and disadvantages over classical GLMs. Their advantage is that they are generally robust, and can usually be applied directly on a much broader range of data while requiring less initial effort in the modelling stage. The drawbacks come in a loss of efficiency that their parameter estimates suffer relative to those produced by correctly specified parametric models, with this difference being especially noticeable when dealing with small to moderate sample sizes. Early attempts at a semiparametric approach to fitting GLMs include the use of quasi-likelihood (QL) methods (Wedderburn 1974), and slightly later, generalized estimating equations (GEEs) (Liang and Zeger 1986), which are described in Section 1.2.1.

In QL methods parameter estimation is done through the formation of the nonparametric "quasi" (log-)likelihood

$$Q(\mu, Y) = \int_y^\mu \frac{y - t}{\phi V(t)}\, dt, \tag{1.25}$$

where $\phi > 0$ is assumed to be a constant and $V(\mu) > 0$ is a given variance function, which can be dependent on the conditional mean. For n independent observations $Y_1, \ldots, Y_n$, the full quasi-likelihood is then formed via the summation $Q(\boldsymbol{\mu}, \boldsymbol{Y}) = \sum_i Q_i(\mu_i, Y_i)$. This is done because the function

$$U(\mu, Y) = \frac{Y - \mu}{\phi V(\mu)}\ , \tag{1.26}$$

has similar necessary properties for asymptotically correct parameter estimation as the

parametric score function. Namely that,

$$\begin{aligned}\mathbb{E}[U(\mu, Y)] &= 0,\\ \text{Var}[U(\mu, Y)] &= \frac{1}{\sigma^2 V(\mu)},\\ -\mathbb{E}\Big[\frac{\partial U(\mu, Y)}{\partial \mu}\Big] &= \text{Var}[U(\mu, Y)].\end{aligned}$$

Due to the general approach, and lack of assumptions of QL, GLMs can be considered a to be a subset of the QL models, as the both admit the same estimating equations. As was the case with classical GLMs, correct specification of the variance function can sometimes pose a challenge during modelling, however methods do exists which allow for the variance functions to be initially left unspecified, after which they are estimated nonparametrically using the process of adaptive estimation (Dewanji and Zhao 2002). A slight downside to semiparametric methods such as QL, which only consider the mean parameters and ignore the data's underlying reference distribution, comes when attempting to conduct inference about the distribution of the data. As these methods only contain information on its first two moments.

1.2 Vector Generalized Linear Models (VGLMs)

According to Yee (2015, pp. 13–14), a vector generalized linear model (VGLM) is a relatively natural extension of a standard GLM to deal with data of the form $\{(\boldsymbol{X}_i, \boldsymbol{Y}_i) \in \mathcal{X} \times \mathcal{Y} \subset \mathbb{R}^q \times \mathbb{R}^K | i = 1, \ldots, n\}$. Now the conditional distribution for $\boldsymbol{Y}_i$ is given by

$$f(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\beta}) = f(\boldsymbol{y}, \boldsymbol{\mu}), \tag{1.27}$$

for some known joint density $f(\cdot)$. Note here that $f(\cdot)$ does not necessarily originate from a multidimensional exponential family, as multivariate generalizations of common exponential families (e.g., Poisson, gamma, ...) are difficult to construct, and most do not have a standard agreed upon form. In this definition of VGLMs, $\boldsymbol{\beta} = (\boldsymbol{\beta}_{(1)}^T, \boldsymbol{\beta}_{(2)}^T, \ldots, \boldsymbol{\beta}_{(K)}^T)$ contains the mean model parameters for each of the K response components, and similarly to the systematic component of GLMs, the means $\boldsymbol{\mu} = (\mu_{(1)}(\boldsymbol{X}^T\boldsymbol{\beta}_{(1)}), \ldots, \mu_{(K)}(\boldsymbol{X}^T\boldsymbol{\beta}_{(K)}))^T$ for each response component are linked to the linear predictors $\boldsymbol{X}^T\boldsymbol{\beta}_{(k)}$ by separate link functions $g_{(1)}(\cdot), g_{(2)}(\cdot), \ldots, g_{(K)}(\cdot)$.

An application of VGLMs is that they can also now be applied to model multiple parameters in a distribution instead of just the conditional mean μ, which is the case with GLMs. It

is also possible in VGLMs for mean models of different components to share regression coefficients. This induces certain constraints across multiple mean models, with the appropriateness of the choice of constraints based in symmetric or interchangeability arguments about the nature of the data.

There is no general approach to the construction of parametric VGLMs, partially due to the non-specific nature of the above definition. In some instances well-defined theory does exists, especially if the data is assumed to originate from a multivariate normal distribution (*e.g.*, linear mixed models, Gałecki and Burzykowski 2013), or for latent variable models. Otherwise there is a limited number of standard parametric VGLMs. The main difficulties of parametric VGLMs is that now we not only have to correctly specify the mean-variance relationship of each marginal distribution, but we must also propose a within-vector association structure for the data. One popular way to introduce the necessary within vector association is through the use of copula methods (Song 2007, pp. 125–129), however such methods also suffer from model mispecification within the copula used and can struggle at producing a valid model for data which has discrete marginal distributions. Successful semiparametric approaches to modeling multivariate data have often used GEEs, which have been shown to produce a robust model when dealing with both longitudinal data and data with mixed marginal distributions (Huang 2017).

1.2.1 Examples of VGLMS

There are some examples of somewhat standard parametric VGLMS which have been proposed to deal with specific types of data. The first example that we will give is of a log-linear VGLM, which is used in modelling correlated discrete responses (Prentice and Zhao 1991). A special case of this model is the quadratic exponential model (QEM) for correlated binary data, which assumes associations between any two response components can be modelled with a single parameter. To fit this model one uses logit-link function on all marginal mean models

$$\mu_{(k)} = \mathbb{P}(Y_{(k)} = 1) = \frac{\exp(\boldsymbol{X}_{(k)}^T \boldsymbol{\beta}_{(k)})}{1 + \exp(\boldsymbol{X}_{(k)}^T \boldsymbol{\beta}_{(k)})}, \quad k = 1, \ldots, K \ . \tag{1.28}$$

Where the QEM probability mass function is of the form

$$\mathbb{P}(\boldsymbol{Y} = \boldsymbol{y}) = c(\boldsymbol{\theta}) \exp\left\{ \sum_{j=1}^{n} \theta_j y_{(j)} + \sum_{j<k} \theta_{jk} y_{(j)} y_{(k)} \right\}, \tag{1.29}$$

which has two vectors of canonical parameters $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$

$$\boldsymbol{\theta}_1 = (\theta_1, \theta_2, \theta_3, \ldots, \theta_m)^T \text{ and } \boldsymbol{\theta}_2 = (\theta_{12}, \theta_{13}, \theta_{14}, \ldots, \theta_{n-1,n})^T \tag{1.30}$$

which describe the association structure between the components of the response $\boldsymbol{Y}$. In particular, it can easily be shown for models of the form given in (1.29), that

$$\begin{aligned} \log\left\{\frac{\mathbb{P}(Y_{(j)}=1|Y_{(k)}=0, k\neq j)}{\mathbb{P}(Y_{(j)}=0|Y_{(k)}=0, k\neq j)}\right\} &= \log\left\{\frac{c(\boldsymbol{\theta})\exp\left\{\theta_j\right\}}{c(\boldsymbol{\theta})}\right\} \\ &= \theta_j \ . \end{aligned} \tag{1.31}$$

The vector of canonical parameters $\boldsymbol{\theta}_1$ encodes the log-odds that each response will be equal to 1 given all the other response components are equal to 0. Similarly, $\boldsymbol{\theta}_2$ contains the parameters which describes the log-odds ratio between any two components given all other components have been withheld.

$$\log\left\{\frac{\mathbb{P}(Y_{(j)}=1, Y_{(k)}=1|Y_{(l)}=y_l, l\neq k, j)\mathbb{P}(Y_{(j)}=0, Y_{(k)}=0|Y_{(l)}=y_l, l\neq k, j)}{\mathbb{P}(Y_{(j)}=1, Y_k=0|Y_{(l)}=y_l, l\neq k, j)\mathbb{P}(Y_{(j)}=0, Y_{(k)}=1|Y_{(l)}=y_l, l\neq k, j)}\right\} = \theta_{jk} \tag{1.32}$$

Other examples of a model that could potentially be classed as a VGLM, is the multivariate Conway-Maxwell-Poisson (MCMP) model (Sellers et al. 2021). The standard univariate CMP distribution has two parameters ν and μ, each relating to the dispersion and rate of the data, and has the probability mass function

$$\mathbb{P}(Y=y|\lambda,\nu) = \frac{\lambda^y}{(y!)^\nu Z(\lambda,\nu)}, \quad y=0,1,\ldots \ . \tag{1.33}$$

Where $Z(\lambda,\nu) = \sum_{k=0}^{\infty}\frac{\lambda^k}{(k!)^\nu}$ is a normalising constant, which is equal to $e^{-\lambda}$ in the case where the dispersion parameter ν is equal to 1. As in this case, (1.33) simplifies to the probability mass function of a Poisson distribution. To generalize this model for multiple count response components, one first forms a multivariate binomial distribution (see Krishnamoorthy (1951)), where we now assume the total number of trials n is distributed via CMP(λ,ν). As noted by Sellers et al. (2021), one limitation of the MCMP model is that it assumes that there is a non-negative correlation between any two components of the response vector. This would be the case of the data was coming from a construction of the multivariate Poisson distribution via the summation of Poisson random variables (Mahamunulu 1967), however this assumption is often violated by physical systems. One example of where negative correlations between components can often occur is in count

data which is collected for multiple species which are in competition for a scarce number of resources, *i.e.*, the increase in counts of one species almost necessarily leads to a decrease in counts of another, due to the depletion of resources caused by the increase in population. Other potential limitations of the MCMP model is that only a single dispersion parameter ν is fit, under the assumption that all marginal components have relatively equal dispersion, which also may be an unrealistic assumption.

A standard example of a semiparametric VGLM are the models fit through the use of GEEs. These models require a "working" correlation matrix $\boldsymbol{V}$ for the responses, which, similar to the mean-variance relationship, is structured by the modeller. An advantage of this approach is that $\boldsymbol{V}$ need not be correctly specified for consistent estimation, however, incorrect specification of $\boldsymbol{V}$ leads to a reduction in the efficiency of the subsequent parameter estimation. In the context of repeated measures, for a given mean model $\mu_{i(j)}$, which has the parameters $\boldsymbol{\beta}$, and models the conditional expectation of subject $i = 1, 2, 3, \ldots, n$ at its jth measurement, the GEEs are of the form

$$\sum_{i=1}^{n} \frac{\partial \boldsymbol{\mu}_i}{\partial \beta} \boldsymbol{V}_i^{-1} \Big(\boldsymbol{Y}_i - \boldsymbol{\mu}_i \Big) = 0. \tag{1.34}$$

Which is a series of nonlinear equations, typically solved using Newton-Raphson iterations. GEEs then produces the estimator $\hat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ which is consistent, and asymptotically normal (Liang and Zeger 1986).

All of these models can be efficient in the correct context, however, as with classical GLMs, both examples of parametric VGLMs have assumptions which can easily lead to a misspecified model and the accompanying loss in estimation efficiency. Firstly, the QEM model has the assumption that the association between any two binary responses can be modelled using a single parameter, and the MCMP model having assumptions which are potentially often violated when modelling physical systems. As for GEEs, Huang (2017) did show that they can form a robust model in a general vector regression scenario. However, this required the standard errors admitted by the GEE model to be corrected using a sandwich estimator of variance, which can still sometimes perform poorly for moderate to large sample sizes even under the correct specification of the correlation matrix $\boldsymbol{V}$(Kauermann and Carroll 2001). What these examples motivate, is the need for a more general VGLM which can be applied in the majority of scenarios and produce an appropriate model under only minimal assumptions. This leads naturally to the model which is introduced in following chapter, which can potentially deal with some of the

issues raised in this section.

Chapter 2

Model and Parameter Estimation

2.1 Introduction

This chapter introduces a vector semiparametric GLM (VSPGLM), which deals with some of the problems of model mispecification common in both the parametric and semiparametric methods mentioned in the previous chapter. This model is a generalisation of a semiparametric GLM (SPGLM), first published by Huang (2014), which was an extension of work conducted by Rathouz and Gao (2009), who introduced this family of SPGLMs. The defining feature of this VSPGLM, and other models in its family, is that it uses an exponential tilt reformulation of the data's underlying reference distribution. The main benefits of this approach is the lack of constraints it places on the true reference distribution, which is initially left unspecified. The model then simultaneously, and efficiently, estimates a nonparametric reference distribution and the parameters of each response's mean model. This deals with problems of specifying within vector correlations and the modelling of mixed data types, as the estimated nonparametric reference distribution is almost completely arbitrary.

This chapter will introduce the model and its semiparametric extension, noting how the vector generalisation differs from the original model of Huang (2014), while comparing and contrasting this approach to those of density ratio models which have also been developed over the last decade using similar methods (Marchese 2018; Luo and Tsai 2012). The chapter will then go on to explain how the models parameters can be estimated using nonparametric maximum empirical likelihood estimation (MELE), and detail how inference can be conducted on the $\boldsymbol{\beta}$ parameters fitted to each component's mean model. This is

done either through the use of a Wald test, using the standard errors from the empirical estimate of the asymptotic covariance matrix, or an empirical likelihood ratio test (ELRT).

2.2 Model

To start this section two definitions are provided, both of which will be used shortly.

Definition 2.1 (Dominating Measure)**.** *Let μ and ν be two measures on the measurable space $(E, \mathcal{E})$. If, for all $A \in \mathcal{E}$*

$$\mu(A) = 0 \implies \nu(A) = 0. \tag{2.1}$$

ν is said to be dominated by μ.

Definition 2.2 (Exponentially Tilted Probability Density (Butler 2007))**.** *Let y be a random variable with probability distribution $F(y)$ and density $dF(y)$, with respect to some dominating measure (Definition 2.1). If y has a finite moment generating function,*

$$\mathbb{E}[e^{\theta y}] = \int_{\mathcal{Y}} e^{\theta y} dF(y) < \infty \tag{2.2}$$

for $\theta \in \Theta$ in some neighbourhood of 0, the exponentially tilted density dF_θ of y, indexed by the tilt parameter θ, is given by

$$dF_\theta(y) = \exp\{\theta y + b(\theta)\} dF(y), \tag{2.3}$$

where $b(\theta)$ is the c.g.f of y, defined by

$$b(\theta) = -\log\left\{ \int_{\mathcal{Y}} e^{\theta y} dF(y) \right\}. \tag{2.4}$$

Let $(\boldsymbol{X}_i, \boldsymbol{Y}_i) \in \mathcal{X} \times \mathcal{Y} \subset \mathbb{R}^q \times \mathbb{R}^K$, $i = 1, \ldots, n$ be data pairs, where each vector valued response $\boldsymbol{Y}_i$ has K components, and is sampled independently from some reference distribution $F_i(\boldsymbol{y})$, conditional on the covariates $\boldsymbol{X}_i$. In this setup, the covariates $\boldsymbol{X}_i$ are either independent samples from some population or fixed by design. Here, each distribution $F_i(\boldsymbol{y})$ is assumed to originate from some multivariate exponential family, and have an associated density, dF_i, with respect to a dominating measure (Definition 2.1). Under these assumptions, the joint density of each response $\boldsymbol{Y}_i \in \mathbb{R}^K$ can be written using the exponential tilt formulation for SPGLMs of Rathouz and Gao (2009). For the VSPGLM,

the formulation has been generalized as to include vector responses $\boldsymbol{Y}_i \in \mathbb{R}^k$, and is shown in (2.5).

$$dF_i(\boldsymbol{y}) = \exp\{b_i + \boldsymbol{\theta}_i^T \boldsymbol{y}\} dF(\boldsymbol{y}), \quad i = 1, \ldots, n \tag{2.5}$$

Here $b_i = b(\boldsymbol{X}_i, \boldsymbol{\beta}, F)$ are the normalizing functions for this distribution, and the canonical parameters are $\boldsymbol{\theta}_i = (\theta_1(\boldsymbol{X}_i, \boldsymbol{\beta}, F), \ldots, \theta_K(\boldsymbol{X}_i, \boldsymbol{\beta}, F))^T$.

Throughout this chapter, unless otherwise specified, the convention that will be used is that $\boldsymbol{\beta} = \{\boldsymbol{\beta}_{(k)} \in \mathbb{R}^{q_k} | k = 1, 2, \ldots, K\}$ is the set of parameters for every component's parametric mean model. The total number of mean model parameters is then $\sum_{k=1}^{K} q_k = Q$, where each vector $\boldsymbol{\beta}_{(k)}$ has $q_k \leqslant q$ elements, and has its own vector of covariates, $\boldsymbol{X}_{(k)} \in \mathbb{R}^{q_k}$, which contains elements of the full covariate vector $\boldsymbol{X}$ for each response. The true size of each covariate vector also being dependent on any constraints that have been placed across different components mean model's. More discussion of these constraints is given in Section 2.6.

As is the case with parametric GLMs, if a Laplace transform exists for the reference distribution F within a neighbourhood of the origin, its c.g.f, in this case the normalising functions $b(\boldsymbol{X}_i, \boldsymbol{\beta}, F)$, exists, and are defined by the K dimensional integral

$$b_i = b(\boldsymbol{X}_i, \boldsymbol{\beta}, F) = -\log \left\{ \int_{\mathcal{Y}} \exp\{\boldsymbol{\theta}_i^T \boldsymbol{y}\} dF(\boldsymbol{y}) \right\} . \tag{2.6}$$

It is easy then to see that the solution for $b(\boldsymbol{X}_i, \boldsymbol{\beta}, F)$ satisfies the normalising constraint implicit in (2.5), as

$$\begin{aligned} \int_{\mathcal{Y}} dF_i(\boldsymbol{y}) &= \int_{\mathcal{Y}} \exp\{b_i + \boldsymbol{\theta}_i^T \boldsymbol{y}\} dF(\boldsymbol{y}) \\ &= \int_{\mathcal{Y}} \frac{\exp\{\boldsymbol{\theta}_i^T \boldsymbol{y}\} dF(\boldsymbol{y})}{\int_{\mathcal{Y}} \exp\{\boldsymbol{\theta}_i^T \boldsymbol{z}\} dF(\boldsymbol{z})} \\ &= 1 . \end{aligned} \tag{2.7}$$

Similarly, the distributions marginal means for each observation can be used to implicitly solve for each vector of canonical, or tilt, parameters $\boldsymbol{\theta}_i$

$$\mu_{(k)}(\boldsymbol{X}_{(k)i}^T \boldsymbol{\beta}_{(k)}) = \int_{\mathcal{Y}} y_{(k)} \exp\{b_i + \boldsymbol{\theta}_i^T \boldsymbol{y}\} dF(\boldsymbol{y}), \quad i = 1, \ldots, n, k = 1, \ldots, K . \tag{2.8}$$

Here, equations (2.6) and (2.8) show that each of the densities $dF_i(\boldsymbol{y})$, with form given by (2.5), are an exponential tilt of the reference distribution $dF(\boldsymbol{y})$, where the amount of tilt given to each response is dependent upon the response's mean, $\mu_{(k)}(\boldsymbol{X}_{(k)i}^T \boldsymbol{\beta}_{(k)})$, for each of the K components.

Another property of the SPGLM of Rathouz and Gao (2009), which generalizes to this model, is that the density shown in (2.5) also has the form of the following density ratio model

$$dF_i(\boldsymbol{y}) = \frac{w(\boldsymbol{y}, \boldsymbol{\theta}_i)}{W(\boldsymbol{\theta}_i, F)} dF(\boldsymbol{y}), \;\; i = 1, \ldots, n, \tag{2.9}$$

where each $w(\boldsymbol{y}, \boldsymbol{\theta}_i) = \exp\{\boldsymbol{\theta}_i^T \boldsymbol{y}\}$ can be considered to be the models weight functions, and each $W(\boldsymbol{\theta}_i, F) = \int_y \exp\{\boldsymbol{\theta}_i^T \boldsymbol{y}\} dF(\boldsymbol{y}) = \exp\{-b_i\}$ the associated normalizing functions for $i = 1, \ldots, n$. A similar model to this VSPGLM is then the multivariate density ratio model (MDRM) proposed by Marchese (2018), who expanded an earlier density ratio model of Luo and Tsai (2012) to multiple responses. In the MDRM the joint distribution $dF_i(\boldsymbol{y})$ for a vector response $\boldsymbol{y} \in \mathbb{R}^K$ is of the form

$$dF(\boldsymbol{y}|\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{\beta}, F) = \frac{\exp\left\{ \sum_{i=1}^{k} h_{(k)}(y_{(k)})(\boldsymbol{x}^T \boldsymbol{\beta})_{(k)} \right\} dF_0(\boldsymbol{y})}{\int \exp\left\{ \sum_{i=1}^{k} h_{(k)}(z_{(k)})(\boldsymbol{x}^T \boldsymbol{\beta})_{(k)} \right\} dF_0(\boldsymbol{z})}, \tag{2.10}$$

where $\boldsymbol{x} \in \mathbb{R}^j$, $\boldsymbol{\beta} \in \mathbb{R}^{J \times K}$ is a matrix containing the regression parameters of each mean model and $dF_0(\boldsymbol{y})$ the baseline cumulative distribution function when $\boldsymbol{x} = \boldsymbol{0}$. In the MDRM, the functions $h_{(k)}(\cdot)$ are then mappings from the response vector $\boldsymbol{y} \in \mathbb{R}^K$ to the scale of the distributions sufficient statistics. The main difference between the MDRM and the proposed VSPGLM is that the MDRM attempts regression which puts its linear predictors $\boldsymbol{x}^T \boldsymbol{\beta}$ on the scale of the canonical parameters, $\boldsymbol{\theta} = (\boldsymbol{x}^T \boldsymbol{\beta})$, whereas the VSPGLM instead attempts to model the mean functions $\mu_{(k)}(\boldsymbol{X}_{(k)}^T \boldsymbol{\beta}_{(k)})$ directly, using a parametric model. Doing regression specifically in terms of the canonical parameter results in less parameters to optimize over (the same number of parameters remain in the final model), however has downsides in terms of the interpretability. As Rathouz and Gao (2009) noted, regressing on the mean function $\mu = \mu(\boldsymbol{X}^T \boldsymbol{\beta})$ directly, allows any interpretation of changes in the regression coefficients contained in $\boldsymbol{\beta}$ to be independent of the underlying reference distribution $dF(\boldsymbol{y})$. As in the case of the MDRM, for continuous multivariate response data, the regression parameters $\boldsymbol{\beta}$, once estimated, must be re-scaled using an estimate of the asymptotic covariance matrix for $\boldsymbol{\beta}$. No such re-scaling of parameters is necessary when modelling using this VSPGLM, with the use of parametric mean models also being advantageous as it allows $\boldsymbol{\beta}$ to be interpreted as mean contrasts, or as the "treatment effect" present in the data (see Chapter 4, Section 4.6 for an example of this).

2.3 Semiparametric Extension

The exponential tilt formulation of the joint density, described in the previous section, easily leads itself to a semiparametric extension which is used in the estimation of the parameters of this model. Here the reference distribution F is now left unspecified, and treated as an infinite-dimensional parameter which is to be estimated along with the set of $\boldsymbol{\beta}$ parameters for each component's mean model. This step is one of the major advantages that semiparametric models of this form have over other competing parametric approaches, as it removes a major cause of model misspecification. With F and $\boldsymbol{\beta}$ now both treated as parameters, the semiparametric log-likelihood to jointly estimate them has the following form

$$\ell(\boldsymbol{\beta}, F|\boldsymbol{X}, \boldsymbol{Y}) = \sum_{i=1}^{n} \left\{ \log dF(\boldsymbol{Y}_i) + b_i + \boldsymbol{\theta}_i^T \boldsymbol{Y}_i \right\} . \tag{2.11}$$

One thing to note about this formulation, is that the infinitely dimensional parameter F is actually non-identifiable, as the log-likelihood is invariant to any tilting of the reference distribution. Hence, to ensure F is identifiable, it is constrained to have a some mean $\boldsymbol{\mu}$ which lies in the interior of $\mathcal{Y}$, the response space for the data. Although the exact location of $\boldsymbol{\mu}$ is unimportant, due to the invariance of the log-likelihood to any re-tilting. Therefore, F is said to be located in the parameter space of distributions functions $F_{\boldsymbol{\mu}}$, which have mean $\boldsymbol{\mu}$, and for which a Laplace transformation exists within a neighbourhood of the origin.

For a single variable response, it was shown in Huang and Rathouz (2017) that the score function for $\boldsymbol{\beta}$ is orthogonal to the nuisance tangent space of the infinite-dimensional parameter F. In this paper the following proposition was shown for any parametric or semiparametric GLM.

Proposition 2.1 (Orthogonality)**.** *The mean-model parameters $\boldsymbol{\beta}$ and the error distribution F in any generalized linear model are orthogonal.*

The implication of this result is that both the estimation of $\boldsymbol{\beta}$ and any inferences conducted on its parameters will, asymptotically, be independent of any estimation of the reference distribution F, ensuring the efficiency of this method. The orthogonality of F and $\boldsymbol{\beta}$ generalises to the proposed VSPGLM, with this proposition implicitly used in Section 2.6 to derive an empirical estimate for the asymptotic covariance matrix of $\boldsymbol{\beta}$.

2.4 Maximum Empirical Likelihood Estimation

Both the reference distribution F and mean model parameters $\boldsymbol{\beta}$ are estimated using maximum empirical likelihood estimation (MELE) (Owen 2001). To do so, the densities dF_i are replaced in the semiparametric log-likelihood (2.11) across the observed support $\{\boldsymbol{Y}_i \in \mathbb{R}^K | i = 1, \ldots, n\}$ by non-negative point probability masses $\boldsymbol{p} = \{p_i \geqslant 0 | \sum_{i=1}^n p_i = 1\}$. The probability masses $\boldsymbol{p} = (p_1, \ldots, p)^T$ are known as Vardi's or histogram estimators, and are often used in the estimation of biased sampling or density ratio models (Gill, Vardi, and Wellner 1988). Once this is done, the empirical log-likelihood has the following form

$$\ell(\boldsymbol{\beta}, \boldsymbol{p}) = \sum_{i=1}^{n} \log p_i + b_i + \boldsymbol{\theta}_i^T \boldsymbol{Y}_i \ . \tag{2.12}$$

As was the case with the original exponentially tilted reference distribution, the empirical log-likelihood is once again subject to the mean and normalisation constraints for each pair of normalising and tilt parameters $(b_i, \boldsymbol{\theta}_i) \in \mathbb{R} \times \mathbb{R}^K$. With equation (2.13) ensuring that each tilted nonparametric distribution is normalised

$$1 = \sum_{i=1}^{n} p_i \exp\{b_j + \boldsymbol{\theta}_j^T \boldsymbol{Y}_i\}, \quad j = 1, \ldots, n, \tag{2.13}$$

and equation (2.14) once again controls the amount of tilt given to each observation via the mean $\mu_{(k)}(\boldsymbol{X}_{(k)j}^T \boldsymbol{\beta}_k)$ of each response component.

$$\mu_{(k)}(\boldsymbol{X}_{(k)j}^T \boldsymbol{\beta}_k) = \sum_{i=1}^{n} p_i Y_{(k)i} \exp\{b_j + \boldsymbol{\theta}_j^T \boldsymbol{Y}_i\} \quad j = 1, \ldots, n, k = 1, \ldots, K \tag{2.14}$$

These constraints being an empirical adaption of the constraints displayed in equations (2.8) and (2.6), with the K dimensional integral now collapsing to a singular summation, due to the discretization of the densities across all K response components marginal distributions. Similar to the constraints placed on F in the previous section, a condition on equation (2.14) being solvable is that the mean vector $\boldsymbol{\mu}$ of the reference distribution estimators $\boldsymbol{p}$ lies in the convex hull of the observed support for $\boldsymbol{Y}$ (Definition 2.3). Where it is convention that the empirical likelihood $\ell(\boldsymbol{\beta}, \boldsymbol{p})$ be set to $-\infty$ if this condition is violated (Owen 2001).

Definition 2.3 (Convex Hull)**.** *The series of points $\boldsymbol{Y}_i \in \mathbb{R}^K, i = 1, \ldots, n$, form the convex hull $\mathcal{C}$, defined by*

$$\mathcal{C} = \left\{ \sum_{i=1}^{n} p_i \boldsymbol{Y}_i \ : \ p_i \in \mathbb{R}_+ \ \forall i \quad \textit{and} \quad \sum_{i=1}^{n} p_i = 1 \right\} \tag{2.15}$$

Identifiability constraints can also be placed on the vector of probability masses $\boldsymbol{p}$, by enforcing the following mean constraint

$$\boldsymbol{\mu}_0 = \sum_{i=1}^{n} \boldsymbol{Y}_i p_i, \tag{2.16}$$

for some $\boldsymbol{\mu}_0$ which lies in the interior of the convex hull formed by $\boldsymbol{Y}$. This is of little consequence however, as once again the empirical log-likelihood remains invariant to any particular choice of $\boldsymbol{\mu}$. Enforcing a specific $\boldsymbol{\mu} = \boldsymbol{\mu}_0$ has been found to be useful to increase the numerical stability of the univariate case of this model when it is fit computationally (Wurm and Rathouz 2018), but was not adopted in the approach to fit the VSPGLM outlined in Chapter 3.

To estimate $\boldsymbol{\beta}$ the profile empirical log-likelihood is formed, which is defined to be the supremum of the empirical log-likelihood over all nonparametric reference distributions $\boldsymbol{p}$, whose means are equal to $\boldsymbol{\mu}$.

$$pl(\boldsymbol{\beta}) = \sup_{\boldsymbol{p}} \ \ell(\boldsymbol{\beta}, \boldsymbol{p}) \tag{2.17}$$

The estimates for $\boldsymbol{\beta}$ and F, $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{p}}$ are then taken to be the joint maximisers $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{p}})$ of the empirical log-likelihood.

$$(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{p}}) = \arg\max \ell(\boldsymbol{\beta}, \boldsymbol{p}) \tag{2.18}$$

Using $\hat{\boldsymbol{p}}$, an empirical estimate of the reference distribution F, $\hat{F}$, can be produced via the joint empirical distribution function

$$\hat{F}_n(\boldsymbol{y}) = \sum_{i=1}^{n} \hat{p}_i \mathbb{I}\{\boldsymbol{Y}_i \leqslant \boldsymbol{y}\}. \tag{2.19}$$

When dealing with a single response, one could use the empirical distribution function and the Kalmagorov-Smirnov statistic $D_n = \sup_{\boldsymbol{y}} |F(\boldsymbol{y}) - \hat{F}_n(\boldsymbol{y})|$ to determine an adequate parametric GLM for the data (Huang 2014), however, outside of the multivariate normal distribution, there will most likely be no adequate parametric distribution function available to compute the statistic when modelling data with multiple response components.

2.5 Asymptotic Results for F and β

For the estimates $\hat{F}$ and $\hat{\boldsymbol{\beta}}$, produced by the MELE procedure, there are the following pair of results from Huang (2014). Some details of the conditions necessary for these results

are given in Appendix A.2, with the steps to prove these results given in the paper's supplementary material. More details are not given here, as full proof lies outside the scope of this chapter.

Firstly, in relation to the consistency of both $\hat{F}$ and $\hat{\boldsymbol{\beta}}$, there is the following lemma.

Lemma 2.1. *As* $n \to \infty$,

$$\hat{\boldsymbol{\beta}} \to \boldsymbol{\beta}^* \text{ and } \hat{F} \to F^* \tag{2.20}$$

in probability, relative to the weak topology. Here, the weak topology on the infinite-dimensional parameter space for F *uses the distance* $\|F_1 - F_2\|_{\mathcal{H}_l} = \sup_{h \in \mathcal{H}_l} \int h(dF_1 - dF_2)$, *where* $\mathcal{H}_l \coloneqq \{\mathbb{I}_{\{y \leqslant r\}} : r \in \mathcal{Y} \subset \mathbb{R}\}$ *is the set of all left indicator functions.* F^* *is then the true non computable reference distribution, which belongs to the class of functions with mean* $\boldsymbol{\mu}$ *and a Laplace transform in a neighbourhood of the origin.*

Secondly, for the the univariate case of this model, the following proposition was shown using previous results from Murphy and Van Der Vaart (2000).

Proposition 2.2. *As* $n \to \infty$ *the the combined vector of regression and distribution parameters,* $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*, \hat{F} - F^*)^T$, *satisfies*

$$\sqrt{n} \begin{pmatrix} \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \\ \hat{F} - F^* \end{pmatrix} \xrightarrow{\mathcal{D}} \begin{pmatrix} G_1 \\ G_2 \end{pmatrix} \tag{2.21}$$

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

$$W_1 = \left(\mathbb{E}_{\boldsymbol{X}} \left[\frac{\mu'(\boldsymbol{X}^T\boldsymbol{\beta})^2 \boldsymbol{X}\boldsymbol{X}^T}{V(\boldsymbol{X}, \boldsymbol{\beta}^*, F^*)} \right] \right)^{-1} \tag{2.22}$$

where $V(\boldsymbol{X}, \boldsymbol{\beta}^*, F^*)$ *is the conditional variance function of* $\boldsymbol{Y}$ *given* $\boldsymbol{X}$, *under the true values of the parameters* $\boldsymbol{\beta}^*$ *and* F^*, *and* G_2 *is mean zero Gaussian process which is independent of* G_1.

Both of these results, the consistency and joint asymptotic normality of $\hat{F}$ and $\hat{\boldsymbol{\beta}}$, as well as the asymptotic covariance matrix for $\sqrt{n}\,(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)$, generalise to the VSPGLM, with an empirical estimate of W_1 introduced in the following section.

2.6 Score and Covariance Matrix for β

In this section we introduce and derive the empirical score function $\boldsymbol{S}(\boldsymbol{\beta})$ and the empirical Fisher information matrix $\mathcal{I}(\boldsymbol{\beta})$ for the mean model parameters $\boldsymbol{\beta}$. The derivation of both are similar to those of Rathouz and Gao (2009), however, we are now dealing with multiple sets of mean model parameters. Note that, due to the orthogonality of F and $\boldsymbol{\beta}$, $\boldsymbol{p}$ can be treated as fixed in these derivations.

For ease of notation, and to clean up subsequent equations, during the remainder of this section we will place all vectors of parameters $\boldsymbol{\beta}_{(k)} \in \mathbb{R}^{q_k}$ into a single column vector

$$\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_{(1)} \\ \boldsymbol{\beta}_{(2)} \\ \vdots \\ \boldsymbol{\beta}_{(K)} \end{pmatrix} \in \mathbb{R}^{Q}, \quad \text{where } \sum_{k=1}^{K} q_k = Q. \tag{2.23}$$

As will be shown in Chapter 3, constraints can be placed on $\boldsymbol{\beta}$ such that parameters can be common across multiple mean models. These constraints can range from having a common intercept coefficient across all mean models to having the entirety of the regression coefficients constrained to be equal across each model. For the components with constrained parameters, the same link function is used on each constrained mean model, as constraining mean parameters but using different link functions destroys the interpretability of the fitted model. Therefore, in this section, K is no longer the number of components in the response vector $\boldsymbol{Y}_i$, but the number of separate mean models. For this section, the notation of $K' \geqslant K$ will be used to denote the number of components in the vector responses $\boldsymbol{Y}_i$, with $k' \in k$ denoting the set of response components which are modelled using the kth set of separate mean model parameters $\boldsymbol{\beta}_{(k)}$, each using the same link function.

To derive the empirical Fisher information matrix, first, the empirical score for $\boldsymbol{\beta}$ must be derived. To do so, suppose $\boldsymbol{y}$ originates from the joint density

$$dF(\boldsymbol{y};\boldsymbol{\theta}) = \exp\Big\{\boldsymbol{y}^T\boldsymbol{\theta} + b(\boldsymbol{\theta})\Big\}dF(\boldsymbol{y}).$$

Then, as was the case in equation (1.9), $\boldsymbol{y}$ has the log likelihood

$$\begin{aligned} \ell(\boldsymbol{\theta};\boldsymbol{y}) &= \ln dF(\boldsymbol{y};\boldsymbol{\theta}) \\ &= \boldsymbol{y}^T\boldsymbol{\theta} + b(\boldsymbol{\theta}) + \ln dF(\boldsymbol{y}). \end{aligned} \tag{2.24}$$

The score for the the vector of canonical parameters $\boldsymbol{\theta}$ is given by

$$\begin{aligned}\boldsymbol{S}(\boldsymbol{\theta}) &= \frac{\partial \ell}{\partial \boldsymbol{\theta}} \\ &= \boldsymbol{y} + \nabla b(\boldsymbol{\theta}).\end{aligned} \tag{2.25}$$

and

$$\begin{aligned}\boldsymbol{S}(\boldsymbol{\theta}) &= \nabla \ln dF(\boldsymbol{y};\boldsymbol{\theta}) \\ &= \frac{1}{dF(\boldsymbol{y};\boldsymbol{\theta})}\nabla dF(\boldsymbol{y};\boldsymbol{\theta}).\end{aligned} \tag{2.26}$$

Under mild regularity conditions (Appendix A.1), which are satisfied if the data originates from an exponential family, the following result holds

$$\begin{aligned}\mathbb{E}[\boldsymbol{S}(\boldsymbol{\theta})] &= \int_{\mathcal{Y}} \boldsymbol{S}(\boldsymbol{\theta}) dF(\boldsymbol{y};\boldsymbol{\theta}) \\ &= \int_{\mathcal{Y}} \frac{1}{dF(\boldsymbol{y};\boldsymbol{\theta})}\nabla dF(\boldsymbol{y};\boldsymbol{\theta}) dF(\boldsymbol{y};\boldsymbol{\theta}) \\ &= \int_{\mathcal{Y}} \nabla dF(\boldsymbol{y};\boldsymbol{\theta}) \\ &= \nabla \int_{\mathcal{Y}} dF(\boldsymbol{y};\boldsymbol{\theta}) \\ &= \nabla 1 \\ &= \mathbf{0}.\end{aligned} \tag{2.27}$$

Using this fact, (2.25) can be rearranged to obtain

$$\begin{aligned} 0 &= \mathbb{E}[\boldsymbol{S}(\boldsymbol{\theta})] \\ &= \mathbb{E}[\boldsymbol{Y}] + \nabla b(\boldsymbol{\theta}) \\ \implies \mathbb{E}[\boldsymbol{Y}] &= -\nabla b(\boldsymbol{\theta}).\end{aligned} \tag{2.28}$$

Similarly, a standard result for the score function is that

$$\begin{aligned} 0 &= \nabla \mathbb{E}[\boldsymbol{S}(\boldsymbol{\theta})] \\ &= \nabla \int_{\mathcal{Y}} \nabla \ln \Big(dF(\boldsymbol{y};\boldsymbol{\theta}) \Big) dF(\boldsymbol{y};\boldsymbol{\theta}) \\ &= \int_{\mathcal{Y}} \Big(\nabla^2 \ln \Big(dF(\boldsymbol{y};\boldsymbol{\theta}) \Big) dF(\boldsymbol{y};\boldsymbol{\theta}) + \nabla \ln \Big(dF(\boldsymbol{y};\boldsymbol{\theta}) \Big) \Big(\nabla dF(\boldsymbol{y};\boldsymbol{\theta}) \Big)^T \Big) \\ &= \int_{\mathcal{Y}} \nabla^2 \ln \Big(dF(\boldsymbol{y};\boldsymbol{\theta}) \Big) dF(\boldsymbol{y};\boldsymbol{\theta}) + \int_{\mathcal{Y}} \nabla \ln \Big(dF(\boldsymbol{y};\boldsymbol{\theta}) \Big) \Big(\nabla dF(\boldsymbol{y};\boldsymbol{\theta}) \Big)^T \\ &= \int_{\mathcal{Y}} \nabla^2 \ln \Big(dF(\boldsymbol{y};\boldsymbol{\theta}) \Big) dF(\boldsymbol{y};\boldsymbol{\theta}) + \int_{\mathcal{Y}} \nabla \ln \Big(dF(\boldsymbol{y};\boldsymbol{\theta}) \Big) \Big(\nabla \ln \Big(dF(\boldsymbol{y};\boldsymbol{\theta}) \Big) \Big)^T dF(\boldsymbol{y};\boldsymbol{\theta}) \\ &= \mathbb{E}\Big[\nabla^2 \ln \Big(dF(\boldsymbol{y};\boldsymbol{\theta}) \Big) \Big] + \mathbb{E}\Bigg[\Big(\nabla \ln \Big(dF(\boldsymbol{y};\boldsymbol{\theta}) \Big) \Big) \Big(\nabla \ln \Big(dF(\boldsymbol{y};\boldsymbol{\theta}) \Big) \Big)^T \Bigg] \\ &= \mathbb{E}\Big[\nabla^2 \boldsymbol{S}(\boldsymbol{\theta}) \Big] + \mathbb{E}\Big[\boldsymbol{S}(\boldsymbol{\theta}) \boldsymbol{S}(\boldsymbol{\theta})^T \Big].\end{aligned} \tag{2.29}$$

Therefore,

$$\mathbb{E}\Big[\nabla^2 \boldsymbol{S}(\boldsymbol{\theta})\Big] = -\mathbb{E}\Big[\boldsymbol{S}(\boldsymbol{\theta})\boldsymbol{S}(\boldsymbol{\theta})^T\Big]. \tag{2.30}$$

Hence, (2.28) implies

$$\begin{aligned}\mathbb{E}\Big[\nabla \boldsymbol{S}(\boldsymbol{\theta})\Big] &= \mathbb{E}\Big[\nabla^2 b(\boldsymbol{\theta})\Big] \\ &= -\mathbb{E}\Big[(\boldsymbol{y} + \nabla b(\boldsymbol{\theta}))(\boldsymbol{y} + \nabla b(\boldsymbol{\theta}))^T\Big],\end{aligned} \tag{2.31}$$

and as $\mathbb{E}\Big[\nabla \boldsymbol{S}(\boldsymbol{\theta})\Big] = \mathbf{0}$,

$$\begin{aligned} 0 =& \nabla^2 b(\boldsymbol{\theta}) + \Sigma_{\boldsymbol{Y}} \\ \implies \Sigma_{\boldsymbol{Y}} =& -\nabla^2 b(\boldsymbol{\theta}),\end{aligned} \tag{2.32}$$

where

$$\Sigma_{\boldsymbol{Y}} = \mathbb{E}\Big[(\boldsymbol{y} + \nabla b(\boldsymbol{\theta}))(\boldsymbol{y} + \nabla b(\boldsymbol{\theta}))^T\Big]\ . \tag{2.33}$$

Using the chain rule on the log-likelihood of the joint density, under the assumption that each observation $\boldsymbol{Y}_i \in \mathbb{R}^{K'}$ is independent, the score function for each *iid* vector observation is

$$\begin{aligned}\boldsymbol{S}(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \frac{\partial \ell_i}{\partial \boldsymbol{\beta}} \\ &= \sum_{i=1}^{n} \frac{\partial \ell_i}{\partial \boldsymbol{\theta}_i} \frac{\partial \boldsymbol{\theta}_i}{\partial \boldsymbol{\mu}_i} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}.\end{aligned} \tag{2.34}$$

As was previously stated, the joint density for each observation $\boldsymbol{Y}_i$ is an exponential tilting of the reference distribution F, where now each tilted density dF_i has been replaced by the estimators $\boldsymbol{p} = (p_1, \ldots, p_n)^T$, and for which the empirical log-likelihood is

$$\ell(\boldsymbol{\beta}, \boldsymbol{p}) = \sum_{i=1}^{n} \log p_i + b_i + \boldsymbol{\theta}_i^T \boldsymbol{Y}_i. \tag{2.35}$$

To calculate each derivative term in (2.34), one can make the normalising constant b_i a function of $\boldsymbol{\theta}_i$ using an empirical analogue of the normalising constraints shown in (2.13)

$$b(\boldsymbol{\theta}) = -\log\Big\{\sum_{i=1}^{n} p_i \exp\Big\{\boldsymbol{\theta}^T \boldsymbol{Y}_i\Big\}\Big\}. \tag{2.36}$$

The gradient of $b(\boldsymbol{\theta})$ for a single observation is then

$$\begin{aligned}\nabla b(\boldsymbol{\theta}) &= -\nabla\left(\log\left\{\sum_{i=1}^{n} p_i \exp\left\{\boldsymbol{\theta}^T \boldsymbol{Y}_i\right\}\right\}\right) \\ &= -\frac{1}{\sum_{i=1}^{n} p_i \exp\left\{\boldsymbol{\theta}^T \boldsymbol{Y}_i\right\}} \nabla\left(\sum_{i=1}^{n} p_i \exp\left\{\boldsymbol{\theta}^T \boldsymbol{Y}_i\right\}\right) \\ &= -\frac{\sum_{j=1}^{n} \boldsymbol{Y}_j p_j \exp\left\{\boldsymbol{\theta}_i^T \boldsymbol{Y}_j\right\}}{\sum_{j=1}^{n} p_j \exp\left\{\boldsymbol{\theta}_i^T \boldsymbol{Y}_j\right\}}. \end{aligned} \tag{2.37}$$

Substituting this new relation for $b(\boldsymbol{\theta})$ into the the empirical log-likelihood results in

$$\ell(\boldsymbol{\beta}, \boldsymbol{p}) = \sum_{i=1}^{n} \log p_i - \log\left\{\sum_{j=1}^{n} p_j \exp\left\{\boldsymbol{\theta}_i^T \boldsymbol{Y}_j\right\}\right\} + \boldsymbol{\theta}_i^T \boldsymbol{Y}_i \,. \tag{2.38}$$

Using this new expression, which is now only in terms of the nonparametric reference distribution $\boldsymbol{p}$ and the tilt parameters $\boldsymbol{\theta}_i$, one can separately evaluate each derivative in the score function for $\boldsymbol{\beta}$ (2.34). The first derivative present in the score function, $\frac{\partial \ell_i}{\partial \boldsymbol{\beta}}$, now evaluates to

$$\begin{aligned}\frac{\partial \ell_i}{\partial \boldsymbol{\theta}_i} &= \frac{\partial}{\partial \boldsymbol{\theta}_i}\left(\log p_i + b_i + \boldsymbol{\theta}_i^T \boldsymbol{Y}_i\right) \\ &= \boldsymbol{Y}_i - \frac{\sum_{j=1}^{n} \boldsymbol{Y}_j p_j \exp\left\{\boldsymbol{\theta}_i^T \boldsymbol{Y}_j\right\}}{\sum_{j=1}^{n} p_j \exp\left\{\boldsymbol{\theta}_i^T \boldsymbol{Y}_j\right\}} \\ &= \boldsymbol{Y}_i - \sum_{j=1}^{n} \boldsymbol{Y}_j p_j \exp\left\{\boldsymbol{\theta}_i^T \boldsymbol{Y}_j + b_i\right\} \\ &= \boldsymbol{Y}_i - \boldsymbol{\mu}_i. \end{aligned} \tag{2.39}$$

Then, using (2.28) we have that

$$\nabla b(\boldsymbol{\theta}) = -\boldsymbol{\mu}, \tag{2.40}$$

therefore, the next derivative is

$$\frac{\partial \boldsymbol{\theta}_i}{\partial \boldsymbol{\mu}_i} = \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\theta}_i}\right)^{-1} = \left(\nabla^2 b_i(\boldsymbol{\theta}_i)\right)^{-1}. \tag{2.41}$$

Using the result derived for $\Sigma_{\boldsymbol{Y}}$ (2.32), this derivative is equal to the inverse of the empirical covariance matrix for $\boldsymbol{Y}$. That is,

$$\begin{aligned}\left(\nabla^2 b(\boldsymbol{\theta})\right)^{-1} &= \Sigma_{\boldsymbol{Y}}^{-1} \\ &= \left\{\sum_{j=1}^{n} p_j (\boldsymbol{Y}_j - \boldsymbol{\mu}_j)(\boldsymbol{Y}_j - \boldsymbol{\mu}_j)^T \exp\left\{\boldsymbol{\theta}^T \boldsymbol{Y}_j + b\right\}\right\}^{-1}. \end{aligned} \tag{2.42}$$

The final derivative in the score is the partial derivative of the mean functions with respect to the regression coefficient vector $\boldsymbol{\beta}$.

$$\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}}$$

This resulting derivative is denoted as $\boldsymbol{T}$, where $\boldsymbol{T}$ is a $K' \times Q$ matrix for which each of the K' rows are given by the following vector

$$\boldsymbol{T}_{(k)} = \frac{\partial \mu_{(k)}}{\partial \boldsymbol{\beta}} = (\mu'_{(k)}(\boldsymbol{X}_{(k)}^T \boldsymbol{\beta}_{(1)})\boldsymbol{X}_{(k)}^T, \cdots, \mu'_{(k)}(\boldsymbol{X}_{(k)}^T \boldsymbol{\beta}_{(K)})\boldsymbol{X}_{(k)}^T), k = 1, \ldots, K'.$$

Multiplying these three derivatives together then produces the empirical score function for $\boldsymbol{\beta}$

$$\boldsymbol{S}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{T}_i^T \Sigma_{\boldsymbol{Y}_i}^{-1} (\boldsymbol{Y}_i - \boldsymbol{\mu}_i). \tag{2.43}$$

If the data originates from an exponential family, the Fisher information matrix is defined as being the covariance of $\boldsymbol{S}(\boldsymbol{\beta})$

$$\mathcal{I}(\boldsymbol{\beta}) = \mathbb{E}[\boldsymbol{S}(\boldsymbol{\beta})\boldsymbol{S}(\boldsymbol{\beta})^T] \ . \tag{2.44}$$

For a single observation $\boldsymbol{Y} \in \mathbb{R}^{K'}$, the score has the outer product

$$\boldsymbol{S}(\boldsymbol{\beta})\boldsymbol{S}(\boldsymbol{\beta})^T = \boldsymbol{T}^T \Sigma_{\boldsymbol{Y}}^{-1} (\boldsymbol{Y} - \boldsymbol{\mu})(\boldsymbol{Y} - \boldsymbol{\mu})^T \Sigma_{\boldsymbol{Y}}^{-1} \boldsymbol{T}. \tag{2.45}$$

Taking the expectation of this expression and using the results derived previously

$$\begin{aligned}
\mathcal{I}(\boldsymbol{\beta}) &= \mathbb{E}[\boldsymbol{S}(\boldsymbol{\beta})\boldsymbol{S}(\boldsymbol{\beta})^T] \\
&= \mathbb{E}[\boldsymbol{T}^T \Sigma_{\boldsymbol{Y}}^{-1} (\boldsymbol{Y} - \boldsymbol{\mu})(\boldsymbol{Y} - \boldsymbol{\mu})^T \Sigma_{\boldsymbol{Y}}^{-1} \boldsymbol{T}] \\
&= \boldsymbol{T}^T \Sigma_{\boldsymbol{Y}}^{-1} \underbrace{\mathbb{E}[(\boldsymbol{Y} - \boldsymbol{\mu})(\boldsymbol{Y} - \boldsymbol{\mu})^T]}_{\Sigma_{\boldsymbol{Y}}} \Sigma_{\boldsymbol{Y}}^{-1} \boldsymbol{T} \\
&= \boldsymbol{T}^T \Sigma_{\boldsymbol{Y}}^{-1} \Sigma_{\boldsymbol{Y}} \Sigma_{\boldsymbol{Y}}^{-1} \boldsymbol{T} \\
&= \boldsymbol{T}^T \Sigma_{\boldsymbol{Y}}^{-1} \boldsymbol{T},
\end{aligned} \tag{2.46}$$

and in the case of n independent observations of $\boldsymbol{Y} \in \mathbb{R}^{K'}$,

$$\mathcal{I}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{T}_i^T \Sigma_{\boldsymbol{Y}_i}^{-1} \boldsymbol{T}_i \ . \tag{2.47}$$

It is easy to check that in the case that $k = 1$, this expression collapses into the estimate of the Fisher information matrix originally derived by Rathouz and Gao (2009), when introducing this family of SPGLMs. However, in this, the multivariate case, the expression

does not simplify as easily, partially due to the consideration of constraints across mean models. The inverse of (2.47), which will be denoted as $\hat{\Sigma}$, can then be used as an estimate of the asymptotic covariance matrix W_1, from Proposition 2.2. One can also see that the another way to express $\hat{\Sigma}$ is as the inverse of the variance of the gradient of the profile empirical log-likelihood with respect to $\boldsymbol{\beta}$

$$\hat{\Sigma} = \left\{ \text{Var}\Big[\nabla pl(\boldsymbol{\beta})\Big] \right\}^{-1}. \tag{2.48}$$

2.7 Parameter Inference

Inference on the regression parameter estimate $\hat{\boldsymbol{\beta}}$ can be conducted in several ways. Firstly, for testing the null hypothesis $H_0 : \beta_{(k)j} = 0$, for any $k = 1, 2, \ldots, K, j = 1, 2, \ldots, q_k$, a Wald test can be used, as was done for the univariate case of this model in Huang and Rathouz (2012), and was implied by Proposition 2.2. To do so, it is assumed that under H_0

$$\frac{\hat{\beta}_{(k)i}}{\text{se}\big(\hat{\beta}_{(k)i}\big)} = \frac{\hat{\beta}_{(k)i}}{\sqrt{\hat{\Sigma}_{\hat{\beta}_{(k)i}}}} \sim t_{n-q_k}, \tag{2.49}$$

where q_k is the number of parameters in the kth marginal model and $\hat{\Sigma}_{\hat{\beta}_{(k)i}}$ are the standard errors from the estimate of the asymptotic covariance matrix (2.48).

One can also perform likelihood based inference on the $\boldsymbol{\beta}$ parameters using the profile empirical likelihood function $pl(\boldsymbol{\beta})$. This is done by profiling out the F parameters from the empirical likelihood function for a fixed $\boldsymbol{\beta}$, from which the profile empirical log-likelihood ratio statistic can be used as a method to provide inference on $\boldsymbol{\beta}$. Inference of this kind was the preferred method of Huang and Rathouz (2012), over the the use of a Wald test, as it has the correct asymptotic properties necessary for inference without requiring the computation of the standard errors $\text{se}\big(\hat{\beta}_{(k)i}\big)$, and also allows for non-symmetric confidence intervals to be constructed for the mean model parameters. Based on results from Murphy and Van Der Vaart (2000) (Appendix A.2), we have the following propositions for the asymptotic χ^2 distribution of the profile empirical log-likelihood ratio statistic.

Proposition 2.3 (Point Hypothesis)**.** *Under the null hypothesis $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}^* \in \mathbb{R}^Q$, the profile empirical log-likelihood ratio statistic*

$$2\{pl(\hat{\boldsymbol{\beta}}) - pl(\boldsymbol{\beta}^*)\},$$

is asymptotically χ^2 in distribution with Q degrees of freedom as $n \to \infty$.

Details of how a proof to this result could be constructed are given in the supplementary materials of Huang (2014), which outlines the steps to verify results from Murphy and Van Der Vaart (2000) (Appendix A.2). This thesis will provide simulations to verify this, and the following proposition, in Chapter 5.

Proposition 2.3 then naturally generalises to the testing of composite hypothesis on $\boldsymbol{\beta}$. Where, now we consider the null hypothesis where $\boldsymbol{\beta}$ lies in the r dimensional subspace of $\mathbb{R}^Q$, $1 \leqslant r \leqslant Q$, defined by

$$B_0 = \{\boldsymbol{\beta} \in \mathbb{R}^Q : \ \boldsymbol{M}\boldsymbol{\beta} = \gamma\}, \tag{2.50}$$

where $\boldsymbol{M}$ is a given matrix with rank r. If we now define

$$\hat{\boldsymbol{\beta}}_0 = \arg\max_{\boldsymbol{\beta} \in B_0} pl(\boldsymbol{\beta}), \tag{2.51}$$

then we have the following corollary.

Corollary 2.1 (Composite Hypothesis)**.** *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*

$$2\{pl(\hat{\boldsymbol{\beta}}) - pl(\hat{\boldsymbol{\beta}}_0)\}$$

is asymptotically χ^2 *in distribution with* r *degrees of freedom as* $n \to \infty$*.*

For finite samples the empirical, likelihood ratio can instead be compared to a $rF_{r,n-Q}$ distribution to perform joint inference on a model. This result following from the fact that if a random variable X follows a F distribution with degrees of freedom r and $n - Q$, then as $n \to \infty$, X converges in probability to $\frac{\chi^2_r}{r}$. That is,

$$rF_{r,n-Q} = \chi^2_r + o_P(1), \tag{2.52}$$

which is the calibration that will be used for the ELRTs conducted in the applications and examples given in Chapter 4, and the simulation studies shown in Chapter 5, Section 5.8.

Chapter 3 will now outline the details of the current computational procedures used to fit this model in MATLAB, and what constraints have been considered.

Chapter 3

Fitting the Model

3.1 Introduction

This chapter gives details of how VSPGLMs mean model and reference distribution parameters $\boldsymbol{\beta}$ and $\boldsymbol{p}$ can be estimated computationally. The estimation procedure is currently written in `MATLAB`, which uses the function `fmincon` to solve the nonlinear constrained optimization problem induced by the mean and normalization equality constraints present in the VSPGLMs formulation. Currently the **R** package **gldrm** (generalized linear density ratio model), written by Wurm and Rathouz (2018) and available on CRAN [1], computes parameter estimates for the univariate case of this model. Therefore, throughout this chapter we will also be comparing the approach presented here and that taken by the **gldrm** package.

A GitHub repository containing this method's code and example usages, some of which are the applications described in Chapter 4, is also publicly available [2]. However, due to updates in the code and the construction of the data used, some of the syntax may vary slightly to the code examples presented to fit each model.

3.2 Model Interface

Currently the model is fitted using a wrapper function `fit_vspglm`, which acts as an interface to the user. The main modelling argument to `fit_vspglm` is a string formula

[1] https://cran.r-project.org/web/packages/gldrm/index.html
[2] https://github.com/gden173/vspglm

object, which uses similar syntax to the formulas used in common **R** functions for fitting regression models, such as **lm** and **glm**, with separate formulas placed in a string array for every separate unconstrained mean model. The formula syntax to fit the following separate marginal mean models

$$g_1(\mu_{(1)}) = \boldsymbol{X}_{(1)}^T \boldsymbol{\beta}_{(1)} \quad g_2(\mu_{(2)}) = \boldsymbol{X}_{(2)}^T \boldsymbol{\beta}_{(2)}, \tag{3.1}$$

for link functions $g_1(\cdot)$ and $g_2(\cdot)$, is

```
# Separate models
model = fit_vspglm(["y_1 ~ x_1", "y_2 ~ x_2"], tbl, links);
```

Here the `tbl` argument is a `MATLAB` table which contains all the response components and covariates of the model, and `links` is a cell array of strings which specify the link functions used for each mean model. Currently `fit_vspglm` only allows the identity, log, logit and inverse link functions as options for this argument.

$$\mu, \ \log(\mu), \ \log\left(\frac{\mu}{1-\mu}\right), \ \frac{1}{\mu} \tag{3.2}$$

The object returned by `fit_vspglm` is a structure which has all the necessary information to analyze the fitted model. This includes the log-likelihood achieved, as well summary of which mean model parameters were found to be significant and their estimates. The structure also contains the estimated means, empirical covariance matrices, reference distribution parameters, and the tilt and normalising parameters for every observation. This should allow users to dig deep into each model and perform extra analysis or hypothesis testing, however, the most relevant object returned in most cases will be the regression coefficient summary tables.

Interfacing with the user through a formula object allows for the model's constraints to be specified inside the formula, without any additional optional arguments. A minimal example for a model which can be constrained to have the same $\boldsymbol{\beta}$ parameter over both margins but different covariates, *i.e.*, the model

$$g(\mu_{(1)}) = \boldsymbol{X}_{(1)}^T \boldsymbol{\beta}, \quad g(\mu_{(2)}) = \boldsymbol{X}_{(2)}^T \boldsymbol{\beta}, \tag{3.3}$$

can be fit using the formula syntax

```
# Models constrained to have the same regression parameters
# using different covariates x_1 and x_2
model = fit_vspglm(["(y_1, y_2) ~ ((x_1&x_2))], tbl, links);
```

where the ampersand indicates a common regression coefficient for each covariate grouped inside the nested set of parentheses. In the case where we only want to constrain the intercept coefficient across both models,

$$g(\mu_{(1)}) = \beta_0 + \beta_{(1)1}\boldsymbol{X}_{(1)}, \quad g(\mu_{(2)}) = \beta_0 + \beta_{(2)1}\boldsymbol{X}_{(2)}, \tag{3.4}$$

the formula syntax becomes

```
# Models constrained to have equal intercepts only
model = fit_vspglm(["(y_1, y_2) ~ ((x_1&0), (0&x_2))"], tbl, links);
```

here the 0 inside the nested parentheses indicates that no covariate for the associated mean model should be included in that particular grouping of covariates for a common regression coefficient. Finally, in the case where both response are modelled with equal covariates and common and equal $\boldsymbol{\beta}$ parameters,

$$g(\mu_{(1)}) = \boldsymbol{X}^T\boldsymbol{\beta} = g(\mu_{(2)}), \tag{3.5}$$

the constrained formula syntax collapses to

```
# Models constrained to have equal regression parameters
# using using the same covariate x
model = fit_vspglm(["(y_1, y_2) ~ x"], tbl, links);
```

More documentation on how to use the `fit_vspglm` function can be found on the GitHub repository.

3.3 Optimization

As was mentioned in the introduction, the constrained nonlinear optimization problem which is necessary to fit the VSPGLM is performed by the `MATLAB` function `fmincon` [3].

[3]Documentation is located at this address https://au.mathworks.com/help/optim/ug/fmincon.html#busog7r_seealso.

`fmincon` takes a series of optional arguments, and for this model the ones which are used are the objective function and its gradient, the initial parameter values, their upper and lower bounds, and a nonlinear constraint argument for both the mean and normalisation constraints and their gradients. An optional argument to `fit_vspglm` is which algorithm `fmincon` will use to perform the underlying optimization procedure. The main options are an interior point algorithm and sequential quadratic programming (SQP) algorithm. For best performance the underlying optimization routine can be relatively problem specific, with the SQP algorithm performing best when dealing with discrete response data, and the interior point algorithm having better performance on datasets with a larger number of continuous vector responses. This is all relative however, with the computational time taken to fit data generated during simulation studies performed in Chapter 5 fluctuating between datasets generated by the same model.

The current optimization procedure is to use `fmincon` to solve for all the VSPGLMs parameters simultaneously. Therefore, for each observation the optimization simultaneously solves for the observations normalising function b_i, probability mass p_i and vector of tilt parameters $\boldsymbol{\theta}_i$. Hence, for n observations of a K dimensional response vector $\boldsymbol{Y}_i$, where each response $Y_{(k)i}, k = 1, \ldots, K$ depends on q_k covariates $\boldsymbol{X}_{(k)i}$, the current method estimates $(2+K)n + \sum_{k=1}^{K} q_k = (2+K)n + Q$ parameters. These are,

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

In the case where the modeller would like constraints on the mean model parameters across responses, the number of variables which are optimized over remains unchanged. A linear equality constraint matrix is instead used, which is part of `fmincon`'s optional arguments.

The objective function passed to `fmincon` is the empirical log-likelihood

$$\ell(\boldsymbol{\beta}, \tilde{\boldsymbol{p}}, \boldsymbol{\theta}, \boldsymbol{b}) = -\sum_{i=1}^{n} \tilde{p}_i + b_i + \boldsymbol{\theta}_i^T \boldsymbol{Y}_i \ , \tag{3.6}$$

where now $\tilde{p}_i = \log(p_i)$, as a log-transformation of the reference distribution probability masses ensures that $p_i \geqslant 0, i = 1, \ldots, n$ without the need for any additional constraints. As was discussed in Chapter 2, the log-likelihood function shown in (3.6) is minimised

subject to both the mean and normalisation constraints for each observation $j = 1, \ldots, n$, which are shown in (3.7)

$$0 = 1 - \sum_{i=1}^{n} \exp\{b_j + \boldsymbol{\theta}_j^T \boldsymbol{Y}_i + \tilde{p}_i\}, \quad j = 1, \ldots, n, \tag{3.7}$$

and (3.8), now using $\tilde{p}_i$.

$$0 = \mu_{(k)}(\boldsymbol{X}_{(k)j}^T \boldsymbol{\beta}_k) - \sum_{i=1}^{n} Y_{(k)i} \exp\{b_j + \boldsymbol{\theta}_j^T \boldsymbol{Y}_i + \tilde{p}_i\} \quad j = 1, \ldots, n, k = 1, \ldots, K \ . \tag{3.8}$$

This set of constraints also means that we do not need to specify that $\boldsymbol{p}$ satisfies the normalising constraint

$$\sum_{i=1}^{n} p_i = 1, \tag{3.9}$$

as each exponentially tilted distribution is already normalised (3.7).

The arguments passed to `fmincon` also include the gradient of the empirical log-likelihood, as well as the gradients of both the mean and normalising constraints. The gradient of empirical log-likelihood is a vector of length $Q + n(K + 2)$, with the form

$$\nabla \ell(\boldsymbol{\beta}, \tilde{\boldsymbol{p}}, \boldsymbol{\theta}, \boldsymbol{b}) = \begin{pmatrix} \frac{\partial \ell}{\partial \boldsymbol{\beta}} \\ \frac{\partial \ell}{\partial \tilde{\boldsymbol{p}}} \\ \frac{\partial \ell}{\partial \boldsymbol{\theta}} \\ \frac{\partial \ell}{\partial \boldsymbol{b}} \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ -\mathbf{1} \\ -\boldsymbol{Y}_i, \ i = 1, \ldots, n \\ -\mathbf{1} \end{pmatrix}. \tag{3.10}$$

With the gradient of the normalising constraint shown in (3.11)

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\beta}} &\implies \mathbf{0} \\ \frac{\partial}{\partial b_j} &\implies -\sum_{i=1}^{n} \exp\{b_j + \boldsymbol{\theta}_j^T \boldsymbol{Y}_i + \tilde{p}_i\}, \quad j = 1, \ldots, n \\ \frac{\partial}{\partial \boldsymbol{\theta}_j} &\implies -\sum_{i=1}^{n} \boldsymbol{Y}_i \exp\{b_j + \boldsymbol{\theta}_j^T + \tilde{p}_i\}, \quad j = 1, \ldots, n \\ \frac{\partial}{\partial \tilde{p}_i} &\implies -\sum_{j=1}^{n} \exp\{b_j + \boldsymbol{\theta}_j^T \boldsymbol{Y}_i + \tilde{p}_i\}, \quad i = 1, \ldots, n, \end{aligned} \tag{3.11}$$

and the mean constraint gradient shown in (3.12).

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\beta}} &\implies \mu'_{(k)}(\boldsymbol{X}^T_{(k)j}\boldsymbol{\beta}_{(k)})\boldsymbol{X}_{(k)j}, \quad k = 1, \ldots, K, j = 1, \ldots, n \\
\frac{\partial}{\partial b_i} &\implies -\sum_{i=1}^{n} Y_{(k)i} \exp\{b_j + \boldsymbol{\theta}_j^T \boldsymbol{Y}_i + \tilde{p}_i\} \quad j = 1, \ldots, n, k = 1, \ldots, K \\
\frac{\partial}{\partial \boldsymbol{\theta}_i} &\implies -\sum_{i=1}^{n} \boldsymbol{Y}_i \boldsymbol{Y}_i^T \exp\{b_j + \boldsymbol{\theta}_j^T \boldsymbol{Y}_i + \tilde{p}_i\} \quad j = 1, \ldots, n \\
\frac{\partial}{\partial \tilde{p}_i} &\implies \mathbf{0}, \quad i = 1, \ldots, n
\end{aligned} \tag{3.12}$$

To ensure that the starting point of the algorithm is feasible, and the initial mean vector $\boldsymbol{\mu}^{(0)}$ lies in the interior of the the convex hull, the initial point $\boldsymbol{z}^{(0)}$ is used

$$\begin{aligned}
\boldsymbol{z}^{(0)} &= (\boldsymbol{\beta}^{(0)T}, \tilde{\boldsymbol{p}}^{(0)T}, \boldsymbol{b}^{(0)T}, \boldsymbol{\theta}^{(0)T})^T \\
&= (\underbrace{g_{(1)}(\bar{Y}_{(1)}), 0, \ldots, 0, g_{(2)}(\bar{Y}_{(2)}), 0, \ldots, 0, g_{(K)}(\bar{Y}_{(K)}), 0, .., 0}_{\boldsymbol{\beta}^{(0)T}}, \ldots \\
&\underbrace{-\ln(n), \ldots, -\ln(n)}_{\tilde{\boldsymbol{p}}^{(0)T}}, \underbrace{\mathbf{0}_n^T}_{\boldsymbol{b}^{(0)T}}, \underbrace{\mathbf{0}_{nK}^T}_{\boldsymbol{\theta}^{(0)T}})^T.
\end{aligned} \tag{3.13}$$

where every probability mass is initially set to $p_i^{(0)} = \frac{1}{n}$, the initial tilts $\boldsymbol{\theta}^{(0)}$ and normalising constants $\boldsymbol{b}^{(0)}$ to 0, along with nearly all initial regression coefficients. The exception being the intercept coefficient, which is initially set equal to the sample mean of the kth response $\bar{Y}_{(k)}$. This ensures $\boldsymbol{\mu}^{(0)}$ satisfies the convex hull condition, as

$$\boldsymbol{\mu}^{(0)} = \sum_{i=1}^{n} p_i^{(0)} \boldsymbol{Y}_i = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{Y}_i \ . \tag{3.14}$$

In the case where the intercept regression coefficient has been constrained to be shared across multiple models, the individual response means are instead replaced with the grand mean of that subset of constrained mean models, as each constrained model should have the same data type and link function. Some attempts have been made to speed up the optimization procedure by first fitting independent parametric GLMs to each mean model and using the regression coefficients fitted by these models as the initial guesses $\boldsymbol{\beta}^{(0)}$, in a hope that these parameter values are close to the solution produced by the VSPGLM. These attempts were unsuccessfully, with these new initial points for $\boldsymbol{\beta}^{(0)}$ producing a non feasible starting point for the algorithm, due to $\tilde{p}^{(0)}$ remaining unchanged, as this initial point for $\boldsymbol{\beta}^{(0)}$ will not satisfy the mean constraints to a desired tolerance.

To increase the numerical stability of the solution, each response component $\boldsymbol{Y}_{(k)i}$ is also centered and re-scaled onto the interval $[-1, 1]$, as was done by Wurm and Rathouz (2018)

for computing the univariate model. This process is done via the invertible transformation

$$\begin{aligned} \tilde{Y}_{(k)i} &= \left(Y_{(k)i} - \frac{M_{(k)} + m_{(k)}}{2}\right) \cdot \frac{2}{M_{(k)} - m_{(k)}} \\ \implies Y_{(k)i} &= \frac{M_{(k)} + m_{(k)}}{2} + \frac{M_{(k)} - m_{(k)}}{2}\tilde{Y}_{(k)i}, \end{aligned} \tag{3.15}$$

where $M_{(k)} = \max_{1 \leqslant i \leqslant n} \boldsymbol{Y}_{(k)}$ and $m_{(k)} = \min_{1 \leqslant i \leqslant n} \boldsymbol{Y}_{(k)}$. The link function for each response is also then modified to be

$$\tilde{g}_{(k)}(\mu_{(k)i}) = g\left(\frac{M_{(k)} + m_{(k)}}{2} + \frac{M_{(k)} - m_{(k)}}{2}\mu_{(k)i}\right), \tag{3.16}$$

which has the inverse

$$\tilde{\mu}_{(k)i} = \left(\mu_{(k)i} - \frac{M_{(k)} + m_{(k)}}{2}\right) \cdot \frac{2}{M_{(k)} - m_{(k)}}. \tag{3.17}$$

The parameter estimates for $\boldsymbol{\beta}$ and $\boldsymbol{p}$ obtained using the transformed response $\tilde{\boldsymbol{Y}}$ are equal to those obtained using the original response, as the log-likelihood is invariant under this transformation. This can easily be shown by placing the modified observations $\tilde{\boldsymbol{Y}}$ into the empirical log-likelihood (3.6). For a single observation this is

$$\ell_j(\boldsymbol{\beta}, \tilde{\boldsymbol{p}}, \boldsymbol{\theta}, \boldsymbol{b}) = \tilde{p}_j + b_j + \sum_{k=1}^{K} \theta_{(k)j}\left(\nu_{(k)} + \rho_{(k)}\tilde{Y}_{(k)i}\right), \tag{3.18}$$

where

$$\nu_{(k)} = \frac{M_{(k)} + m_{(k)}}{2} \text{ and } \rho_{(k)} = \frac{M_{(k)} - m_{(k)}}{2}. \tag{3.19}$$

Using the definition given in (2.36) for $b(\boldsymbol{\theta})$

$$b(\boldsymbol{\theta}) = -\log\left\{\sum_{i=1}^{n} p_i \exp\left\{\boldsymbol{\theta}^T \boldsymbol{Y}_i\right\}\right\}, \tag{3.20}$$

using the transformed response, this is now

$$b(\boldsymbol{\theta}) = -\log\left\{\sum_{i=1}^{n} p_i \exp\left\{\sum_{k=1}^{K} \theta_{(k)}\left(\nu_{(k)} + \rho_{(k)}\tilde{Y}_{(k)i}\right)\right\}\right\}. \tag{3.21}$$

Substituting (3.21) into the log-likelihood results in

$$\ell_j = -\tilde{p}_j - \log\left\{\sum_{i=1}^{n} p_i \exp\left\{\sum_{k=1}^{K} \theta_{(k)}\left(\nu_{(k)} + \rho_{(k)}\tilde{Y}_{(k)i}\right)\right\}\right\} + \sum_{k=1}^{K} \theta_{(k)j}\left(\nu_{(k)} + \rho_{(k)}\tilde{Y}_{(k)i}\right). \tag{3.22}$$

Here, cancellation from (3.21) and the summation in (3.18) occurs. This produces

$$\ell_j = -\tilde{p}_j - \log\left\{\sum_{i=1}^{n} p_i \exp\left\{\tilde{\boldsymbol{\theta}}_j^T \tilde{\boldsymbol{Y}}_j\right\}\right\} + \tilde{\boldsymbol{\theta}}_j^T \tilde{\boldsymbol{Y}}_j, \tag{3.23}$$

where $\tilde{\boldsymbol{\theta}}_i = (\rho_{(1)}\theta_{(1)i}, \ldots, \rho_{(K)}\theta_{(k)i})^T$. The mean constraint (3.8) can then be written as

$$\begin{aligned}
\mu_{(k)}(\boldsymbol{X}_{(k)j}^T\boldsymbol{\beta}_k) &= \sum_{i=1}^{n} Y_{(k)i} \exp\left\{b_j + \boldsymbol{\theta}_j^T\boldsymbol{Y}_i + \tilde{p}_i\right\} \\
&= \frac{\sum_{i=1}^{n}(\nu_{(k)} + \rho_{(k)}\tilde{Y}_{(k)i}) \exp\left\{\sum_{k=1}^{K}\theta_{(k)}\left(\nu_{(k)} + \rho_{(k)}\tilde{Y}_{(k)i}\right) + \tilde{p}_i\right\}}{\sum_{i=1}^{n} p_i \exp\left\{\sum_{k=1}^{K}\theta_{(k)}\left(\nu_{(k)} + \rho_{(k)}\tilde{Y}_{(k)i}\right)\right\}} \\
&= \nu_{(k)} + \rho_{(k)}\frac{\sum_{i=1}^{n}\tilde{Y}_{(k)i}\exp\left\{\tilde{\boldsymbol{\theta}}_j^T\tilde{\boldsymbol{Y}}_j + \tilde{p}_i\right\}}{\sum_{i=1}^{n} p_i \exp\left\{\tilde{\boldsymbol{\theta}}_j^T\tilde{\boldsymbol{Y}}_j\right\}} \quad , \qquad (3.24)
\end{aligned}$$

and modified mean function, $\tilde{\mu}_{(k)i} = \frac{1}{\rho_{(k)}}(\mu_{(k)i} - \nu_{(k)})$, then becomes

$$\begin{aligned}
\tilde{\mu}_{(k)}(\boldsymbol{X}_{(k)j}^T\boldsymbol{\beta}_k) &= \frac{\sum_{i=1}^{n}\tilde{Y}_{(k)i}\exp\left\{\tilde{\boldsymbol{\theta}}_j^T\tilde{\boldsymbol{Y}}_j + \tilde{p}_i\right\}}{\sum_{i=1}^{n} p_i \exp\left\{\tilde{\boldsymbol{\theta}}_j^T\tilde{\boldsymbol{Y}}_j\right\}} \\
&= \sum_{i=1}^{n}\tilde{Y}_{(k)i}\exp\left\{\tilde{\boldsymbol{\theta}}_j^T\tilde{\boldsymbol{Y}}_j + \tilde{p}_i + \tilde{b}_i\right\},
\end{aligned}$$

where

$$\tilde{b}_i = -\log\left\{\sum_{i=1}^{n} p_i \exp\left\{\tilde{\boldsymbol{\theta}}_j^T\tilde{\boldsymbol{Y}}_j\right\}\right\} .$$

Which shows that both the original and transformed link functions produce the same implicit definition for $\boldsymbol{\theta}$ and b_i.

Constraints are also placed on $\boldsymbol{\theta}_i$ such that

$$-\frac{500}{K} \leqslant \theta_{(k)i} \leqslant \frac{500}{K}, \quad i = 1, \ldots, n, k = 1, \ldots, K, \qquad (3.25)$$

where K is the total number of components in $\boldsymbol{Y}$. This constraint helps bound the quantity $|b_i + \boldsymbol{\theta}^T\tilde{\boldsymbol{Y}}_i|$, and keeps $\exp\{b_i + \boldsymbol{\theta}^T\tilde{\boldsymbol{Y}}_i\}$ within the numerical tolerance of MATLAB's floating point arithmetic. To get convergence of this method and a stable solution for some of the applications shown in Chapter 4, similar extra constraints sometimes had to also be placed on the transformed reference distribution parameters $\tilde{\boldsymbol{p}}$ and the normalising constants b_i. Another method that can be used to increase the numerical stability of the solution, and speed up the computational time, is to re-scale any covariates which are too large. For the smaller data sets, either in the number of observations or response components, these extra constraints have not been found to be necessary, however, scaling the covariates will usually produce a decrease in the computational time taken to fit the model.

The **gldrm** package of Wurm and Rathouz (2018) does not quite take the same approach to the one described in this chapter. Instead, they reduce the number of parameters in the optimization by sequentially solving for θ, $\boldsymbol{p}$ (they refer to the reference distribution as $\tilde{f}_0$) and then $\boldsymbol{\beta}$, using the Quasi-Newton method BFGS and Newton-Raphson iterations. They do not solve directly for the normalising constants b_i, as its relationship to the tilt parameters $\boldsymbol{\theta}$ (3.21) allows for each b_i to be empirically estimated at each iteration. This approach will potentially be adopted in future iterations of the code. The method of Wurm and Rathouz (2018) also reduces the number of reference distribution parameters by only attaching probability masses to unique values in the observed support for $\boldsymbol{Y}$. For Bernoulli data this can reduce the number of reference distribution parameters estimated to 2, however, when dealing with a vector response with mixed components there will most likely not be a noticeable decrease in the number of parameters to estimate, especially if one of the responses components is continuous. This may be added in future updates of the code, where truncation will be used to reduce the number of unique continuous responses. However, a trade off is present in the reduction in the number of unique observations and a lowering of the apparent variance present in the data. This will most likely have no effect on the methods point estimations of the mean model parameters, but may have some influence on parameter inference derived from the standard errors of the estimate of the asymptotic covariance matrix.

The next chapter will showcase several applications of the proposed VSPGLM, and how each differently structured model can be fit using the `fit_vspglm` interface via the formula syntax.

Chapter 4

Applications and Examples

4.1 Introduction

This chapter will provide several example applications of the proposed VSPGLM to showcase its versatility, and compare its parameters estimates and inference, using both the Wald test and the ELRT, to existing methods. This includes applications to some of the motivating examples mentioned in Chapter 1, these being the Burn Injury dataset (Fan and Gijbels 1996), which has response components of mixed data types, and the Butterfly dataset (Oliver, Prudic, and Collinge 2006), which has a large number of response components, and has previously been modelled by Hui et al. (2013) using both species distribution models (SDMs) and species archetype models (SAMs). This chapter also includes applications of the VSPGLM to several smaller datasets which are common examples in existing VGLM literature. The first of which is an application to data from a cerebrovascular drug trial with a two-way crossover design. This datasets has bivariate binary responses, and can also be tackled using a full parametric VGLM, specifically the quadratic exponential model outlined in Chapter 1 (Diggle 2013). The last two applications of the VSPGLM shown in this chapter are to the hospital visit dataset, which contains four count responses and was modelled in Song (2007) using a $4-$variate Poisson VGLM, and data from a sorbinil retinopathy trial (Rosner, Glynn, and Lee 2006), which has bivariate ordinal responses, and was also modelled in Huang (2017) using GEEs.

4.2 Burn Injury

The first example that we will look at in this chapter is an application of the VSPGLM to data from the Burn Injury dataset, which has two responses of different data types. This data has previously been used as a test case for general vector regression by Huang (2017) using GEEs, and also has analysis by Fan and Gijbels (1996) and Song (2007). The latter of which used copula functions to produce correlations in the marginal distributions. The dataset contains 981 observations of hospital patients admitted to hospital, each with some amount of 3rd degree burns. The data has two response variables, a binary response Y_1 indicating the patient's disposition of death (0 = No, 1 = Yes), and the percentage of the patients body which was covered in 3rd degree burns at the time of admittance to hospital, Y_2. In the dataset the burn severity response has been placed on a scale of $0 - 10,000$ (Table 4.1), however, the transformation $Y_2 \to \log(Y_2 + 1)$ will instead be modelled, as a relative burn area measurement. This transformation is unnecessary when using the VSPGLM, but is done to be consistent with existing literature and compare parameter estimates to those of Huang (2017). In both the analysis presented in Huang (2017) and Song (2007), the main question of interest this dataset presented was how the age of patient effects their burn severity and probability of death.

Table 4.1: Snippet of the Burn Injury dataset.

patient	age	gender	burns	death
1	3	1	1000	1
2	39	0	1675	1
3	42	0	50	1
4	21	0	450	1
5	33	0	7200	0
6	68	1	9200	0
⋮	⋮	⋮	⋮	⋮

In Huang (2017) the following marginal mean models were used (4.1), using only the patients age as a covariate and a working assumption of independence between the prevalence of death and burn severity.

$$\begin{aligned} \mu_1 &= \mathbb{E}[Y_1|\text{age}] = \frac{\exp\{\beta_{(1)1} + \beta_{(1)2}\text{age}\}}{1 + \exp\{\beta_{(1)1} + \beta_{(1)2}\text{age}\}} \\ \mu_2 &= \mathbb{E}[Y_2|\text{age}] = \beta_{(2)1} + \beta_{(2)2}\text{age}. \end{aligned} \tag{4.1}$$

The parameter values of the GEE fitted models were

$$\begin{aligned}\hat{\mu}_1 &= \mathbb{E}[Y_1|\text{age}] = \frac{\exp\{-3.6891 + 0.0508\text{age}\}}{1 + \exp\{-3.6891 + 0.0508\text{age}\}} \\ \hat{\mu}_2 &= \mathbb{E}[Y_2|\text{age}] = 6.7118 + 0.0035\text{age} \ ,\end{aligned} \tag{4.2}$$

and to perform joint inference, a sandwich correction of variance was performed on the working independence covariance structure. This was done to to obtain consistent estimates of the true standard errors for the joint model, under the assumption that the mean models (4.1) are correctly specified. Using the same marginal mean models, with a logit and identity link, the VSPGLM fits the models

$$\begin{aligned}\hat{\mu}_{(1)} &= \mathbb{E}[Y_1|\text{age}] = \frac{\exp\{-3.657 + 0.0509\text{age}\}}{1 + \exp\{-3.657 + 0.0509\text{age}\}} \\ \hat{\mu}_{(2)} &= \mathbb{E}[Y_2|\text{age}] = 6.7318 + 0.0027\text{age},\end{aligned} \tag{4.3}$$

using the code

```
# Burn Injury model
burn_injury = fit_vspglm(["death ~ age", "burn_severity ~ age"],...
                  data,{'logit', 'id'});
```

which has very similar parameter estimates to those shown in (4.2), which are also the point estimates presented in Song (2007) for fitting two univariate GLMs to each margin. Moreover, the sandwich corrected standard errors for each slope in Huang (2017) are $\text{se}(\hat{\boldsymbol{\beta}}_{(1)2}) = 0.0051$ and $\text{se}(\hat{\boldsymbol{\beta}}_{(2)2}) = 0.0017$, which are relatively similar to $\text{se}(\hat{\boldsymbol{\beta}}_{(1)2}) = 0.0045$ and $\text{se}(\hat{\boldsymbol{\beta}}_{(2)2}) = 0.0019$ found using the VSPGLM's standard errors. The GEE model found that the marginal effect of age on both a patients burn severity and probability of death was significant, however, the VSPGLM does not find that a patients age has a significant effect on the severity of their burns ($p = 0.16$).

Table 4.2: VSPGLM non-intercept coefficient summary for burns model (4.3)

	estimate	se	t	p
death	0.0509	0.004	11.0	$< 2.22 \times 10^{-16}$
burns	0.0027	0.002	1.42	0.16
LogLikelihood(ℓ)	-6668.4			

4.3 Butterfly

The next example application of the VSPGLM comes via a butterfly species distribution dataset, originally reported in Oliver, Prudic, and Collinge (2006). This dataset contains counts of 33 species of butterfly which were observed at 66 locations in a national park in Boulder County, Colorado (Table 4.3). This example was included as it shows the VSPGLM's performance when dealing with an arbitrary number of correlated response components. Previous analysis of this dataset by Hui et al. (2013) used a variation of SDMs and SAMs on the 14 butterfly species which have greater than 10 total observations across the 66 locations (Appendix A.3, Table A.1). In their analysis, separate independent GLMs were fit to all 14 species using the covariates of habitat type (hay-field, mixed, short, tall), and measures of the height of buildings and concentration of vegetation at each location.

Table 4.3: Snippet of the butterfly dataset .

pieris rapae	colias philodice	colias eurytheme	...	habitat	building	urban vegetation
0	0	0	...	mixed	2.12	10.88
1	1	2	...	mixed	2.12	11.83
1	1	2	...	mixed	19.80	1.68
0	2	1	...	mixed	5.33	0
0	0	0	...	mixed	3.94	0
⋮	⋮	⋮	...	⋮	⋮	⋮

As was mentioned in Chapter 1, this is an example of when joint modelling should be preferred, but cannot be tackled using standard parametric methods due to both negative and positive correlations being present in the responses (Figure A.1). To fit the VSPGLM to these same species, 14 separate mean models were used (4.4), which have separate mean parameters $\boldsymbol{\beta}_{(k)}$ for each species and the same set of covariates $\boldsymbol{X}_i \in \mathbb{R}^6$ across all 66 locations.

$$\mu_{(k)i} = \mathbb{E}[Y_{(k)i}|\boldsymbol{X}_i] = \exp\{\boldsymbol{X}_i^T\boldsymbol{\beta}_{(k)}\}, k = 1, \ldots, 14, \ i = 1, \ldots, 66 \ . \tag{4.4}$$

The resulting model coefficients for the 14 dimensional model is shown in Table A.2 in Appendix A.3. A visualisation of how well the VSPGLM marginally fits each species is also shown in Figure 4.1.

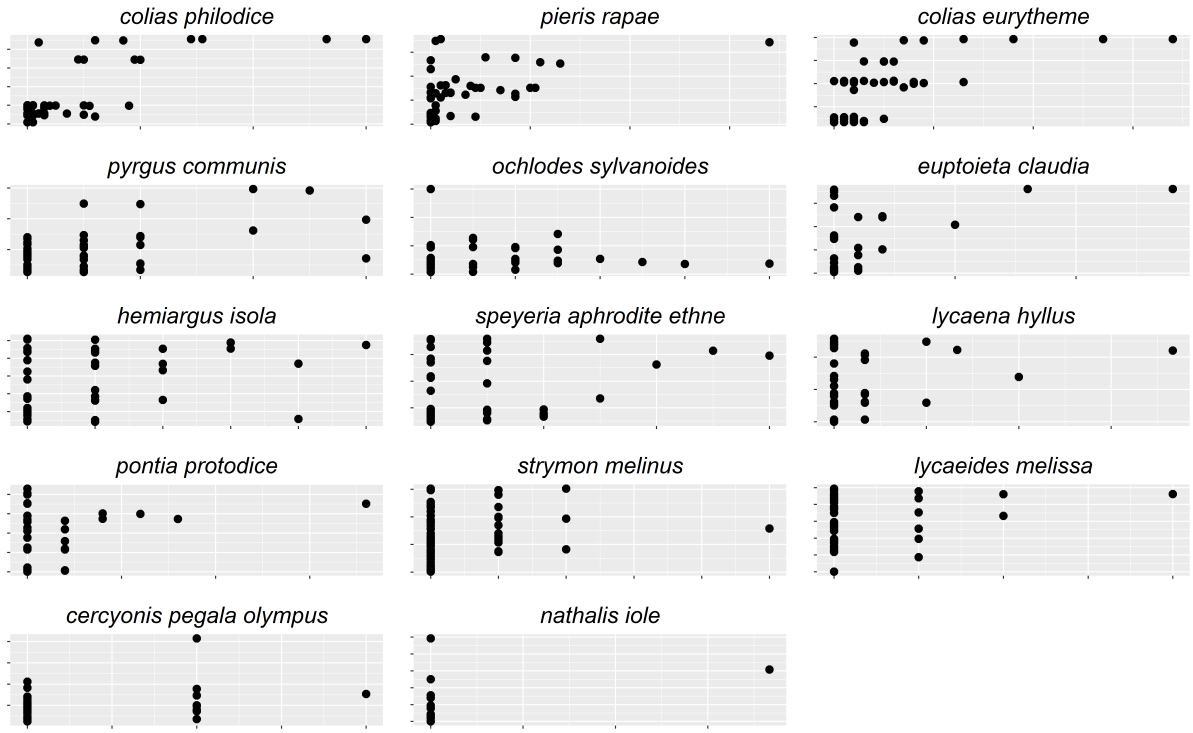

Figure 4.1: Predictions ($y-$axis) vs counts ($x-$axis) for 14−dimensional butterfly model shown in (4.4).

Due to the low number of counts for the majority of species, this analysis will from now on will only consider the 3 most frequently observed butterfly species in the dataset (*Colias philodic, Pieris rapae, Colias eurytheme*). Using the same mean models as before, the VSPGLM can be fit to these species using the following code

```
# Three Species Butterfly Model
butterfly_3 = fit_vspglm(["colias_philodice ~ (building,vegetation,habitat)",...
                          "pieris_rapae ~ (building,vegetation,habitat)", ...
                          "colias_eurytheme ~ (building,vegetation,habitat)"],...
                           data,{'log', 'log','log' });
```

The results of fitting the three separate log-link models is presented in A.3 (Table A.3), and Figure 4.2 compares the predicted means against the observed counts. Using this model, one can now undertake joint significance testing for common effects across species. The full model shows that the height of surrounding buildings has a non significant effect for all three species (Table A.3), and achieves the log-likelihood $\ell_{\text{full}} = -228.9$. To test the hypothesis $H_0 : \boldsymbol{\beta}_{\text{building}} = 0$ for these species, the building covariate is removed from the model. The resulting model has the log-likelihood $\ell_{\text{no building}} = -231.8$. Calibrating the

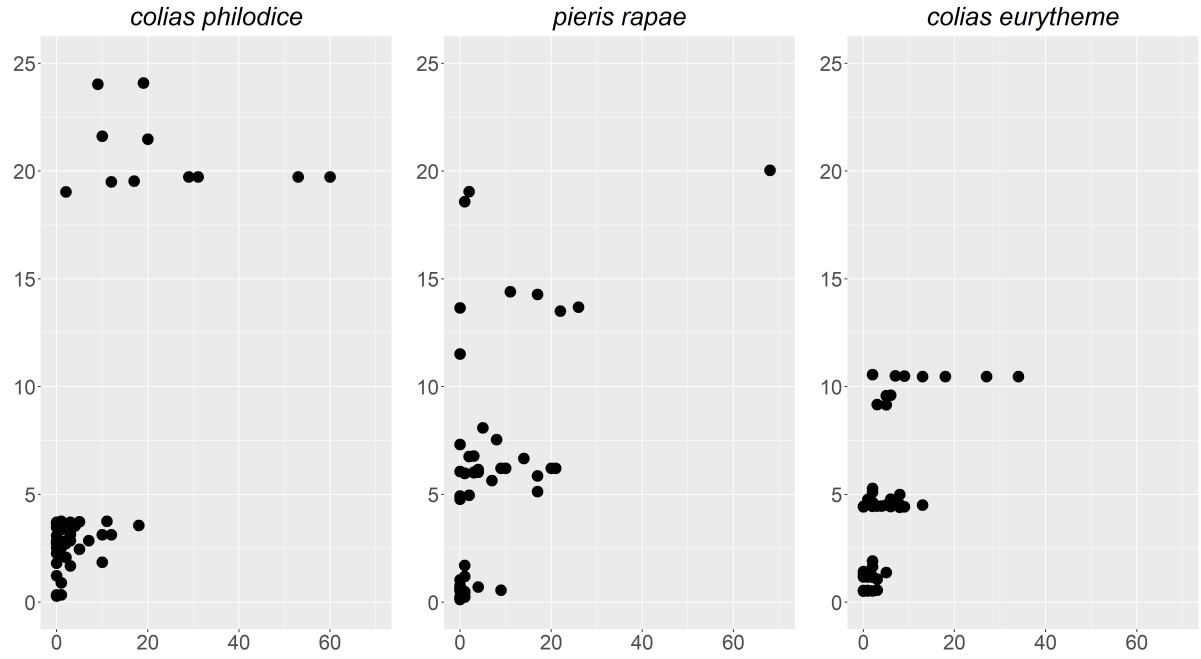

Figure 4.2: Predictions ($y-$axis) vs counts ($x-$axis) for 3−dimensional butterfly model.

ELRT against a $F_{3,66-18}$ distribution

$$2 \cdot (\ell_{\text{full}} - \ell_{\text{no building}}) = 5.8 \implies \mathbb{P}(F_{3,66-18} \geqslant 5.8/3) = 0.14, \tag{4.5}$$

one finds that there is no evidence that the building size has a significant effect on counts. By dropping the building and vegetation covariates out of each separate model, we find only the habitat type to have a significant effect across all three species (Table 4.4).

Table 4.4: Coefficient summary for 3 species habitat only model.

	coefficient	estimate	se	t	p
Colias philodice	intercept	3.17	0.20	15.74	2.220×10^{-16}
	mixed	−2.33	0.47	−4.95	5.940×10^{-6}
	short	−4.27	1.67	−2.56	0.013
	tall	−1.94	0.44	−4.41	4.220×10^{-5}
Pieris rapae	intercept	2.22	0.26	8.53	4.840×10^{-12}
	mixed	−2.86	1.93	−1.48	0.144
	short	−2.73	1.43	−1.90	0.062
	tall	0.02	0.42	0.05	0.960
Colias eurytheme	intercept	2.46	0.24	10.20	6.900×10^{-16}
	mixed	−2.23	0.49	−4.54	2.670×10^{-5}
	short	−3.09	0.84	−3.68	4.960×10^{-4}
	tall	−0.95	0.30	−3.23	0.002
LogLikelihood(ℓ)	-234.4				

With only the habitat covariates remaining in each mean model, we can now investigate if

the model

$$\mu_{(k),i} = \exp\{\boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 \boldsymbol{x}_{\text{mixed},i} + \boldsymbol{\beta}_2 \boldsymbol{x}_{\text{short},i} + \boldsymbol{\beta}_3 \boldsymbol{x}_{\text{tall},i}\}, \; i = 1, \ldots, 66,$$
$$k \in \{\textit{Colias philodic, Pieris rapae, Colias eurytheme}\}, \tag{4.6}$$

is sufficient, *i.e.*, the habitat types have an equal effect on the counts of all three species. For the constrained model we find $\ell_{\text{habitat, constrained}} = -254.65$ (Table A.4), calibrating the ELRT against a $F_{6,54}$ distribution produces the result

$$2 \cdot (\ell_{\text{habitat}} - \ell_{\text{habitat, constrained}}) = 40.5 \implies \mathbb{P}(F_{6,66-12} \geqslant 40.5/6) = 2.24 \times 10^{-5}, \tag{4.7}$$

which shows significant evidence of different habitat effects. The syntax for fitting the model (4.6) is:

```
# Three Species Constrained Habitati Only  Butterfly Model
butterfly_3 = fit_vspglm(["(colias_philodice,pieris_rapae,...
                           colias_eurytheme)  ~ habitat"],...
                           data,{'log'});
```

4.4 Two-Period Cross-Over Drug Trial

In the next example, the VSPGLM will be applied to data from a 2×2 cross-over design trial for a drug intended to increase cerebrovascular health. The trial consisted of $n = 67$ subjects, and compared the subjects outcomes when given either of drug A (the active drug) or drug B (the placebo). In this case $n = 34$ subjects were randomly assigned to the drug administering sequence AB and $n = 33$ assigned to the sequence BA. The response variables for this dataset are then pairs of binary variables $(Y_{(1)i}, Y_{(2)i})$, indicating whether the patients echo-cardiogram was normal after delivery of the active drug or placebo (1 = Normal, 0 = Abnormal), at each stage of the trial.

Table 4.5: Two-period cross-over cerebrovascular drug trial data ($n = 67$) (Jones and Kenward 2003)

Y_1	Y_2	**treatment 1**	**treatment 2**	$(n =)$
0	0	0	1	9
0	0	1	0	6
0	1	0	1	4
1	0	0	1	2
1	0	1	0	6
1	1	0	1	18
1	1	1	0	22
			total	67

Previous analysis in Song (2007) models both responses using a VGLM, which has marginal logistic mean models. If $Y_{(k)i}, k = 1, 2, i = 1, \ldots, 67$ is the binary response indicating if experimental subject i had an normal echo-cardiogram after treatment with drug k, then the following constrained model was fit (4.8), which has an additional interaction term between the treatment and period the drug was administered.

$$\mu_{(k)i} = \frac{\exp(\beta_0 + \beta_1 \text{treatment} + \beta_2 \text{period} + \beta_3 \text{treatment}{\cdot}\text{period})}{1 + \exp(\beta_0 + \beta_1 \text{treatment} + \beta_2 \text{period} + \beta_3 \text{treatment}{\cdot}\text{period})} \tag{4.8}$$

The model shown in (4.8) found significant evidence of a treatment effect for the drug ($Z = 1.98$), with $\hat{\beta}_1 = 1.17$ and $\text{se}(\hat{\beta}_1) = 0.59$, however, both the period and interaction effects were found to be non-significant.

To compare results, we chose to parametrise the VSPGLM slightly differently, fitting two separate mean models for each time point using only the treatment covariate, both using a logistic link (4.9).

$$\begin{aligned} \mu_{(1)i} &= \frac{\exp(\beta_{(1)0} + \beta_{(1)1} \text{treatment})}{1 + \exp(\beta_{(1)0} + \beta_{(1)1} \text{treatment})}, \quad i = 1, \ldots, 67 \\ \mu_{(2)i} &= \frac{\exp(\beta_{(2)0} + \beta_{(2)1} \text{treatment})}{1 + \exp(\beta_{(2)0} + \beta_{(2)1} \text{treatment})}, \quad i = 1, \ldots, 67 \end{aligned} \tag{4.9}$$

The code syntax to fit this model is shown below, and the VSPGLM results for both mean models are reported in Table 4.6.

```
# Two-Way Drug Trial Two Period Model
two_way = fit_vspglm(["y_1 ~ treatment1", "y_2 ~ treatment2"],...
                     data,{'logit', 'logit'});
```

Table 4.6: Coefficient summary for the two-period cross-over drug trial model (4.9).

period	coefficient	estimate	se	t	p
1	intercept	0.43	0.35	1.23	0.22
	treatment	1.11	0.58	1.92	0.06
2	intercept	0.61	0.36	1.69	0.10
	treatment	0.09	0.51	0.17	0.87
	LogLikelihood (ℓ)	-278.12			

These results appear to show that the treatment effect is only marginally significant in the first period, which suggests the symmetric model

$$\mu_{(k)i} = \frac{\exp(\beta_0 + \beta_1 \text{treatment})}{1 + \exp(\beta_0 + \beta_1 \text{treatment})}, \; k = 1, 2, \; i = 1, \ldots, 73, \tag{4.10}$$

which has a common treatment effect across both time periods. We can then use this model to test the hypothesis

$$H_0 : \; \beta_{(1)0} = \beta_{(2)0} \text{ and } \beta_{(1)1} = \beta_{(2)1} \tag{4.11}$$

for the model (4.9), *i.e.*, there is no difference in the subjects response to the drug when administered in either sequence. The code to fit the symmetric model uses the syntax,

```
# Two-Way Drug Trial Symmetric  Model
two_way = fit_vspglm(["(y_1, y_2)  ~  ((treatment_1 & treatment_2))"],...
                     data,{'logit', 'logit'});
```

and Table 4.7 contains the results.

Table 4.7: Coefficient summary for the cross-over drug trial symmetric model (4.10).

coefficient	estimate	se	t	p
intercept	0.52	0.25	2.07	0.04
treatment	0.56	0.23	2.40	0.02
LogLikelihood (ℓ)	-279.48			

To test the null hypothesis outlined in (4.11), the ELRT is calibrated against a $F_{2,67-4}$ distribution

$$2 \cdot (\ell - \ell_{\text{symmetric}}) = 2.72 \implies \mathbb{P}(F_{2,67-4} \geqslant 2.72/2) = 0.26, \tag{4.12}$$

which shows no evidence of different period or treatment effects over both time periods. Therefore, the symmetric model (4.10) is sufficient, which finds a significant common treatment effect of 0.56 across both drug administering sequences

4.5 Hospital Visit

Another model from Song (2007), for which the VSPGLM results can be compared to, is a 4-variate Poisson VGLM fit to a longitudinal dataset containing counts of children's hospital visits. This data set consisted of the number of visits $n = 73$ children made to a particular hospital over a year's four quarters. The covariates reported include the age of each child (0 to 68 months), whether the child's biological mother was a regular smoker (1 = Yes, and 0 = No) as well as the child's gender (1 = Female and 0 = Male) (Table 4.8).

Table 4.8: Snippet of the hospital visit dataset.

child	age	gender	smoking	quarter 1	quarter 2	quarter 3	quarter 4
1	63	1	1	5	0	0	1
2	8	1	1	3	4	2	2
3	31	1	1	0	0	1	2
4	33	1	1	2	0	1	1
5	24	0	1	1	0	0	0
6	34	0	1	2	0	2	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

The underlying motivation behind this analysis was to determine whether any of the previously stated baseline covariates have a significant effect on the number of quarterly hospital visits. In the original analysis by Song, it was noted that there was a significant difference in the number of hospital visits recorded in the winter quarter. This is visualised in Figure A.2 (Appendix A.4), which shows a proportion of children who did visit the hospital was significantly higher in the first quarter (winter), when compared with the remaining three quarters. Due to the significant difference, Song also introduced a seasonal covariate to indicate if the response was coming from the winter quarter. Four mean models were then fit (4.13), one for each quarter, using an interchangeable structure which constrains $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)$ to be common across all four time points.

$$\mu_{(k)i} = \exp\Big(\beta_0 + \beta_1 \text{age} + \beta_2 \text{smoking} + \beta_3 \text{season}\Big), \quad k = 1, \ldots, 4, \; i = 1, \ldots, 73 \tag{4.13}$$

This model only found that the seasonal effect to be significant, with the estimated regression coefficient for the season covariate being $\hat{\beta}_3 = 0.61$, which had the standard error $\text{se}(\hat{\beta}_3) = 0.1$.

To compare the results obtained using the VSPGLM with those obtained by Song, the same constrained mean models were used, each with a log-link and common regression coefficients. This model can be fit using the constrained formula syntax,

```
# Hospital Visit Model
hospital_visit = fit_vspglm(["(quarter1, quarter2, quarter3, quarter4) ~ .
                              (month, smoking, (season&0&0&0))"],...
                              data,{'log', 'log','log','log'});
```

where only the seasonal covariate is included in the mean model for the winter quarter (quarter 1), and all other coefficients are constrained to be common across all four quarters. The results for this model are summarized in Table 4.9.

Table 4.9: Coefficient summary for the VSPGLM hospital visit model (4.13).

coefficient	estimate	se	t	p
intercept	-0.496	0.34	-1.48	0.14
month	0.004	0.01	0.61	0.54
smoking	0.201	0.25	0.81	0.42
season	0.757	0.15	5.08	3.02×10^{-6}
LogLikelihood (ℓ)	-321.862			

The regression coefficients and standard errors found by the VSPGLM are relatively similar to those found previously by Song. Once again, only the seasonal effect was found to be significant, causing an expected multiplicative increase of $\exp(0.76) \approx 2.138$ hospital visits in the first quarter when compared to the remaining three quarters. To determine if any of the baseline covariates have a significant effect on hospital visits, both the age and smoking covariates are jointly removed from the model shown in (4.13). The log-likelihood for the resulting model is $\ell_{\text{season}} = -322.26$ (Table 4.9), with the ELRT giving the result

$$2 \times (\ell - \ell_{\text{season}}) = 0.91 \implies \mathbb{P}(F_{2,73-4} \geqslant 0.91/2) = 0.63. \tag{4.14}$$

Hence, we can reduce the model to

$$\mu_{(k)i} = \exp\Big(\beta_0 + \beta_2 \text{season}\Big), \quad k = 1, \ldots, 4, \ i = 1, \ldots, 73, \tag{4.15}$$

and conclude that non of the baseline covariates reported in the original dataset appear to have a significant effect on the hospital visit counts.

Table 4.10: VSPGLM hospital visit model coefficient summary, using only the seasonal covariate.

coefficient	estimate	se	t	p
intercept	-0.24	0.13	-1.80	0.08
season	0.77	0.15	5.17	1.99×10^{-6}
LogLikelihood (ℓ)	-322.26			

4.6 Sorbinil Trial

The next example application uses the VSPGLM to analyze data which originates from a sorbinil retinopathy drug trial, described by Rosner, Glynn, and Lee (2006). This trial consisted of $n = 41$ subjects who were randomly assigned to four treatment groups to have a combination of their left or right eye treated with sorbinil, or a placebo, after exposure to an allergen. Once treated, the subjects had itching scores for both eyes recorded on a 9 point Likert scale, starting from 0, the subject had no desire to itch, and increasing in increments of 0.5 to 4, the subject was experiencing an incapacitating itch to that eye. The full dataset is shown below in Table 4.11.

Table 4.11: Itching scores for each of the four treatment groups on the 9 point Likert scale, in increments of 0.5 from 0 to 4 (Rosner, Glynn, and Lee 2006).

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		

The ordinal nature of this data makes it relatively difficult to analyse using standard VGLM methods, as the mean functions for each margin have to be constrained to the support $[0,4] \times [0,4]$ for all combinations of treatments. Previous attempts at modelling this dataset have therefore employed some form of transformation of each response to the interval $[0,1]$, after which logistic mean models can be used. In this case, the responses are being treated as pseudo-proportions, as it is a somewhat reasonable assumption that the variances of each observation should approach zero for the extreme itching scores (McCullagh and Nelder 1989, pp. 328–332). This shows the versatility of the proposed VSPGLM, as there is no need to transform the response. Instead separate mean models can be fit using identity link functions, which increases the interpretability of the final model's parameters.

For each subject the response data can be denoted as the pair (Y_L, Y_R), the itchiness scores for the left and right eye, with the treatments denoted by $\mathcal{I}_L$ and $\mathcal{I}_R$, indicating which eye sorbinil was applied to. The first model that can be fit to this data are the separate mean models shown below (4.16), which allow for sorbinil to have a different treatment effect across both the left and right eye.

$$\begin{aligned} \mu_L &= \mathbb{E}[Y_L|\mathcal{I}_L] = \beta_{(L)0} + \beta_{(L)1}\mathcal{I}_L \\ \mu_R &= \mathbb{E}[Y_R|\mathcal{I}_R] = \beta_{(R)0} + \beta_{(R)1}\mathcal{I}_R \end{aligned} \tag{4.16}$$

The model results are displayed in Table 4.12.

Table 4.12: Coefficient summary for the separate sorbinil models (4.16)

coefficient	estimate	se	t	p
$\beta_{(L)0}$	2.20	0.16	14.02	2.22×10^{-16}
$\beta_{(L)1}$	−0.20	0.21	−0.95	0.35
$\beta_{(R)0}$	2.40	0.13	18.48	2.22×10^{-16}
$\beta_{(R)1}$	−0.67	0.20	−3.35	1.80×10^{-3}
LogLikelihood(ℓ)	−146.97			

Of the two slopes, only $\hat{\beta}_{(R)2} = -0.67$ was found to be significant. This appears to show that sorbinil was only effective in significantly reducing the itchiness score on the subjects right eye, which is a relatively improbable result. To deal with this issue, the next model that can be fit is the symmetric model (4.17), where both marginal mean models are

constrained to have a single intercept and slope.

$$\begin{aligned}\mu_L &= \mathbb{E}[Y_L|\mathcal{I}_L] = \beta_0 + \beta_1 \mathcal{I}_L \\ \mu_R &= \mathbb{E}[Y_R|\mathcal{I}_R] = \beta_0 + \beta_1 \mathcal{I}_R \end{aligned} \tag{4.17}$$

The code syntax to fit the model shown in (4.17) is:

```
# Symmetric Sorbinil Model
sorbinal = fit_vspglm(["(left_eye, right_eye)~((left_treatment&right_treatment))"],.
                                                data,{'id', 'id'});
```

and results are displayed in Table 4.13.

Table 4.13: Coefficient summary for the symmetric sorbinil model (4.17).

coefficient	estimate	se	t	p
β_0	2.30	0.11	20.45	2.22×10^{-16}
β_1	-0.43	0.14	-3.08	3.74×10^{-3}
LogLikelihood(ℓ)	-147.94			

Using the results of this model, and that of (4.16), we can test if the two separate models can be simplified into a single model, with a common sorbinil treatment effect across both eyes. This is the null hypothesis $H_0 : \beta_{(L)0} = \beta_{(R)0}$ and $\beta_{(L)1} = \beta_{(R)1}$. Calibrating the ELRT against a $F_{2,41-4}$ distribution

$$2 \cdot (\ell_{\text{separate}} - \ell_{\text{symmetric}}) = 1.94 \implies \mathbb{P}(F_{2,41-4} \geqslant 1.94/2) = 0.38,$$

we find that symmetry is an adequate model for the data. The next most natural thing to check, is if there is a significant interference effect which occurs if sorbinil has been administered to the eye for which the response is not collected from. To test this, the following models can be fit (4.18), which have a common additive interference term across both eyes.

$$\begin{aligned}\mu_L &= \mathbb{E}[Y_L|\mathcal{I}_L] = \beta_0 + \beta_1 \mathcal{I}_L + \underbrace{\beta_2 \mathcal{I}_R}_{\text{additive interference}} \\ \mu_R &= \mathbb{E}[Y_R|\mathcal{I}_R] = \beta_0 + \beta_1 \mathcal{I}_R + \underbrace{\beta_2 \mathcal{I}_L}_{\text{additive interference}} \end{aligned} \tag{4.18}$$

The syntax for fitting this model is:

```
# Additive Interference Sorbinil Model
sorbinal = fit_vspglm(["(left_eye,right_eye)~((left_treatment&right_treatment), .
                                              (right_treatment&left_treatment))"],.
                                                  data,{'id', 'id'});
```

The results are summarized in Table 4.14, which show no evidence of a significant additive interference effect.

Table 4.14: Coefficient summary for the additive interference sorbinil model (4.18).

coefficient	estimate	se	t	p
β_0	2.29	0.20	11.69	3.73×10^{-14}
β_1	−0.42	0.19	−2.27	0.03
β_2	0.02	0.19	0.10	0.92
LogLikelihood	−147.94			

Lastly, we can test if there is any interaction interference present between the left and right eye's itching scores

$$\begin{aligned} \mu_L = \mathbb{E}[Y_L|\mathcal{I}_L] &= \beta_0 + \beta_1 \mathcal{I}_L + \underbrace{\beta_2 \mathcal{I}_R \mathcal{I}_L}_{\text{interaction interference}} \\ \mu_R = \mathbb{E}[Y_R|\mathcal{I}_R] &= \beta_0 + \beta_1 \mathcal{I}_R + \underbrace{\beta_2 \mathcal{I}_R \mathcal{I}_L}_{\text{interaction interference}} , \end{aligned} \tag{4.19}$$

via the following syntax:

```
# Interaction Interference Sorbinil Model
sorbinal = fit_vspglm(["(left_eye,right_eye)~((left_treatment&right_treatment),.
                                                interaction)"],.
                                               data,{'id', 'id'});
```

[1] The results of this model are shown in Table 4.15.

Table 4.15: Coefficient summary for the interaction interference sorbinil model (4.19).

coefficient	estimate	se	t	p
β_0	2.31	0.12	20.00	2.26×10^{-16}
β_1	−0.43	0.15	−2.79	0.01
β_2	−0.04	0.23	−0.20	0.84
LogLikelihood	−147.93			

[1] Currently interaction covariates must be hard coded in the table argument to `fit_vspglm`. This will updated shortly.

Here, once again, there is no significant evidence of any interaction effect. This was also the conclusion of Huang (2017), who found the symmetric model (4.17) to be adequate for this data.

To visualise how the VSPGLM is modelling the joint distribution of (Y_L, Y_R), the estimated probability masses can also be plotted, as it will exists on a discrete 9×9 lattice with points at every half integer increment between 0 and 4. In Figure 4.3 the fitted probability masses are shown, re-weighted by the median observation of each treatment group, along with the location of the predicted mean for each of the four treatment groups. This shows that, in the treatment group where both eyes are treated with sorbinil, the distribution is relatively symmetric around the main diagonal, however, in the cases when only one eye is receives sorbinil the distribution is tilted to attach greater probability masses to higher itching scores from the eye which only receives a placebo. Finally, in the placebo treatment group the distribution has been tilted to attach greater probability masses to observations for which both the left and right eye have high itching scores.

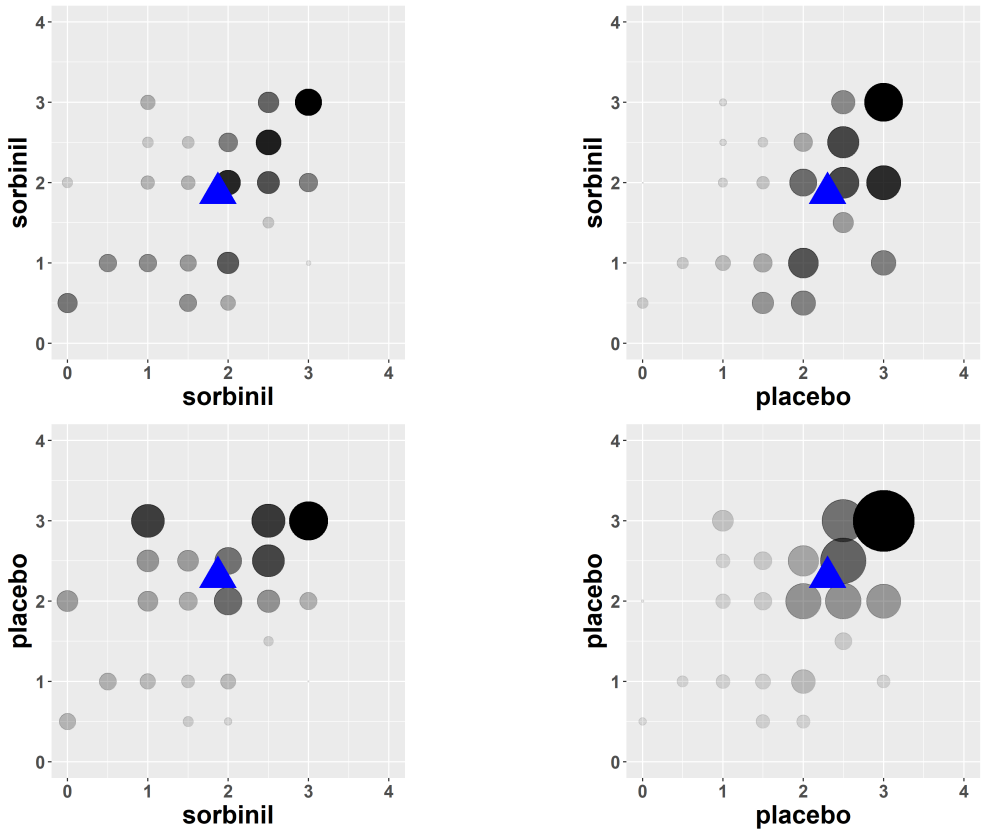

Figure 4.3: Estimated joint pmf for the symmetric sorbinil model tilted at the median itching scores of each treatment group. Itching scores for the right and left eye are shown on the x and y axis, with their treatments given via the axis label. Each probability mass is sized and coloured dependent on its value, with larger darker circles indicated greater mass. The predicted means for each of the four treatment groups are located at the blue triangular marker.

4.7 Cochlear

The last application that we will consider, but the VSPGLM will not actually be applied to, is the Cochlear dataset. This dataset contains data from a study measuring the effectiveness of noise reduction algorithms in improving the speech receptive threshold (SRT) of hearing impaired individuals. In the study, the SRT scores of 12 participants with cochlear implants were measured twice using six varieties of noise reduction algorithms. The algorithms used in this study consisted of a baseline standard noise reduction algorithm (Beam), and five variants (SpS0, SPz-3, Spz+3, SpZ+6, and SpZ). Each subject's SRT scores for each algorithms were measured in a randomised order during the first measurement period, with the order used during the initial period reversed for the second set of measurements.

Table 4.16: Snippet of the cochlear dataset SRT scores for subjects 1 to 12 (SO1-S12) .

	algorithm	SRT		algorithm	SRT	...		algorithm	srt
S01	SpZ+6	-7.3	S02	SpZ-3	-4.7	...	S12	SpZ+6	7.5
S01	SpZ0	-6.6	S02	SpS0	-0.8	...	S12	SpZ-3	7.2
S01	Beam	-4.0	S02	Beam	0.1	...	S12	SpZ+3	7.9
S01	SpS0	-6.5	S02	SpZ0	-4.5	...	S12	SpS0	7.9
S01	SpZ-3	-8.5	S02	SpZ+3	-3.7	...	S12	SpZ0	8.8
S01	SpZ+3	-8.8	S02	SpZ+6	-5.9	...	S12	Beam	11.4
S01	SpZ+3	-8.5	S02	SpZ+6	-5.8	...	S12	Beam	9.5
S01	SpZ-3	-7.7	S02	SpZ+3	-6.3	...	S12	SpZ0	3.2
S01	SpS0	-7.2	S02	SpZ0	-6.5	...	S12	SpS0	4.9
S01	Beam	-5.4	S02	Beam	-2.3	...	S12	SpZ+3	4.2
S01	SpZ0	-8.4	S02	SpS0	-3.8	...	S12	SpZ-3	5.0
S01	SpZ+6	-7.8	S02	SpZ-3	-4.2	...	S12	SpZ+6	4.9

This dataset is relatively unique, due to the design of the study, and allows a series of hypothesis to be investigated relating to the within subject correlations present during each measurement period. Currently, similarly structured studies are modelled using repeated measures multiple analysis of variance (MANOVA), which has an assumption of sphericity which is marginally violated in the case for this dataset, and requires certain corrections to be made to keep the type I errors at an acceptable level. We believe the VSPGLM may be a viable alternative to the use of repeated measures MANOVA, due to the lack of parametric assumptions necessary. Unfortunately, the VSPGLM code does not yet converge numerically on this dataset, with the mean and normalisation constraints consistently violated outside an acceptable tolerance level for the model to be fit adequately. Therefore, this particular application's results is not included in this thesis, but will be the subject of future research.

Chapter 5 will now present the results of several simulation studies which have been run using the VSPGLM.

Chapter 5

Simulations

5.1 Introduction

To test properties of the VSPGLM outlined in Chapter 2, in Sections 5.2 to 5.7 of this chapter a range of simulation studies are conducted to verify the consistency and asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$. For each simulation the bias, standard errors, type I error rate (or power) and 90%, 95% and 99% confidence interval coverage rates of each parameter in $\hat{\boldsymbol{\beta}}$ is reported. In Section 5.8 several simulations studies are also run to verify the power and the type I error rates of the ELRT when calibrated to an F distribution.

To test this models performance across a variety of settings, the first set of simulation studies are conducted using an array of standard and non-standard models to generate the data to which the VSPGLM is fit to. The first model simulated from is a correlated bivariate normal distribution, having both different and common mean parameters across both margins. Generalized linear mixed models (GLMMs) are then used to simulate both positively a negatively correlated trivariate responses which are marginally Poisson, Bernoulli, and gamma. A generalized linear mixed effects model was also used to simulate data which has correlated Poisson, gamma and normal margins, with a variation of this model simulating data for both studies conducted in the second set of simulations. For similar sample sizes, the coverage levels and type I error rates obtained from the simulations are comparable to those reported for the simulation studies conducted using similar models (Marchese 2018; Huang 2014; Wurm and Rathouz 2018).

In the simulation results for the VSPGLM fit to the Poisson and gamma GLMMs,

(Sections 5.4 and 5.6) some of the regression parameters show over-coverage relative to the 90%, 95% and 99% nominal levels, with the biggest discrepancy occurring for coverage at the 90% confidence level. This is a property of the SPGLM family of models from which this VSPGLM originates, and occurs when dealing with responses which lie on the boundary of the convex hull. At this point the tilted distribution experiences little to no variance, which inflates the standard errors derived from the estimate of the asymptotic covariance matrix for $\hat{\boldsymbol{\beta}}$. Some discussion of how to correct this problem is given in Huang and Rathouz (2012), who added a stability parameter to the optimization, penalizing iterations which are too near to the boundary of the convex hull. Wurm and Rathouz (2018) also enforced an identifiability constraint at each iteration of their method, which achieved a similar result. This chapter uses a slightly different method, correcting the standard errors after the VSPGLM is fit. The procedure used in this chapter to adjust a particular margins inflated standard errors to a nominal level is outlined in Section 5.9.

5.2 Multivariate Normal Simulation

The most standard multivariate distribution that data will be simulated from in this chapter is the multivariate normal distribution. In this case, the data is generated from a bivariate normal distribution with a non diagonal covariance matrix, inducing correlated marginal distributions. The data $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ were sampled via the procedure

$$(Y_{(1)i}, Y_{(2)i}) \sim \mathcal{N}((1 + \boldsymbol{X}_{(1)i}^T\boldsymbol{\beta}_{(1)}, 1 + \boldsymbol{X}_{(2)i}^T\boldsymbol{\beta}_{(2)})^T, \boldsymbol{\Sigma}),\ i = 1, \ldots, n, \tag{5.1}$$

where the covariance matrix $\boldsymbol{\Sigma}$ is given by

$$\boldsymbol{\Sigma} = \begin{pmatrix} 0.8 & 0.4 \\ 0.4 & 1.2 \end{pmatrix}, \tag{5.2}$$

and the true mean model parameters for both margins are

$$\begin{aligned} \boldsymbol{\beta}_{(1)} &= (-1, 0)^T \\ \boldsymbol{\beta}_{(2)} &= (0.5, 2.2)^T. \end{aligned} \tag{5.3}$$

These values were chosen to compare the VSPGLM's results to those obtained by the MDRM (Marchese 2018), which used a similar true covariance matrix $\boldsymbol{\Sigma}$ and values for $\boldsymbol{\beta}$. The covariates for each simulation, $\boldsymbol{X}_{(1)}$ and $\boldsymbol{X}_{(2)}$, were independently sampled uniformly between -1 and 1. This covariate sampling procedure was used in all simulations conducted in this chapter, mainly to reduced the total computational time, as in most cases covariates

can be centred and scaled with minimal impact on the final models interpretability. For each simulation the VSPGLM was fit to the data using following two mean models

$$\begin{aligned}\mu_{(1)} &= \boldsymbol{\beta}_{(1)0} + \boldsymbol{X}_{(1)}^T\boldsymbol{\beta}_{(1)} \\ \mu_{(2)} &= \boldsymbol{\beta}_{(2)0} + \boldsymbol{X}_{(2)}^T\boldsymbol{\beta}_{(1)}.\end{aligned} \tag{5.4}$$

Overall, $N = 1000$ simulations were conducted, each time using a sample size of $n = 200$. The results of these simulations are reported in Table 5.1, which show good agreement between $\boldsymbol{\beta}$ and $\hat{\boldsymbol{\beta}}$, with an average bias of the order of 10^{-3}. The numerical standard errors, $\hat{\sigma}$, and those derived from the asymptotic covariance matrix are also in close agreement $\|\hat{\sigma} - \bar{\text{se}}(\hat{\boldsymbol{\beta}})\|_\infty = 0.0071$, and the type I error for true null parameter $\boldsymbol{\beta}_{(1)2}$ is at the nominal 5% level. The coverage levels for all three of the 90%, 95% and 99% confidence intervals is close to the theoretical values for all regression coefficients, with all results reported also agreeing with those obtained by the MDRM in Marchese (2018).

Table 5.1: Simulation coverage for $N = 1000$ simulations, each using $n = 200$ samples from a bivariate normal model with covariance matrix $\Sigma = ((0.8, 0.4)^T|(0.4, 1.2)^T)$ and true regression coefficients $\boldsymbol{\beta}_{(1)} = (-1, 0)^T$ and $\boldsymbol{\beta}_{(2)} = (0.5, 2.2)^T$.

$\boldsymbol{\beta}$	$\hat{\boldsymbol{\beta}}$	$\lvert\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\rvert$	$\hat{\sigma}$	$\bar{\text{se}}(\hat{\boldsymbol{\beta}})$	$p \leqslant 0.05$	90%	95%	99%
−1	−0.9996	−0.0004	0.11	0.11	1	0.91	0.95	0.99
0	−0.0005	0.0005	0.11	0.11	0.05	0.90	0.95	0.99
0.5	0.4906	0.0094	0.14	0.13	0.95	0.88	0.94	0.99
2.2	2.1969	0.0031	0.14	0.14	1	0.90	0.94	0.99

5.3 Constrained Multivariate Normal Simulation

The next model that we will simulate data from is a bivariate normal distribution, which now has common true $\boldsymbol{\beta}$ parameters across both models

$$(Y_{(1)i}, Y_{(2)i}) \sim \mathcal{N}((1 + \boldsymbol{X}_{(1)i}^T\boldsymbol{\beta}, 1 + \boldsymbol{X}_{(2)i}^T\boldsymbol{\beta})^T, \Sigma),\ i = 1, \ldots, n, \tag{5.5}$$

where $\Sigma = ((0.8, 0.4)^T|(0.4, 1.2)^T)$ is unchanged from the last simulation and $\boldsymbol{\beta} = (-1, 0, 0.67)^T$. For each simulation, the VSPGLM was fit to the data using the following two mean models

$$\begin{aligned}\mu_{(1)} &= \boldsymbol{\beta}_0 + \boldsymbol{X}_{(1)}^T\boldsymbol{\beta} \\ \mu_{(2)} &= \boldsymbol{\beta}_0 + \boldsymbol{X}_{(2)}^T\boldsymbol{\beta}.\end{aligned} \tag{5.6}$$

The results of these simulations are reported in Table 5.2. Here, once again we see

Table 5.2: Simulation results for $N = 1000$ simulations, each with a sample size of $n = 200$ from a bivariate normal model with covariance matrix $\Sigma = ((0.8, 0.4)^T | (0.4, 1.2)^T)$ and regression coefficients $\boldsymbol{\beta} = (-1, 0, 0.67)^T$

$\boldsymbol{\beta}$	$\hat{\boldsymbol{\beta}}$	$\lvert\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\rvert$	$\hat{\sigma}$	$\bar{\text{se}}(\hat{\boldsymbol{\beta}})$	$p \leqslant 0.05$	90%	95%	99%
-1	-0.9957	0.0043	0.07	0.07	1	0.89	0.95	0.99
0	0.0030	0.0030	0.07	0.07	0.07	0.88	0.93	0.98
0.67	0.6668	0.0032	0.07	0.07	1	0.91	0.96	0.99

that the average estimates show little bias, with relatively good agreement between the standard errors for all three parameters, $\|\hat{\sigma} - \bar{\text{se}}(\hat{\boldsymbol{\beta}})\|_\infty = 0.0047$. The coverage at the each of the 90%, 95% and 99% confidence intervals is close to the theoretical values, with slightly lower coverage at the 90% and 95% levels for the true null parameter $\boldsymbol{\beta}_2$, resulting in type I error rates for $\boldsymbol{\beta}_2$ which are slightly higher than the nominal $\alpha = 0.05$ significance level.

5.4 Multivariate Poisson Simulation

The process of simulating from a multivariate Poisson distribution is slightly more involved than simulating from a multivariate normal distribution, although several different multivariate generalisations of the Poisson distribution have previously been proposed (*e.g.*, Krishnamoorthy 1951). A common problem across these generalisations is that they only allow for positive correlation between each of their k components, as most are constructed by the addition of a common Poisson random variable to k other independent Poisson random variables. This was also mentioned as one of the downsides of the multivariate CMP model of Sellers et al. (2021), which also has this assumption.

In this case, the goal is to simulate data from a model which has both negatively and positively correlated marginal Poisson distributions, as the non-negative correlation assumption would be violated by the Butterfly dataset analysed in Section 4.3 (Figure A.1). To do this, data is simulated from the following non-standard trivariate GLMM.

$$\begin{aligned}
\alpha &\sim \mathcal{N}(0, \sigma_0^2), \quad \sigma \in \mathbb{R}_+ \\
\boldsymbol{X} &\sim \ \mathcal{U}[-1, 1]^3 \\
Y_1 | \alpha, \boldsymbol{X}_{(1)} &\sim \text{Pois}(\exp(\boldsymbol{X}_{(1)}^T \boldsymbol{\beta}_{(1)} + \alpha)) \\
Y_2 | \alpha, \boldsymbol{X}_{(2)} &\sim \text{Pois}(\exp(\boldsymbol{X}_{(2)}^T \boldsymbol{\beta}_{(2)} + 0.5\alpha)) \\
Y_3 | \alpha, \boldsymbol{X}_{(3)} &\sim \text{Pois}(\exp(\boldsymbol{X}_{(3)}^T \boldsymbol{\beta}_{(3)} - 0.3\alpha))
\end{aligned} \tag{5.7}$$

In this case, all three marginal distributions have an association, both positive and negative, which is induced by the common intercept parameter $\alpha \sim \mathcal{N}(0, \sigma_0^2)$, which has different weightings for each response component and σ_0^2 is fixed to 1 for all simulations. Once again, the covariates were independently sampled from a $\mathcal{U}[-1, 1]$ distribution for each simulation, and the α weightings used were $1, 0.5$, and -0.3. A total of $N = 1000$ simulations were run, each with a sample size of $n = 200$ and true $\boldsymbol{\beta}$ parameter values of

$$\boldsymbol{\beta}_{(1)} = 0, \ \boldsymbol{\beta}_{(2)} = -0.8, \ \boldsymbol{\beta}_{(3)} = 0.3 \ . \tag{5.8}$$

The log link was used to fit all three correctly specified mean models, and the results are reported below in Table 5.3.

Table 5.3: Simulation results for $N = 1000$ simulations using a sample size of $n = 200$ from the trivariate Poisson model (5.7), with true $\boldsymbol{\beta}$ parameters $\boldsymbol{\beta}_{(1)} = 0.4, \ \boldsymbol{\beta}_{(2)} = -0.8, \ \boldsymbol{\beta}_{(3)} = 0$.

$\boldsymbol{\beta}$	$\hat{\boldsymbol{\beta}}$	$\lvert\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\rvert$	$\hat{\sigma}$	$\bar{\text{se}}(\hat{\boldsymbol{\beta}})$	$p \leqslant 0.05$	90%	95%	99%
0.4	0.3947	0.0053	0.12	0.12	0.89	0.90	0.95	0.99
-0.8	-0.8063	0.0063	0.14	0.14	1.00	0.92	0.97	0.99
0	0.0075	0.0075	0.12	0.12	0.04	0.91	0.96	0.99

These results show the consistency of the estimates for $\boldsymbol{\beta}$, with $\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\|_\infty = 0.004$. They also show relatively high power for both the non-null $\boldsymbol{\beta}$ parameters, and the type I error for $\boldsymbol{\beta}_{(3)} = 0$ is right at the nominal 0.05 level. One discrepancy occurs for the coverage levels shown for the marginal Poisson for Y_2, which has a true $\boldsymbol{\beta}$ parameter of -0.8 and the α coefficient 0.5. The coverage rates for estimates derived from empirical likelihood methods tend to converge from below (Diciccio, Hall, and Romano 1991), therefore, the high coverage rates for both the 90% (0.93) and 95% (0.96) confidence interval are slightly unusual. This problem was discussed in the introduction to this chapter, and when using a standard error adjustment at the 0.9 quantile (Section 5.9) coverage levels of $0.89, 0.95$ and 0.99 are obtained.

5.5 Multivariate Bernoulli Simulation

To simulate from a correlated multivariate binary distribution, a similar procedure was followed to the previous Poisson simulation study. The VSPGLM was fit to the following non-standard trivariate GLMM, using a logit link for each of the three mean models, with

$\sigma_0^2 = 1$.

$$
\begin{aligned}
\alpha &\sim \mathcal{N}(0, \sigma_0^2), \quad \sigma \in \mathbb{R}_+ \\
\boldsymbol{X} &\sim \ \mathcal{U}[-1, 1]^3 \\
Y_1|\alpha, \boldsymbol{X}_{(1)} &\sim \text{Ber}(1/(1 + \exp(-\boldsymbol{X}_{(1)}^T \boldsymbol{\beta}_{(1)} + \alpha))) \\
Y_2|\alpha, \boldsymbol{X}_{(2)} &\sim \text{Ber}(1/(1 + \exp(-\boldsymbol{X}_{(2)}^T \boldsymbol{\beta}_{(2)} + 0.5\alpha))) \\
Y_3|\alpha, \boldsymbol{X}_{(3)} &\sim \text{Ber}(1/(1 + \exp(-\boldsymbol{X}_{(3)}^T \boldsymbol{\beta}_{(3)} - 0.3\alpha)))
\end{aligned}
\tag{5.9}
$$

In this case, the true mean model parameters are the same set of parameters used in the previous simulation, and once again $N = 1000$ simulations were conducted each with a sample size of $n = 200$. The results from this simulation study are shown in Table 5.4, which shows relatively low bias for each of the parameter estimates, as well as good coverage at each of the $90\%, 95\%$ and 99% nominal levels for $\boldsymbol{\beta}_{(1)}, \boldsymbol{\beta}_{(2)}, \boldsymbol{\beta}_{(2)}$.

Table 5.4: Simulation results for $N = 1000$ simulations using a sample size of $n = 200$ from the trivariate correlated Bernoulli model (5.9), with true $\boldsymbol{\beta}$ parameters $\boldsymbol{\beta}_{(1)} = 0.4$, $\boldsymbol{\beta}_{(2)} = -0.8$, $\boldsymbol{\beta}_{(3)} = 0$.

$\boldsymbol{\beta}$	$\hat{\boldsymbol{\beta}}$	$\lvert\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\rvert$	$\hat{\sigma}$	$\bar{\text{se}}(\hat{\boldsymbol{\beta}})$	$p \leqslant 0.05$	90%	95%	99%
0.4	0.3929	0.0071	0.25	0.25	0.34	0.92	0.96	0.99
−0.8	−0.8148	0.0148	0.28	0.27	0.88	0.90	0.95	0.99
0	−0.0086	0.0086	0.27	0.25	0.07	0.88	0.94	0.99

All of these results are very similar to those found for the MDRM (Marchese 2018), when simulating from a hierarchical binary normal mixture model, however, these simulation results appear to show lower bias in the mean parameter estimates for the binary responses, when using the same sample size. One area where the VSPGLM parameter estimates performs poorly, relative to the Poisson simulation, is in the power levels achieved (the MDRM simulations did not report power), with $\boldsymbol{\beta}_{(1)} = 0.4$ only detected as significant ($\alpha = 0.05$) in 34% of simulations. This is most likely due to the scale of the covariates, with the random $\alpha \sim \mathcal{N}(0, \sigma_0^2)$ term potentially being much larger than $\boldsymbol{X}_{(2)}^T \boldsymbol{\beta}_{(2)}$ in some cases, controlling the magnitude of the expit$(\cdot)$ function used to generate the data. This issue can easily be fixed by using greater sample sizes or scaling the covariates, due to the good agreement between $\hat{\boldsymbol{\beta}}$ and the nominal coverage rates.

5.6 Multivariate Gamma Simulation

To simulate from a multivariate gamma distribution we follow a similar process to the previous two simulations, using the following non-standard trivariate GLMM.

$$\begin{aligned} \gamma &\sim \mathcal{N}(0, \sigma_0^2), \quad \sigma_0 \in \mathbb{R}_+, \alpha \in \mathbb{R}_+ \\ \boldsymbol{X} &\sim \mathcal{U}[-1, 1]^3 \\ Y_1|\gamma, \boldsymbol{X}_{(1)} &\sim \text{Gamma}(\alpha, \exp(\boldsymbol{X}_{(1)}^T \boldsymbol{\beta}_{(1)} + \gamma)) \\ Y_2|\gamma, \boldsymbol{X}_{(2)} &\sim \text{Gamma}(\alpha, \exp(\boldsymbol{X}_{(2)}^T \boldsymbol{\beta}_{(2)} + 0.5\gamma)) \\ Y_3|\gamma, \boldsymbol{X}_{(3)} &\sim \text{Gamma}(\alpha, \exp(\boldsymbol{X}_{(3)}^T \boldsymbol{\beta}_{(3)} - 0.3\gamma)) \end{aligned} \tag{5.10}$$

Here, the syntax $\text{Gamma}(\alpha, \beta)$ is used, where $\alpha > 0$ is the distribution's shape parameter, and $\beta > 0$ is the scale parameter. The true marginal means of each model, using a log-link, then should be

$$\begin{aligned} \mu_{(k)} &= \alpha\beta = \alpha \exp(\boldsymbol{X}_{(k)}^T \boldsymbol{\beta}_{(k)} + c\gamma) \\ \implies \ln(\mu_{(k)}) &= \ln(\alpha) + c\gamma + \boldsymbol{X}_{(k)}^T \boldsymbol{\beta}_{(k)}, \end{aligned} \tag{5.11}$$

where c is an arbitrary constant. This will allow for easy comparison of $\hat{\boldsymbol{\beta}}$ to the true parameter values.

For all simulations α is some fixed positive constant, and γ is now a standard normal random variable ($\sigma_0^2 = 1$) which introduces negative or positive correlations, and has the weightings of 1, 0.5 and -0.3 across each marginal component. Here, the true values of the $\boldsymbol{\beta}$ parameters are

$$\boldsymbol{\beta}_{(1)} = 0.5, \ \boldsymbol{\beta}_{(2)} = -1, \ \boldsymbol{\beta}_{(3)} = 0, \tag{5.12}$$

and for all simulations the gamma shape parameter α was fixed to be 1.25. The results from $N = 1000$ simulations, each using a sample size of $n = 200$ are shown in Table 5.5. For the first and third marginal models, which have true mean parameters $\boldsymbol{\beta}_{(0)} = 0.5$ and $\boldsymbol{\beta}_{(3)} = 3$, the results are expected. As was the case with the multivariate Poisson simulation, over coverage is observed in the 90% and 95% confidence intervals for the marginal model with the true mean parameter, $\boldsymbol{\beta}_{(3)} = -1$, which has the largest magnitude. This is the result of a discrepancy between the numerical standard deviation $\hat{\sigma}_{(2)} = 0.11$ and $\bar{\text{se}}(\hat{\boldsymbol{\beta}}_{(2)}) = 0.14$. Using a corrective quantile of 0.75 (Section 5.9), nominal coverage levels of $0.91, 0.95, 0.99$ with a corrected standard error of 0.11 can be obtained. Here the coverage rates at the 90% level also slightly outperform those found in Wurm and Rathouz

(2018), when simulating from a gamma GLM with only three non intercept coefficients and a sample size of $n = 100$.

Table 5.5: Simulation results for $N = 1000$ simulations using a sample size of $n = 200$ from the trivariate gamma model (5.10), with $\alpha = 1.25$ and true $\boldsymbol{\beta}$ parameters $\boldsymbol{\beta}_{(1)} = 0.5,\ \boldsymbol{\beta}_{(2)} = -1,\ \boldsymbol{\beta}_{(3)} = 0$.

$\boldsymbol{\beta}$	$\hat{\boldsymbol{\beta}}$	$\lvert\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\rvert$	$\hat{\sigma}$	$\bar{\text{se}}(\hat{\boldsymbol{\beta}})$	$p \leqslant 0.05$	90%	95%	99%
0.5	0.4911	0.0089	0.11	0.11	1.00	0.91	0.95	0.99
-1	-0.9951	0.0049	0.11	0.14	1	0.96	0.99	1.00
0	-0.0062	0.0062	0.11	0.11	0.06	0.89	0.94	0.99

5.7 Trivariate Mixed Normal, Poisson, Gamma Simulation

To show the general applicability of this model, it can also be applied to the following hierarchical trivariate generalized mixed effects model, which has correlated marginal distributions of mixed data types. In this case, the trivariate model is comprised of correlated normal, gamma, and Poisson marginal distributions, which are once again associated due to some random effect $\alpha \sim \mathcal{N}(0, \sigma_0^2)$, with different weightings across each margin.

$$\begin{aligned}
\alpha &\sim \mathcal{N}(0,\ \sigma_0^2) \\
\boldsymbol{X} &\sim \mathcal{U}[-1, 1]^3 \\
\boldsymbol{Y}_1 | \boldsymbol{X}_{(1)}, \alpha &\sim \mathcal{N}(\boldsymbol{X}_{(1)}^T \boldsymbol{\beta}_{(1)} + \alpha,\ \sigma_1^2) \\
\boldsymbol{Y}_2 | \boldsymbol{X}_{(2)}, \alpha &\sim \text{Pois}(\exp(\boldsymbol{X}_{(2)}^T \boldsymbol{\beta}_{(2)} + 0.5\alpha)) \\
\boldsymbol{Y}_3 | \boldsymbol{X}_{(3)}, \alpha &\sim \text{Gamma}(\lambda,\ \exp(\boldsymbol{X}_{(3)}^T \boldsymbol{\beta}_{(3)} - 0.3\alpha))
\end{aligned} \tag{5.13}$$

Once again, $N = 1000$ simulations were conducted using a sample size of $n = 200$. For all simulations σ_0^2 was fixed equal to 1, and the gamma distributions shape parameter λ set to 1.25. When fitting the VSPGLM, the identity-link was used for the normal margin, and log-links where used for both the Poisson and gamma mean models. The results for this simulation is shown in Table 5.6.

Table 5.6: Simulation results for $N = 1000$ simulations using a sample size of $n = 200$ from the a trivariate mixed effects model (5.13) with correlated normal, Poisson and gamma margins.

margin	$\boldsymbol{\beta}$	$\hat{\boldsymbol{\beta}}$	$\lvert\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\rvert$	$\hat{\sigma}$	$\bar{\text{se}}(\hat{\boldsymbol{\beta}})$	$p \leqslant 0.05$	90%	95%	99%
Normal	1	1.0058	0.0058	0.17	0.18	1	0.91	0.96	0.99
Poisson	-0.5	-0.4994	0.0006	0.13	0.13	0.96	0.83	0.88	0.94
Gamma	0.4	0.3956	0.0044	0.12	0.13	0.85	0.89	0.93	0.98

These results show good performance for all marginal models, with the $\boldsymbol{\beta}$ estimates showing levels of bias on the order of 10^{-3}. The worst performing margin was that of the Poisson, having coverage of $0.83, 0.88, 0.93$ for the $90\%, 95\%$ and 99% confidence intervals.

5.8 F Simulations

The last set of simulation studies that will be conducted in this chapter are done to verify properties of the profile empirical likelihood ratio statistic for finite sample sizes, stated in Chapter 2, Corollary 2.1. Specifically that, when testing the composite hypothesis of the form $\boldsymbol{M}\boldsymbol{\beta} = \boldsymbol{\gamma}$, for finite samples the ELRT approximately follows a $rF_{r,n-Q}$ distribution, where $r = \text{rank}(\boldsymbol{M})$. The first simulation study in this section will focus on the type I error rates when using the finite sample adjustment, and the second study will look at the power of this test for joint inference.

In the first study, data is once again simulated from the mixed effects model (5.13). The VSPGLM is then fit twice to the data, once using the following set of mean models (5.14), and once dropping the covariates of $\boldsymbol{\beta}_{(2)2}$ and $\boldsymbol{\beta}_{(3)2}$ from the model.

$$\begin{aligned} \mu_{(1)} &= \boldsymbol{\beta}_{(1)0} + \boldsymbol{\beta}_{(1)1}\boldsymbol{X}_{(1)} \\ \mu_{(2)} &= \exp(\boldsymbol{\beta}_{(2)0} + \boldsymbol{\beta}_{(2)1}\boldsymbol{X}_{(2)} + \boldsymbol{\beta}_{(2)2}\boldsymbol{X}_{(3)}) \\ \mu_{(3)} &= \exp(\boldsymbol{\beta}_{(3)0} + \boldsymbol{\beta}_{(3)1}\boldsymbol{X}_{(3)} + \boldsymbol{\beta}_{(3)2}\boldsymbol{X}_{(1)}) \end{aligned} \tag{5.14}$$

For each simulation the null hypothesis $H_0 : \boldsymbol{\beta}_{(2)2} = \boldsymbol{\beta}_{(3)2} = 0$ is tested at a different significance levels by calibrating the ELRT between (5.14) and the true nested model to a $2F_{2,n-8}$ distribution. The matrix $\boldsymbol{M} \in \mathbb{R}^{2\times 8}$ and vector $\gamma \in \mathbb{R}^2$ for this composite hypothesis are

$$\boldsymbol{M} = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \quad \boldsymbol{\gamma} = (0, 0)^T. \tag{5.15}$$

In total, $N = 3000$ simulations were run for sample sizes of $n = 75, 150$, with the type I error rates at significance levels of $0.1, 0.05$ and 0.01 reported below in Table 5.7.

Table 5.7: Type I errors at the $0.10, 0.05$ and 0.01 significance levels for testing $H_0 : \boldsymbol{\beta}_{(2)2} = \boldsymbol{\beta}_{(3)2} = 0$, using sample sizes of $n = 75, 150$ and $N = 3000$ simulations.

	Type I Errors		
n	0.10	0.05	0.01
75	0.119	0.060	0.012
150	0.097	0.050	0.013

For $n = 75$, the type I error rates are close to the nominal levels, with very close agreement for $n = 150$. This shows the validity of the finite sample correction, with the type I error rates for $n = 75$ also being are very similar to those obtained in the simulation studies conducted in Huang (2014), who used a sample size of $n = 66$ and found rates of $11.7\%, 5.9\%$ and 1.3% when comparing the ELRT to a $F_{1,66-3}$ distribution..

The final simulation study of this chapter investigates the power of the proposed method for joint inference. The full model that is simulated from is shown below (5.16), which is very similar to the model used in the previous simulation, however, now there is a common covariate across both the Poisson and gamma margins.

$$\begin{aligned} \alpha &\sim \mathcal{N}(0,\ \sigma_0^2) \\ \boldsymbol{X} &\sim \mathcal{U}[-1, 1] \\ \boldsymbol{Y}_1|\boldsymbol{X}, \alpha &\sim \text{Pois}(\exp(\boldsymbol{X}^T\boldsymbol{\beta}_{(1)} + 0.5\alpha)) \\ \boldsymbol{Y}_2|\boldsymbol{X}, \alpha &\sim \text{Gamma}(\lambda,\ \exp(\boldsymbol{X}^T\boldsymbol{\beta}_{(2)} - 0.3\alpha)) \end{aligned} \tag{5.16}$$

The true values of $\boldsymbol{\beta}_{(1)}$ and $\boldsymbol{\beta}_{(2)}$ were varied between -1 and 1 in increments of 0.1, and at each pair of values 100 simulations were run, using a sample size of $n = 75$. After each simulation the null hypothesis $H_0 : \boldsymbol{\beta}_{(1)} = \boldsymbol{\beta}_{(2)} = 0$ was tested at a 0.05 significance level by comparing the ELRT to a $2F_{2,75-4}$ distribution. Figure 5.1 shows the relative frequency of rejection for each combination of $\boldsymbol{\beta}_{(1)}$ and $\boldsymbol{\beta}_{(2)}$ (the test's "power surface"), with the raw values given in Table A.5, Appendix A.5. The figure shows that the contours of the power surface are relatively circular and are centred at the origin, as expected, and for true values of $\boldsymbol{\beta}_{(1)}$ and $\boldsymbol{\beta}_{(2)}$ which are close to the 0, the power of the test is approximately 0.05, and approaches 1 as $\|\boldsymbol{\beta}\|_\infty \to 1$. With 80% power achieved for true $\boldsymbol{\beta}$ values laying outside a circle about the origin with an approximate radius of 0.6.

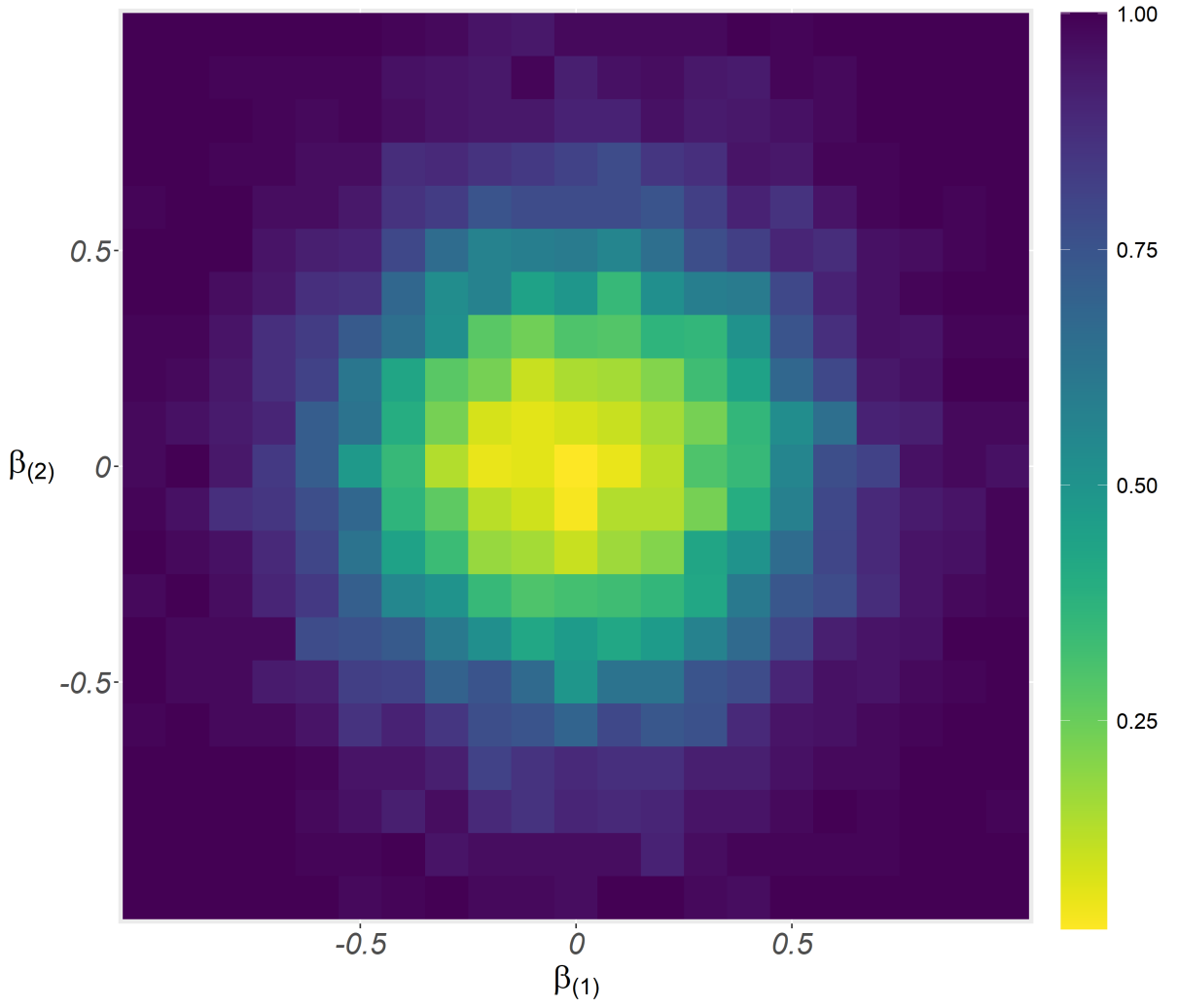

Figure 5.1: Power surface for $\boldsymbol{\beta}_{(1)}$ and $\boldsymbol{\beta}_{(2)}$. Each pixel shows the rejection rate over 100 simulations for testing $H_0 : \boldsymbol{\beta}_{(1)} = \boldsymbol{\beta}_{(2)} = 0$, using a sample size of $n = 75$.

5.9 Standard Error Adjustment

Due to the nature of this method, in that it re-tilts the reference distribution at every data point. In the case where a data point is at the boundary of the convex hull, the re-tilted distribution at that data point will have little to no variance. This is because all other probability masses in the tilted distribution will now be negligible compared to the probability mass which is attached to this point. Therefore, the inverse empirical covariance matrix for each observation

$$\Sigma_{\boldsymbol{Y}}^{-1} = \left\{ \sum_{j=1}^{n} p_j (\boldsymbol{Y}_j - \boldsymbol{\mu}_j)(\boldsymbol{Y}_j - \boldsymbol{\mu}_j)^T \exp\left\{ \boldsymbol{\theta}^T \boldsymbol{Y}_j + b \right\} \right\}^{-1} \tag{5.17}$$

which are used in the calculation of $\hat{\Sigma}$

$$\hat{\Sigma} = \left\{ \text{Var}\Big[\nabla pl(\boldsymbol{\beta})\Big] \right\}^{-1} \tag{5.18}$$

and hence the standard errors $\text{se}(\hat{\boldsymbol{\beta}}_{(k)i}) = \sqrt{\hat{\Sigma}_{(k)i,(k)i}}$, will, on the inverse scale, have a large spectral radius, $\rho(\Sigma_{\boldsymbol{Y}}^{-1})$. This will then have an out-sized influence on the calculated standard error, causing it to be larger than the $\hat{\sigma}$ (the numerical standard deviation), which inflates the confidence interval coverage rates. In the simulation studies for the trivariate Poisson and gamma GLMMs this is especially true at the 90% coverage level for the margins with the greatest discrepancy between $\hat{\sigma}$ and $\text{se}(\hat{\boldsymbol{\beta}})$.

The method that is used in this chapter to adjust the inflated standard errors is to truncate the spectral radius of the empirical covariance matrices before their initial inversion, thus reducing the magnitude of their inverted spectral radius. This method is performed by first computing the eigenvalues $\lambda_{(1)i}, \lambda_{(2)i}, \ldots, \lambda_{(K)i}$ of $\Sigma_{\boldsymbol{Y}_i}$, for all observations $i = 1, \ldots, n$, then for each component the eigenvalues are truncated down to a nominal quantile. We have found that quantile values, $q_{(k)}$, between $0.70 - 0.95$ appear to work well in our experimentation. Currently the truncation is done by first computing the singular value decomposition (SVD) of each $\Sigma_{\boldsymbol{Y}_i}$ matrix (5.19).

$$\Sigma_{\boldsymbol{Y}_i} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T \tag{5.19}$$

As each empirical covariance matrix is square and positive semi-definite, the singular values $\sigma_{(1)i} \ldots \sigma_{(K)i}$ and the eigenvalues $\lambda_{(1)i}, \ldots, \lambda_{(K)i}$ will coincide. The positive semi-definite assumption is only violated when there is no variance at all in the data, which is unlikely in any applied setting. Next, the diagonals of $\boldsymbol{\Sigma}$, for the margin requiring correcting, are truncated to the appropriate quantile $q_{(k)}$, producing the matrix $\tilde{\Sigma}$. The matrix can then be inverted via

$$\Sigma_{\boldsymbol{Y}_i}^{-1} = \boldsymbol{V}\tilde{\Sigma}^{-1}\boldsymbol{U}^T, \tag{5.20}$$

which adjusts the standard errors.

The final chapter will now conclude this thesis and discuss the proposed VSPGLM.

Chapter 6

Conclusion and Discussion

In this thesis a VSPGLM is introduced, which is a generalisation of the models of Huang (2014) and Rathouz and Gao (2009) for dealing with multivariate responses. The main advantage that this VSPGLM has over existing VGLMs is that it does not require any parametric assumptions to be made about the structure of the joint distribution from which the data originates. This allows the VSPGLM to be applied effectively with minimal effort to a wide variety of data, which usually require separate parametric models in each case. Examples of this were given in Chapter 4 (Applications and Examples) where the VSPGLM produced comparable parameter estimates and inference to existing parametric and semiparametric VGLMs across a number of different datasets. Chapter 4 also includes the application of the VSPGLM to the butterfly dataset (Section 4.3) which has both negatively and positively correlated components, and hence cannot be tackled using any standard parametric joint model.

The applicability of the proposed VSPGLM was also shown in the simulations studies conducted in Chapter 5, where the coverage rates and type I errors of the VSPGLM estimator $\hat{\boldsymbol{\beta}}$ were close to the nominal levels for each of the models simulated from, with a method for adjusting the standard errors proposed in the case of over-coverage. In all cases, the results from each simulation were also in agreement with those of the MDRM model of Marchese (2018), and those of Wurm and Rathouz (2018), when a direct comparison could be made. In particular, the finite sample adjustment for the ELRT was shown to produce nominal type I error rates close to those observed in Huang (2014), when using a comparable sample size.

There are several drawbacks to the proposed VSPGLM and the current way it is fit computationally. The first being that it can only give mean predictions for new observations predicted to lie outside the convex hull of the originally observed data, as a new nonparametric reference distribution would have to be estimated to obtain predictions of the variance or covariance present at this new point. A second downside of this VSPGLM relates to the computational fit of this model, with the current method optimizing for $Q + n(2 + K)$ parameters which are subject to $n(K + 1)$ constraints. This means that it is significantly slower to fit the VSPGLM for data which either has a large number of observations or response components. In Chapter 3, several methods from the **gldrm** **R** package were discussed for lowering both the number of parameters, and constraints, with these being potentially implemented in future iterations of the MATLAB code or in a separate **R** package.

The MATLAB code used to fit the VSPGLM is available at https://github.com/gden173/vspglm, along with the data and code necessary to fit each example shown in Chapter 4.

Bibliography

Butler, Ronald W. (2007). *Saddlepoint approximations with applications*. Cambridge: Cambridge University Press.

Dewanji, Anup and Lue Ping Zhao (2002). "An Optimal Estimating Equation with Unspecified Variances". *Sankhya. Series A* 64, 95–108.

Diciccio, T., P. Hall, and J. Romano (1991). "Empirical Likelihood is Bartlett-Correctable". *The Annals of statistics* 19, 1053–1061.

Diggle, Peter (2013). *Analysis of longitudinal data*. 2nd ed. Oxford: Oxford University Press.

Fan, Jianqing and I. Gijbels (1996). *Local polynomial modelling and its applications*. London: Chapman Hall.

Gałecki, Andrzej and Tomasz Burzykowski (2013). *Linear Mixed-Effects Models Using R A Step-by-Step Approach*. 1st ed. 2013. New York, NY: Springer New York.

Gill, Richard D., Yehuda Vardi, and Jon A. Wellner (1988). "Large Sample Theory of Empirical Distributions in Biased Sampling Models". *The Annals of statistics* 16, 1069–1112.

Huang, Alan (2014). "Joint estimation of the mean and error distribution in generalized linear models". *Journal of the American Statistical Association* 109, 186–196.

– (2017). "On generalised estimating equations for vector regression". *Australian New Zealand Journal of Statistics* 59, 195–213.

Huang, Alan and Paul Rathouz (2012). "Proportional likelihood ratio models for mean regression". *Biometrika* 99, 223–229.

– (2017). "Orthogonality of the mean and error distribution in generalized linear models". *Communications in Statistics-Theory and Methods* 46, 3290–3296.

Hui, Francis K. et al. (2013). "To mix or not to mix: comparing the predictive performance of mixture models vs. separate species distribution models". *Ecology* 94, 1913–1919.

Jones, Byron and Michael G. Kenward (2003). *Design and analysis of cross-over trials*. 2nd ed. Boca Raton, Fla.: Chapman Hall/CRC.

Kauermann, Göran and Raymond J. Carroll (2001). "A Note on the Efficiency of Sandwich Covariance Matrix Estimation". *Journal of the American Statistical Association* 96, 1387–1396.

Krishnamoorthy, A.S. (1951). "Multivariate binomial and Poisson distributions". *Sankhyā: the Indian Journal of Statistics*, 117–124.

Liang, Kung-Yee and Scott L. Zeger (1986). "Longitudinal data analysis using generalized linear models". *Biometrika* 73, 13–22.

Luo, Xiadong and Wei Yann Tsai (2012). "A proportional likelihood ratio model". *Biometrika* 99, 211–222.

Mahamunulu, D. M. (1967). "A Note on Regression in the Multivariate Poisson Distribution". *Journal of the American Statistical Association* 62, 251–258.

Marchese, Scott (2018). "Semiparametric Regression Methods for Mixed Type Data Analysis". Thesis.

McCullagh, Peter and John A. Nelder (1989). *Generalized Linear Models*. Vol. 37. CRC Press.

Murphy, S. A. and A. W. Van Der Vaart (2000). "On Profile Likelihood". *Journal of the American Statistical Association* 95, 449–465.

Oliver, J. C., K. L. Prudic, and S. K. Collinge (2006). "Boulder County Open Space Butterfly Diversity and Abundance: "Ecological Archives" E087-061". *Ecology (Durham)* 87, 1066.

Owen, Art B. (2001). *Empirical likelihood*. CRC press.

Prentice, R. L. and Lue Ping Zhao (1991). "Estimating Equations for Parameters in Means and Covariances of Multivariate Discrete and Continuous Responses". *Biometrics* 47, 825–839.

Rathouz, Paul J. and Liping Gao (2009). "Generalized linear models with unspecified reference distribution". *Biostatistics* 10, 205–218.

Rosner, Bernard, Robert J Glynn, and Mei-Ling T Lee (2006). "Extension of the rank sum test for clustered data: Two-group comparisons with group membership defined at the subunit level". *Biometrics* 62, 1251–1259.

Sellers, Kimberly F et al. (2021). "A Flexible Multivariate Distribution for Correlated Count Data". *Stats* 4, 308–326.

Song, Peter X. (2007). *Correlated data analysis: modeling, analytics, and applications*. Springer Science Business Media.

Van Der Vaart, A. W. and J. A. Wellner (1996). *Weak Convergence and Empirical Processes With Applications to Statistics*. 1st ed. 1996. New York, NY: Springer New York.

Wedderburn, R. W. (1974). "Quasi-Likelihood Functions, Generalized Linear Models, and the Gauss-Newton Method". *Biometrika* 61, 439–447.

Wurm, Michael J. and Paul J. Rathouz (2018). "Semiparametric Generalized Linear Models with the gldrm Package". *The R journal* 10, 288.

Yee, Thomas W. (2015). *Vector generalized linear and additive models: with an implementation in R*. Springer.

Appendix A

Appendix

A.1 Regularity Conditions

For all derivations shown in Section 2.6 the following regularity conditions are assumed to be satisfied.

Condition A.1.1. *The dimension of $\boldsymbol{\beta}$ remains fixed.*

Condition A.1.2. *The distributions generating the data have a common support, independent of $\boldsymbol{\beta}$.*

Condition A.1.3. *The parameter space θ is an open set of $\mathbb{R}^Q$.*

Condition A.1.4. *The true value of the parameters $\boldsymbol{\beta}^*$ lies in the interior of Θ.*

Condition A.1.5. *The first three derivatives of the the log likelihood $\ell(\boldsymbol{\beta})$ exist in an open set of Θ which contains the true parameter value $\boldsymbol{\beta}^*$, and within this set the mixed derivatives are bounded.*

Condition A.1.6. *For all $\boldsymbol{y} \in \mathcal{Y} \subset \mathbb{R}^K$ and $\boldsymbol{\beta} \in \Theta$, the log likelihood $\ell(\boldsymbol{\beta})$ is twice differentiable, and*

$$\nabla_{\boldsymbol{\theta}} \int_{\mathcal{Y}} f(\boldsymbol{y}, \boldsymbol{\theta}) d\boldsymbol{y} = \int_{\mathcal{Y}} \nabla_{\boldsymbol{\theta}} f(\boldsymbol{y}, \boldsymbol{\theta}) d\boldsymbol{y}$$
$$\nabla^2_{\boldsymbol{\theta}} \int_{\mathcal{Y}} f(\boldsymbol{y}, \boldsymbol{\theta}) d\boldsymbol{y} = \int_{\mathcal{Y}} \nabla^2_{\boldsymbol{\theta}} f(\boldsymbol{y}, \boldsymbol{\theta}) d\boldsymbol{y}$$

These regularity conditions are adapted from those given by Yee (2015, pp. 537–538).

A.2 Asymptotic Results

The asymptotic results for the univariate model of Huang (2014) were derived under the following three conditions.

Condition A.2.1. *The one dimensional response space* $\mathcal{Y} \subset \mathbb{R}$ *is contained in a finite closed interval and the space of covariates* $\boldsymbol{X} \in \mathcal{X} \subset \mathbb{R}^q$ *is contained with a finite rectangle in* $\mathbb{R}^q$.

Condition A.2.2. *There exists* $\delta_1 > 0$ *such that* $\mu(\boldsymbol{X}^T\boldsymbol{\beta})$ *maps onto* $\mathcal{Y}$*, both the first and second derivatives of* $\mu(\cdot)$ *exist and are continuous on the space*

$$\mathcal{X} \times \{\boldsymbol{\beta} \in \mathbb{R}^q : \ \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\| \leqslant \delta_1\}$$

Condition A.2.3. *There are a* $\delta_2 > 0$ *and a variance function* $V_2 > 0$ *such that* $V(x; \boldsymbol{\beta}, F) \geqslant V_2$ *on the space*

$$\mathcal{X} \times \{(\boldsymbol{\beta}, F) \in \mathbb{R}^q \times \mathcal{F} \ : \|\boldsymbol{\beta} - \boldsymbol{\beta}^*, F - F^*\| \leqslant \delta_2\}$$

Condition A.2.1 is common in GLM theory, with both the covariates and responses originating from some bounded region. Condition A.2.2 relates to the good behaviour of the link function, at least in a neighbourhood close to that of the true parameter values, and A.2.3 states that sufficiently close to the true parameter values the variance function $V(\boldsymbol{X}, \boldsymbol{\beta}, F)$ cannot get arbitrarily small. Conditions A.2.2 and A.2.3 are also used to satisfy sufficient conditions from theorems in Van Der Vaart and Wellner (1996) and Murphy and Van Der Vaart (2000) to show Proposition 2.2.

More information can be found in the supplementary material of the paper (Huang 2014).

A.3 Butterfly Example

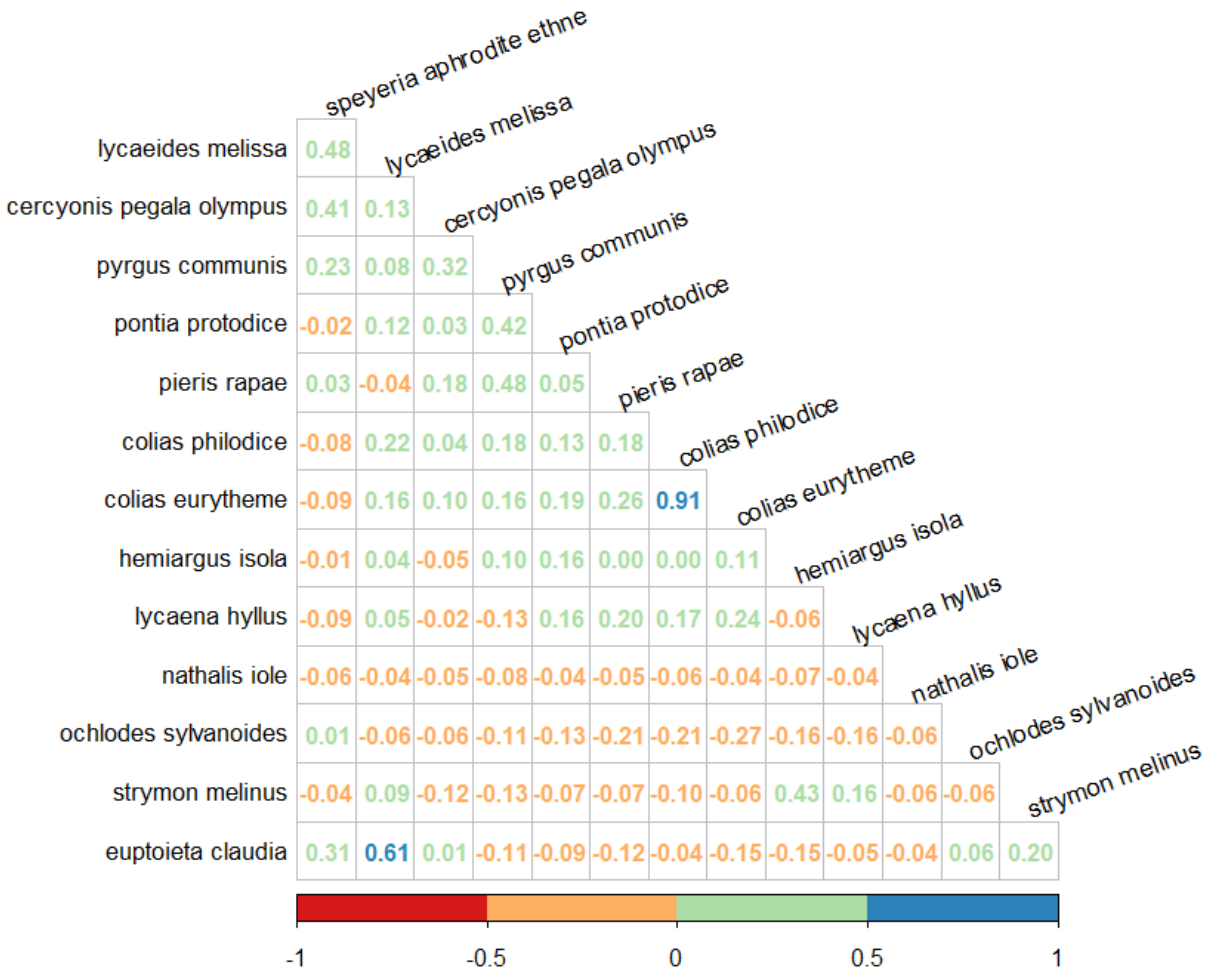

Figure A.1: Correlation plot of the 14 most populous species in the butterfly dataset (Table A.1). This shows both positive and negative correlations were observed among some species. The will obviously happen for species with a few observed counts, however, a negative correlation is also observed between the species *Ochlodes sylvanoides* and *Colias eurytheme* ($\rho = -0.27$) which have the 5th (59) and 3rd (256) highest number of total observed counts in the dataset.

Table A.1: Total counts for all 33 butterfly species across the 66 locations in Boulder County, Colorado (Oliver, Prudic, and Collinge 2006). Only three species (*Colias philodice*, *Pieris rapae*, *Colias eurytheme*) average more then one observation at each site.

	species	count
1	Colias philodice	383
2	Pieris rapae	317
3	Colias eurytheme	256
4	Pyrgus communis	59
5	Ochlodes sylvanoides	54
6	Euptoieta claudia	42
7	Hemiargus isola	41
8	Speyeria aphrodite ethne	39
9	Lycaena hyllus	38
10	Pontia protodice	28
11	Strymon melinus	26
12	Lycaeides melissa	15
13	Cercyonis pegala olympus	12
14	Nathalis iole	11
15	Hesperia leonardus pawnee	10
16	Danaus plexippus	10
17	Phyciodes campestris	8
18	Polites themistocles	5
19	Papilio polyxenes	5
20	Limenitis archippus	5
21	Vanessa cardui	4
22	Danaus gilippus	4
23	Speyeria edwardsii	3
24	Papilio rutulus	2
25	Junonia coenia	2
26	Pholisora catullus	1
27	Hesperia uncas uncas	1
28	Hesperia ottoe	1
29	Polites peckius	1
30	Polites mystic	1
31	Papilio multicaudatus	1
32	Everes comyntas	1
33	Plebejus acmon	1

Table A.2: Fitted regression coefficients of 14−dimensional butterfly species model (4.4). Here mixed, short and tall indicate the habitat covariates.

species	intercept	building	urban.vegetation	mixed	short	tall
colias philodice	3.1	0.00	-1.0×10^{-2}	−2.1	−3.9	−1.5
pieris rapae	2.0	0.11	2.1×10^{-2}	−2.9	−1.6	−0.3
colias eurytheme	2.5	0.00	-1.1×10^{-2}	−2.0	−2.5	−0.6
pyrgus communis	0.1	0.08	2.8×10^{-2}	−1.2	−1.5	−0.5
ochlodes sylvanoides	−2.6	−0.13	4.8×10^{-2}	2.6	2.0	1.6
euptoieta claudia	−1.6	−0.10	-1.6×10^{-2}	2.9	0.2	−0.5
hemiargus isola	−0.0	−0.04	-2.9×10^{-2}	−1.2	−0.7	0.3
speyeria aphrodite.ethne	−1.3	−0.05	-2.4×10^{-3}	1.8	0.4	0.5
lycaena hyllus	−0.5	−0.19	1.9×10^{-2}	−2.2	−2139.8	1.5
pontia protodice	−0.6	−0.08	4.9×10^{-2}	−2.9	−2.5	0.8
strymon melinus	−1.3	−0.26	-2.5×10^{-3}	0.8	−0.1	0.6
lycaeides melissa	−1.6	−0.08	-8.4×10^{-3}	0.9	−4.8	0.9
cercyonis pegala.olympus	−2.6	−0.07	5.2×10^{-2}	1.0	0.8	1.0
nathalis iole	-1.5×10^{5}	0.16	3.5×10^{-2}	1.5×10^{5}	1.4×10^{5}	-1.1×10^{5}

Table A.3: Coefficient summary for 3 species model (*Colias philodice*,*Pieris rapae*, *Colias eurytheme*) using all covariates.

	coefficient	estimate	se	t	p
Colias philodice	intercept	2.98	0.29	10.46	3.8×10^{-15}
	building	−0.06	0.07	−0.87	3.9×10^{-1}
	urban.vegetation	0.01	0.01	0.59	5.5×10^{-1}
	habitat.mixed	−1.93	0.44	−4.42	4.1×10^{-5}
	habitat.short	−4.04	1.62	−2.49	1.5×10^{-2}
	habitat.tall	−1.66	0.45	−3.66	5.3×10^{-4}
Pieris rapae	intercept	1.83	0.45	4.03	1.6×10^{-4}
	building	0.13	0.09	1.47	1.5×10^{-1}
	urban.vegetation	0.03	0.02	1.21	2.3×10^{-1}
	habitat.mixed	−3.88	1.99	−1.96	5.5×10^{-2}
	habitat.short	−2.47	1.18	−2.10	4.0×10^{-2}
	habitat.tall	−0.26	0.53	−0.50	6.2×10^{-1}
Colias eurytheme	intercept	2.35	0.28	8.24	1.9×10^{-11}
	building	0.02	0.04	0.66	5.1×10^{-1}
	urban.vegetation	−0.01	0.01	−0.38	7.0×10^{-1}
	habitat.mixed	−2.18	0.43	−5.08	3.9×10^{-6}
	habitat.short	−3.00	0.83	−3.59	6.7×10^{-4}
	habitat.tall	−0.86	0.32	−2.70	9.1×10^{-3}
	LogLikelihood(ℓ)	−228.87			

Table A.4: Coefficient summary for constrained butterfly model for the species *Colias philodic, Pieris rapae* and *Colias eurytheme* using only the habitat type (mixed, short, tall) as a covariates.

coefficient	estimate	se	t	p
intercept	2.0	0.2	10.7	1.11×10^{-15}
mixed	-1.7	0.4	-4.6	2.05×10^{-5}
short	-2.8	1.0	-2.8	0.01
tall	-0.4	0.2	-1.8	0.07
LogLikelihood(ℓ)	-254.6			

A.4 Hospital Visit Example

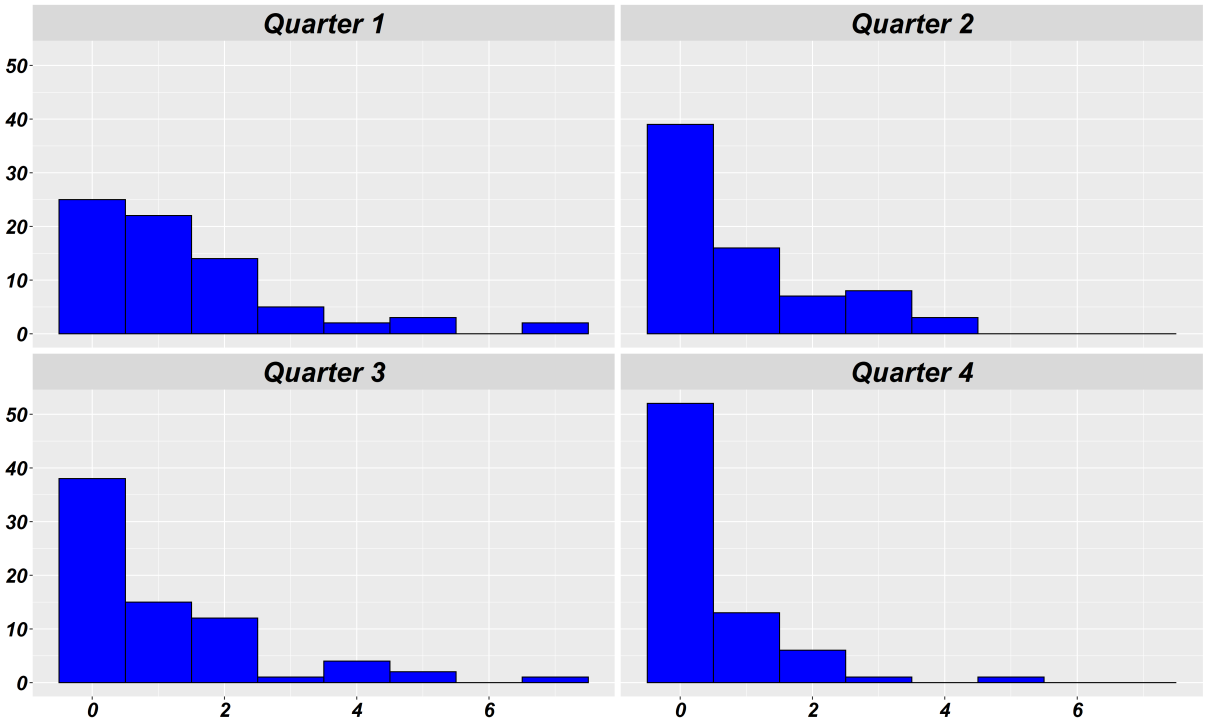

Figure A.2: Counts of hospital visits the $n = 73$ children made per quarter for the hospital dataset (Table 4.8). This shows a significant difference in the counts of children who did not visit the hospital in the winter quarter (25) relative to the remaining three quarters (38, 39, 52).

A.5 F Simulations Power

Table A.5: Number of significant ELRT results out of 100 ($n = 75$) for different values of $\boldsymbol{\beta}_{(1)}$ and $\boldsymbol{\beta}_{(2)}$ at a $\alpha = 0.05$ significance level (Figure 5.1).

$\boldsymbol{\beta}_{(1)}$ \ $\boldsymbol{\beta}_{(2)}$	-1	-0.9	-0.8	-0.7	-0.6	-0.5	-0.4	-0.3	-0.2	-0.1	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
−1	100	100	100	100	100	98	99	100	98	98	97	100	100	98	97	100	100	100	100	100	100
−0.9	100	100	100	100	99	99	100	95	97	97	97	97	91	97	99	99	99	99	100	100	100
−0.8	100	100	100	100	98	96	92	97	89	86	90	89	90	95	95	98	100	99	100	100	99
−0.7	100	100	100	100	99	95	95	92	81	86	89	87	87	92	92	96	98	98	100	100	100
−0.6	99	100	98	98	95	86	91	85	77	75	69	79	74	76	89	95	96	98	99	100	100
−0.5	100	98	98	93	92	82	81	70	75	67	49	63	63	75	78	90	96	95	98	99	100
−0.4	100	98	98	98	78	76	73	61	52	42	47	42	47	57	67	80	92	95	96	100	100
−0.3	98	100	97	90	84	71	55	50	35	30	32	33	36	42	61	74	78	88	95	98	99
−0.2	100	98	96	89	80	63	44	34	18	16	11	17	21	43	50	66	80	89	95	96	99
−0.1	99	96	87	85	77	68	37	27	13	10	4	14	14	23	40	58	79	89	93	95	99
0	98	100	94	84	72	48	35	14	6	7	3	6	13	30	35	56	77	81	96	98	96
0.1	98	96	93	90	72	63	40	23	9	7	9	11	16	23	36	53	65	91	92	98	98
0.2	99	98	94	87	81	62	43	28	23	11	15	16	21	33	44	68	79	94	96	100	100
0.3	99	99	95	87	83	73	65	52	28	24	30	29	37	36	51	75	87	96	95	99	99
0.4	100	100	97	94	87	86	68	53	57	44	49	35	52	59	60	79	91	96	99	100	100
0.5	100	100	100	95	92	91	79	66	57	59	60	56	65	77	82	90	88	96	97	99	100
0.6	99	100	100	97	97	94	86	83	75	78	78	78	75	82	91	86	95	99	100	99	100
0.7	100	100	99	99	97	97	88	89	86	84	81	78	85	87	95	94	99	99	100	100	100
0.8	100	100	100	99	98	99	97	95	94	94	91	91	96	93	94	96	98	100	100	100	100
0.9	100	100	99	99	99	99	96	95	94	99	92	96	97	94	93	99	98	100	100	100	100
1	100	100	100	100	100	100	99	98	95	94	98	98	98	98	100	99	100	100	100	100	100